



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

Αναγνώριση Μουσικών Τίτλων με
τεχνικές Ακουστικού Αποτυπώματος

Γρηγόρης Μπούρδαλας

Επιβλέπων Καθηγητής:
Εμμανουήλ Ψαράκης

Πάτρα, Μάιος 2018

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ.Εμμανουήλ Ψαράκη για την πολύτιμη κατατόπιση στην επιλογή του θέματος της παρούσας διπλωματικής εργασίας, καθώς και συνολικότερα για τις γνώσεις και τον ενθουσιασμό που μου μετέδωσε για τον τομέα της επεξεργασίας σημάτων κατά την διάρκεια των σπουδών μου. Επίσης, δεν θα μπορούσα παρά να ευχαριστήσω την οικογένεια μου και τους φίλους μου για όλη την υποστήριξη που μου έχουν δώσει όλο αυτό το διάστημα.

Περίληψη

Αρχικός στόχος της παρούσας διπλωματικής εργασίας είναι να γίνει μια εισαγωγή στην επιστημονική βιβλιογραφία που υπάρχει γύρω από το θέμα της αναγνώρισης ηχητικών σημάτων με τεχνικές που χρησιμοποιούν ακουστικό αποτύπωμα. Παρουσιάζονται διαφορετικές προσεγγίσεις που έχουν προταθεί για την λύση του προβλήματος της αναγνώρισης μουσικής. Στην συνέχεια, γίνεται αναλυτική περιγραφή και υλοποίηση πάνω σε συγκεκριμένο αλγόριθμο αναγνώρισης μουσικών τίτλων, που χρησιμοποιεί ως βασικό μαθηματικό εργαλείο παραγωγής του ακουστικού αποτυπώματος, το φασματογράφημα του Μετασχηματισμού Φουριέ Σύντομου Χρόνου. Έπειτα, υλοποιείται μια παραλλαγή αυτού, στην οποία το φασματογράφημα προκύπτει από τον Μετασχηματισμό Σταθερού-Q. Ο μετασχηματισμός αυτός έχει σημαντικά πλεονεκτήματα για μουσικές εφαρμογές, αφού μπορεί να μεταφράσει με ακρίβεια το ηχητικό σήμα στην ακολουθία της μουσικής μελωδίας. Επίσης, προτείνεται μια μέθοδος για την μείωση του όγκου δεδομένων της βάσης δεδομένων του συστήματος. Περιλαμβάνει την δημιουργία και τον έλεγχο της παραμέτρου ‘σημαντικότητας’ των κλειδιών που παράγει το ακουστικό αποτύπωμα, τα οποία και αποτελούν τα δεδομένα στην βάση. Τέλος, η απόδοση του συστήματος αναγνώρισης μουσικής εξετάζεται για τα δυο διαφορετικά ακουστικά αποτυπώματα στην ίδια βάση δεδομένων, με διάφορες μονάδες παραμορφώσεων να εφαρμόζονται στα άγνωστα αποσπάσματα εισόδου.

Abstract

The first objective of this thesis is an initial introduction in the field of Music Information Retrieval alongside the study of scientific papers that use Audio Fingerprinting mechanisms. We begin with the presentation of different approaches that solve the problem of MIR and the description of a typical system that is able to retrieve the music information of an unknown excerpt of a music title. Afterward, we continue with the implementation of a system that uses the spectrogram derived by the Short Time Fourier Transform, as the primary tool for the creation of its Audio Fingerprint model. Besides that, the implementation of a variation of this model is implemented, that replaces STFT, with the Constant-Q Transform. This transform induces some advantages in music applications, since the final spectrogram it produces is a direct mapping to the musical notes used in Western Music. Moreover, a technique for reducing the database size is tested, that involves the creation of the parameter "significance" of the database keys, which are the elements of the Audio Fingerprint itself. Finally, the performance of the implemented systems is tested, in the form of several experimental cases, that emulate different distortion cases of the unknown music excerpt input.

Περιεχόμενα

Κατάλογος Σχημάτων	vii
Κατάλογος Πινάκων	viii
1 Εισαγωγή	1
1.1 Πρόλογος	1
1.2 Ορισμός του Προβλήματος	1
1.3 Στόχοι της Εργασίας	1
1.4 Διάρθρωση της Διπλωματικής Εργασίας	2
2 Κύρια Χαρακτηριστικά Προβλήματος	3
2.1 Βασικές Αρχές	3
2.2 Κύριες Ιδιότητες Συστημάτων με Ακουστικό Αποτύπωμα	4
2.3 Σενάρια Εφαρμογών	5
2.4 Γενική Δομή Συστήματος	6
2.5 Εποπτική Περιγραφή Προσεγγίσεων	8
2.6 Σύστημα Δεικτών	9
2.7 Κλασσικές Παραμορφώσεις	10
3 Μαθηματικά Εργαλεία Ψηφιακής Επεξεργασίας Σήματος	12
3.1 Μετασχηματισμός Φουριέ Βραχύ Χρόνου	12
3.1.1 Μαθηματικός Ορισμός Μετασχηματισμού Φουριέ Βραχύ Χρόνου .	12
3.1.2 Πρακτικός Υπολογισμός Μετασχηματισμού Φουριέ Βραχύ Χρόνου	13
3.1.3 Σημασία της Ακολουθίας Παραθύρωσης	14
3.2 Φασματογράφημα Μετασχηματισμού Φουριέ Βραχύ Χρόνου	15

3.3	Μετασχηματισμός Σταθερού-Q	16
3.3.1	Μαθηματική Μοντελοποίηση Σήματος	18
3.3.2	Αποδοτικός Υπολογισμός Μετασχηματισμού Σταθερού-Q	19
4	Η μέθοδος του Shazam	24
4.1	Τι είναι το Shazam	24
4.2	Αναπαράσταση Σήματος	24
4.3	Δημιουργία Κλειδιών	26
4.4	Μηχανισμός Αναζήτησης	26
5	Τλοποίηση	30
5.1	Τλοποίηση του αλγορίθμου Shazam	30
5.1.1	Εξαγωγή τοπικών μεγίστων από το Φασματογράφημα	30
5.1.2	Εξαγωγή και κωδικοποίηση ζευγαριών κορυφών	30
5.1.3	Αποθήκευση των κλειδιών στην βάση δεδομένων	31
5.1.4	Διαχείριση απαντήσεων των ερωτημάτων της βάσης δεδομένων	32
5.1.5	Σύντηξη των τοπικών αποφάσεων	33
5.1.6	Επισκόπηση Αλγορίθμου	34
5.2	Βελτίωση του Αλγορίθμου Shazam	35
5.2.1	Βελτίωση Ευελιξίας στην Μετατόπιση Τονικότητας με την χρήση του CQT	35
5.2.2	Μείωση του Όγκου Κλειδιών στη Βάση Δεδομένων - Σημαντικότητα Κλειδιού	39
6	Πειραματική Διαδικασία	41
6.1	Περιγραφή πλαισίου πειραμάτων - Καθορισμός χριτηρίων απόδοσης	41
6.2	Περιγραφή του Audio Degradition Toolbox	41

6.3 Πειραματική Διαδικασία Αλγορίθμου Shazam	42
6.3.1 Απόδοση αλγορίθμου με καθαρό σήμα	42
6.3.2 Απόδοση αλγορίθμου με θόρυβο	44
6.3.3 Απόδοση αλγορίθμου με Συμπίεση Δυναμικού Εύρους	47
6.3.4 Απόδοση αλγορίθμου με τονική μετατόπιση/χρονική παραμόρφωση	48
6.3.5 Απόδοση Αλγορίθμου σε εξομοίωση διαφορετικών πραγματικών καταστάσεων	51
6.4 Πειραματική Διαδικασία Βελτίωσης Αλγορίθμου με χρήση CQT	51
6.4.1 Απόδοση Βελτίωσης Αλγορίθμου με CQT για καθαρό σήμα	51
6.4.2 Απόδοση Βελτίωσης Αλγορίθμου με CQT με τονική μετατόπιση/χρονική παραμόρφωση	52
6.4.3 Απόδοση Βελτίωσης Αλγορίθμου με CQT σε εξομοίωση διαφορετικών πραγματικών καταστάσεων	53
6.5 Αξιολόγηση της μεθόδου σημαντικότητας των κλειδιών της βάσης δεδομένων	53
7 Επίλογος	57
7.1 Ανακεφαλαίωση	57
7.2 Μελλοντικές επεκτάσεις	58
Βιβλιογραφία	60

Κατάλογος Σχημάτων

2.1	Δομή συστήματος αναγνώρισης μουσικού τίτλου με ακουστικό αποτύπωμα[1].	3
2.2	Τυπική Δομή Συστήματος Απόσπασης του Ακουστικού Αποτύπωματος [3].	8
3.1	Εκτίμηση συχνοτικού περιεχομένου καιθαρού ημιτονικού σήματος, (α) ορθογώνια παραθύρωση, (β) παραθύρωση με παράθυρο Kaiser[6].	14
3.2	(a) Κυματομορφή Ομιλίας. (b) Φασματογράφημα με πλατιές φασματικές μπάντες. (c) Φασματογράφημα με στενές φασματικές μπάντες. [5]	16
3.3	Γραφικές Διαφορές STFT με ένα τυπικό μετασχηματισμό κυμάτιων.	17
3.4	Αντιστοιχία συχνοτήτων - μουσικών νοτών ανά οκτάβα (Hz).	17
3.5	Κάποιοι από τους χρονικούς πυρήνες με τους αντίστοιχους φασματικούς τους, ζεχινώντας από υψηλότερη προς χαμηλότερη συχνότητα. Είναι φανερό πως οι φασματικοί πυρήνες δεν έχουν το ίδιο μήκος.	20
3.6	Τα σχετικά μήκη των ακολουθιών παραθύρωσης N_k του μετασχηματισμού.	21
3.7	Επισκόπηση του συστήματος υπολογισμού του CQT ανά οκτάβα. Το $G(f)$ συμβολίζει το χαμηλοπερατό φίλτρο, ενώ το $\downarrow 2$ συμβολίζει την υποδειγματοληψία ανά παράγοντα 2.	22
3.8	Οι μετατοπισμένοι πυρήνες με τους αντίστοιχους φασματικούς πυρήνες τους. Στο διάγραμμα εμφανίζεται μόνο το πραγματικό μέρος τους.	23
4.1	Αναπαράσταση σήματος 5 δευτερολέπτων με το φασματογράφημα, και τα τοπικά μέγιστα που προκύπτουν	25
4.2	Μηχανισμός δημιουργίας κλειδιών.	27
4.3	Ιστογράμματα σωστού και λάθος κομματιών.	28
4.4	Διαγράμματα χρονικής αντιστοίχησης κλειδιών.	29
5.1	Φασματογράφημα του Μετασχηματισμού Σταθερού-Q.	36
5.2	Απόσπαση τοπικών μεγίστων από το φασματογράφημα του Μετασχηματισμού Σταθερού-Q.	37
6.1	Αποτελέσματα αναζήτησης για σήματα με διαφορετικές τιμές SNR.	45
6.2	Παράδειγμα σύγκρισης φασματογραφήματος με προσθήκη λευκού θορύβου.	45

6.3	Αποτελέσματα αναζήτησης για σήματα με διαφορετικές τιμές SNR.	46
6.4	Παράδειγμα σύγχρισης φασματογραφήματος με προσθήκη θορύβου παρασκηνίου με $SNR = 1$	46
6.5	Παράδειγμα συγχρίσης φασματογραφήματος με προσθήκη συμπίεσης δυναμικού εύρους.	47
6.6	Αποτελέσματα αναζήτησης για σήματα με διαφορετικές τιμές συντελεστή τονικής μετατόπισης.	48
6.7	Παράδειγμα σύγχρισης φασματογραφήματος με τονική μετατόπιση συντελεστή +4%.	49
6.8	Αποτελέσματα αναζήτησης για σήματα με διαφορετικές τιμές συντελεστή τονικής μετατόπισης.	52
6.9	Ταξινομημένη παρουσίαση της κατανομής της σημαντικότητας των χλειδιών στην βάση δεδομένων (αύξουσα ταξινόμηση).	54

Κατάλογος Πινάκων

2.1	Μαθηματική αναπαράσταση των συνηθέστερων ηχητικών παραμορφώσεων.	11
5.1	Πίνακας παραμέτρων αλγορίθμου Shazam	34
5.2	Πίνακας παραμέτρων αλγορίθμου με CQT	39
6.1	Αποτελέσματα αναζήτησης με καθαρό σήμα για το αποθηκευμένο μέρος των κομματιών αναφοράς.	43
6.2	Αποτελέσματα αναζήτησης με καθαρό σήμα για κομμάτι του μη αποθηκευμένου μέρους των κομματιών αναφοράς.	44
6.3	Αποτελέσματα αναζήτησης με σήμα που έχει υποστεί συμπίεση δυναμικού εύρους, για κομμάτι του μη αποθηκευμένου μέρους των κομματιών αναφοράς.	47
6.4	Αντιστοιχία συντελεστή τονικής μετατόπισης κ με την κωδικοποιημένη διακριτικότητα των χρονικών διαφορών $t_2 - t_1$	50
6.5	Μέγιστη τιμή σωστής αντιστοιχισης συχνοτήτων για διαφορετικές τιμές το συντελεστή τονικής μετατόπισης κ.	50

6.6 Αποτελέσματα αναζήτησης σε εξομοίωση διαφορετικών πραγματικών καταστάσεων	51
6.7 Αποτελέσματα αναζήτησης με CQT για είσοδο καθαρού σήματος.	51
6.8 Αποτελέσματα αναζήτησης σε εξομοίωση διαφορετικών πραγματικών καταστάσεων.	53
6.9 Το ποσοστό των μοναδικών κλειδιών της βάσης που ικανοποιούν την συνθήκη σημαντικότητας κλειδιού για διαφορετικές τιμές κατωφλίωσης T_{prune} , μαζί με το αντίστοιχο ποσοστό των συνολικών τιμών που περιέχονται στην βάση δεδομένων. Είναι εμφανές πως αν και τα κλειδιά που διαγράφονται είναι πολύ λίγα, περιέχουν σημαντικό ποσοστό της συνολικής πληροφορίας της βάσης δεδομένων.	54
6.10 Αποτελέσματα Μέσου Όρου χρόνου αναζήτησης με CQT σε βάσεις διαφορετικού κατωφλιού T_{prune}	55
6.11 Αποτελέσματα αναζήτησης με CQT για είσοδο καθαρού σήματος με βάσεις διαφορετικού κατωφλιού T_{prune}	55
6.12 Αποτελέσματα αναζήτησης με CQT για είσοδο βαριά παραμορφωμένου σήματος με βάσεις διαφορετικού κατωφλιού T_{prune}	55

1 Εισαγωγή

1.1 Πρόλογος

Η αναγνώριση μουσικών τίτλων είναι ένα πεδίο το οποίο έχει απασχολήσει την επιστημονική κοινότητα τα τελευταία χρόνια, καθώς αποτελεί σημείο κλειδί σε αρκετές εφαρμογές[1], με τις δύο σημαντικότερες να είναι η αναγνώριση ενός αποσπάσματος μουσικού κομματιού μέσω κινητού τηλεφώνου και η παρακολούθηση περιεχομένου με πνευματικά δικαιώματα.

Μια από τις πηγές της δυσκολίας του προβλήματος προκύπτει από την ίδια την φύση του ήχου, αφού αποτελεί ένα πολυσύνθετο φαινόμενο. Η περιγραφή του αποτελεί πρόβλημα υψηλών διαστάσεων, ενώ μουσικά παρόμοια σήματα διαφέρουν σημαντικά μεταξύ τους από την μαθηματικά αυστηρή πλευρά του υπολογιστή.

1.2 Ορισμός του Προβλήματος

Στην παρούσα διπλωματική εργασία διερευνούνται τρόποι συστηματικής περιγραφής ηχητικών σημάτων από μουσικούς τίτλους, που βασίζονται στην δημιουργία ενός μοναδικού ακουστικού αποτυπώματος του σήματος.

Τα συστήματα αναγνώρισης μουσικών τίτλων με ακουστικό αποτύπωμα αποσπούν μια συμπυκνωμένη και συμπαγής μορφή από ένα μουσικό κομμάτι το οποίο στη συνέχεια αποθηκεύεται σε μια βάση δεδομένων. Χρησιμοποιώντας μοντέλα ακουστικών αποτυπωμάτων και αλγόριθμους αναζήτησης και ταύτισης, παραμορφωμένες εκδόσεις του μουσικού τίτλου είναι σε θέση να αναγνωριστούν επιτυχημένα.

1.3 Στόχοι της Εργασίας

Στόχος της εργασίας είναι αρχικά να γίνει μια εισαγωγή στην επιστημονική βιβλιογραφία που υπάρχει γύρω από αυτό το θέμα και να παρουσιαστούν διαφορετικές προσεγγίσεις που έχουν προταθεί για την λύση του προβλήματος της αναγνώρισης μουσικής.

Στην συνέχεια, στόχος είναι να γίνει αναλυτική περιγραφή και υλοποίηση πάνω σε συγκεκριμένο αλγόριθμο αναγνώρισης μουσικών τίτλων, που χρησιμοποιεί ως βασικό μαθηματικό εργαλείο παραγωγής του ακουστικού αποτυπώματος το φασματογράφημα του Μετασχηματισμού Φουρέ Σύντομου Χρόνου. Τέλος, υλοποιείται και μια παραλλαγή αυτού, στην οποία το φασματογράφημα προκύπτει από τον Μετασχηματισμό Σταθερού-Q.

1.4 Διάρθρωση της Διπλωματικής Εργασίας

Η εργασία είναι οργανωμένη σε επτά κεφάλαια:

Στο κεφάλαιο 2 γίνεται περιγραφή των κύριων χαρακτηριστικών του προβλήματος, καθώς και αναφορά στις κύριες μεθόδους προσέγγισης στον ορισμό του ακουστικού αποτυπώματος. Επίσης, παρουσιάζονται τα βασικά επιμέρους στοιχεία ενός συστήματος που έχει την δυνατότητα να αναγνωρίζει ένα άγνωστο απόσπασμα μουσικής και οι κλασσικές ηχητικές παραμορφώσεις τις οποίες και το ακουστικό αποτύπωμα πρέπει να είναι σε θέση να ξεπερνάει. Στο κεφάλαιο 3, περιγράφονται οι δύο μετασχηματισμοί που χρησιμοποιούνται στον αλγόριθμο που υλοποιείται, δηλαδή τον Μετασχηματισμό Φουριέ Σύντομου Χρόνου και τον Μετασχηματισμό Σταθερού-Q. Στο κεφάλαιο 4, γίνεται περιγραφή του αλγορίθμου του Shazam που βασίζεται σε μια από τις ελάχιστες επιστημονικές δημοσιεύσεις της ομώνυμης εταιρίας. Στο κεφάλαιο 5, περιγράφεται η υλοποίηση του αλγορίθμου καθώς και μια παραλλαγή αυτού, με στόχο την βελτίωση στην απόδοση του αλγορίθμου στην παραμόρφωση της τονικής μετατόπισης. Στο κεφάλαιο 6, περιγράφεται η πειραματική διαδικασία των δυο αλγόριθμων και η απόδοση τους σε μια σειρά διαφορετικών παραμορφώσεων. Τέλος, το κεφάλαιο 7 αποτελείται από τα γενικά συμπεράσματα της έρευνας και γίνεται αναφορά σε μια σημαντική κατηγορία λύσης του προβλήματος της αναγνώρισης μουσικών σημάτων που δεν υλοποιείται στην παρούσα διπλωματική εργασία.

2 Κύρια Χαρακτηριστικά Προβλήματος

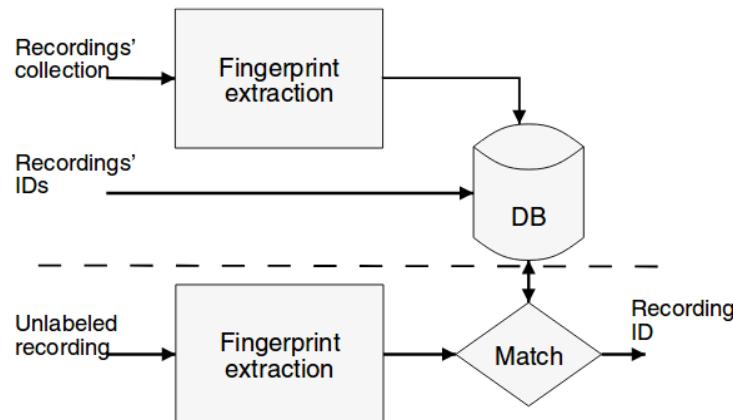
2.1 Βασικές Αρχές

Τι πάρχουν δυο μεγάλες κατηγορίες για την λύση του προβλήματος αυτού: η προσθήκη υδατογραφήματος ή η δημιουργία ενός ακουστικού αποτυπώματος στα ηχητικά σήματα:

Η τεχνική της υδατογράφησης περιλαμβάνει την προσθήκη κάποιων δυαδικών ψηφίων πληροφορίας στο ηχητικό σήμα, τα οποία επιτρέπουν την ανεύρεση των μεταδεδομένων. Το κύριο ζήτημα που προκύπτει είναι πως η προσθήκη της νέας πληροφορίας δεν πρέπει να αλλοιώνει την ακουστική ποιότητα του σήματος.

Η τεχνική του ακουστικού αποτυπώματος περιλαμβάνει την εκμαίευση ενός μοναδικού αποτυπώματος του σήματος, το οποίο και χρησιμοποιείται για την αναγνώριση του σήματος και την μετέπειτα συσχέτιση του με τα μεταδεδομένα. Επίσης, αξίζει να αναφερθεί πως οι δυο αυτές τεχνικές μπορούν και να συνδυαστούν, με το ακουστικό αποτύπωμα να χρησιμοποιείται ως κομμάτι ελέγχου σε μια διαδικασία που προσθέτει υδατογραφήματα σε ηχητικά σήματα.

Τα συστήματα αναγνώρισης ήχου με χρήση ακουστικού αποτυπώματος λειτουργούν με τον εξής τρόπο. Περιέχουν μια βάση δεδομένων με ηχητικά αρχεία αναφοράς, τα οποία είναι και το σύνολο των σημάτων που το σύστημα μπορεί να αναγνωρίσει. Η βασική αρχή είναι η αλλαγή του προβλήματος από ταίριασμα ηχητικών σημάτων σε ταίριασμα αποτυπωμάτων. Η βάση δεδομένων των ηχητικών αρχείων μετατρέπεται σε βάση δεδομένων των αντίστοιχων ακουστικών αποτυπωμάτων, διαδικασία που αποτελεί και το στάδιο της εκπαίδευσης του συστήματος. Επομένως, όταν ζητείται η αναγνώριση ενός άγνωστου ηχητικού αποσπάσματος, το σύστημα υπολογίζει το αποτύπωμα του αποσπάσματος και γίνεται αναζήτηση στην βάση δεδομένων για το καλύτερο ταίριασμα (Σχήμα 1).



Σχήμα 2.1: Δομή συστήματος αναγνώρισης μουσικού τίτλου με ακουστικό αποτύπωμα[1].

2.2 Κύριες Ιδιότητες Συστημάτων με Ακουστικό Αποτύπωμα

Ανάλογα με την φύση της εφαρμογής, επιδιώκεται η ικανοποίηση κάποιων ιδιοτήτων από το σύστημα και ταυτόχρονα προσδιορίζεται και ιεραρχικά η σημαντικότητα τους:

Ευελιξία

Το αποτύπωμα πρέπει να είναι σε θέση να αναγνωρίζει το σήμα αναφοράς με μη αμφισβητήσιμο τρόπο. Ταυτόχρονα όμως, και να πληροί την προϋπόθεση πως ενώ το αποτύπωμα είναι μοναδικό για κάθε αναφορά, μπορεί και να μένει αμετάβλητο για αρκετές τροποποιήσεις που ίσως έχει υποστεί το σήμα του αποσπάσματος προς αναγνώριση. Σε αυτό το σημείο, διαχρίνονται δύο περιπτώσεις που οδηγούν σε αρκετά διαφορετικές στρατηγικές ως προς τον ορισμό των στοιχείων του αποτυπώματος.

Στην πρώτη περίπτωση, θεωρούμε πως το σήμα προς αναγνώριση είναι ακριβές αντίγραφο, ή αλλιώς ισοδύναμο κάποιου σήματος αναφοράς που περιέχεται στην βάση δεδομένων. Όμως, θεωρούμε πως το απόσπασμα μπορεί να έχει υποστεί αρκετές παραμορφώσεις κατά την μετάδοση του ήχου (πχ. επεξεργασία μουσικών κομματιών κατά την αναπαραγωγή τους σε ραδιοφωνικούς σταθμούς). Η περίπτωση αυτή ονομάζεται ακριβές ταίριασμα (exact matching).

Στη δεύτερη περίπτωση, θεωρούμε πως το απόσπασμα δεν υπάρχει επακριβώς ίδιο στη βάση δεδομένων. Αντίθετα, η βάση δεδομένων περιέχει μια αναφορά που είναι πολύ κοντά με το σήμα αναζήτησης στο μουσικό επίπεδο. Για παράδειγμα, τα σήματα της ηχογράφησης ενός κομματιού σε στούντιο, με μια ζωντανή εκτέλεση του ή ακόμα και η επανάληψη μιας ηχογράφησης του ίδιου κομματιού, θεωρούνται παρόμοια σήματα, υπό το πρίσμα της μουσικής τους ομοιότητας. Όμως, αν και ένας άνθρωπος αναγνωρίζει πολύ εύκολα πως πρόκειται για το ίδιο κομμάτι, πολύ μικρές αλλαγές (μικρές καλυστερήσεις, αλλαγές στη μελωδία, διαφορετική εκτέλεση στα φωνητικά, κ.α) κάνουν τα σήματα που προκύπτουν πολύ διαφορετικά μεταξύ τους για τον υπολογιστή. Η κατηγορία αναγνώρισης τέτοιων σημάτων ονομάζεται ταίριασμα κατά προσέγγιση (approximate matching) και μπορεί να θεωρηθεί ως μια αυστηρή επέκταση του ακριβούς ταίριασματος.

Επεκτασιμότητα

Το επόμενο σημαντικό ζήτημα που προκύπτει είναι η επεκτασιμότητα της μεθόδου. Στα πλαίσια μια εμπορικής εφαρμογής, οι βάσεις δεδομένων περιέχουν εκατοντάδες χιλιάδες μουσικούς τίτλους, και το μέγεθος τους αυξάνεται συνεχώς. Η ικανότητα του συστήματος για αποδοτική αναζήτηση σε αυτόν τον μεγάλο όγκο δεδομένων είναι ίσως το πιο κρίσιμο χαρακτηριστικό του.

Για αυτόν τον λόγο, συνήθως προτείνεται ένα μοντέλο ακουστικού αποτυπώματος που να μπορεί να ενσωματωθεί μέσα σε ένα σύστημα δεικτών, καθώς η χρήση δεικτών είναι μια κλασσική προσέγγιση για την υλοποίηση αποδοτικών μοντέλων αναζήτησης.

Διακριτικότητα

Η χρονική διάρκεια της εισόδου του συστήματος, δηλαδή πόσα δευτερόλεπτα πρέπει

να είναι το ηχητικό σήμα προς αναγνώριση.

Ακρίβεια

Η μελέτη και ο διαχωρισμός των αποτελεσμάτων του συστήματος σε σωστές αναγνωρίσεις (true positive), λάθος αναγνωρίσεις (false positive) και άγνωστα αποτελέσματα (μη επίτευξη αναγνώρισης).

Αξιοπιστία

Η δυνατότητα εκτίμησης του κατά πόσο ένα κομμάτι ανήκει ή όχι στη βάση δεδομένων του συστήματος. Η ιδιότητα αυτή είναι σημαντική για εφαρμογές που η μη-αναγνώριση κομματιών ως έξοδος είναι προτιμότερη από μια πιθανώς λάθος αναγνώριση.

Ευστροφία

Η ικανότητα της μεθόδου ακουστικού αποτυπώματος και της βάσης δεδομένων του συστήματος να επαναχρησιμοποιηθεί σε διαφορετικές εφαρμογές, καιώς και η ικανότητα του να αναγνωρίζει ηχητικά αρχεία ανεξαρτήτως του τύπου αρχείου.

2.3 Σενάρια Εφαρμογών

Σε αυτό το σημείο θα αναφερθούν οι κύριες εφαρμογές που χρησιμοποιούν συστήματα βασισμένα στο ηχητικό αποτύπωμα:

Αυτόματη Επίβλεψη Ροών Αναμετάδοσης

Η αυτόματη επίβλεψη ροών αναμετάδοσης ήχου είναι από τις πιο γνωστές εφαρμογές για συστήματα ακουστικών αποτυπωμάτων. Συγκεκριμένα, αναφέρεται στην αυτόματη αναγνώριση ηχητικών σημάτων σε ροές αναμετάδοσης όπως το ραδιόφωνο, η τηλεόραση, ή διαδικτυακοί ραδιοφωνικοί σταθμοί. Η αναγνώριση μπορεί να έχει στόχο να επιστρέψει τα μεταδεδομένα από μουσικούς τίτλους, διαφημιστικά μηνύματα, ή εκπομπές.

Ένα τέτοιο σύστημα μεγάλης κλίμακας αποτελείται από ένα κεντρικό εξυπηρετητή που περιέχει την βάση δεδομένων, και τους διάφορους σταθμούς-πελάτες του συστήματος, οι οποίοι στέλνουν συνεχώς τα αποτυπώματα στον κεντρικό εξυπηρετητή και αυτός με τη σειρά του επιστρέφει τα μεταδεδομένα στους πελάτες. Σε αυτή την περίπτωση, τέτοιου είδους συστήματα χρησιμοποιούνται ευρέως για την εξασφάλιση της κατοχής δικαιωμάτων χρήσης του ηχητικού περιεχομένου στους σταθμούς αναμετάδοσης, ή την εξασφάλιση των συμφωνημένων όρων κατά την προβολή των διαφημιστικών μηνυμάτων από τις διαφημιστικές εταιρίες.

Φιλτράρισμα σε πλατφόρμες διαμοιρασμού ηχητικού περιεχομένου

Οι πλατφόρμες που περιέχουν συστήματα μετάδοσης ηχητικού περιεχομένου είναι νομικά υποχρεωμένες να χρησιμοποιούν συστήματα που να απαγορεύουν τον διαμοιρασμό περιεχομένου που δεν έχει δικαιώματα δημόσιας χρήσης από τον οποιοδήποτε. Σε αυτή την κατηγορία εμπίπτουν πλατφόρμες για μεταφορά αρχείων μέσω Peer2Peer, ή πλατφόρμες αναμετάδοσης πολυμέσων όπως Youtube, Spotify.

Αναγνώριση Μουσικών Κομματιών μέσω Smartphone

Ίσως η πιο γνωστή εφαρμογή από πλευράς μαζικής κατανάλωσης, στην οποία ο χρήστης ηχογραφεί μέσω εφαρμογής με το κινητό του ένα απόσπασμα από ένα μουσικό κομμάτι, το οποίο στέλνεται προς αναγνώριση στην κεντρική βάση δεδομένων και επιστρέφονται ως έξοδος οι πληροφορίες του κομματιού. Τους τελευταίους μήνες, σε ορισμένα μοντέλα η εφαρμογή αυτή είναι πλήρως ενσωματωμένη, με την αναγνώριση μουσικής να γίνεται συνεχώς στο παρασκήνιο.

Αυτόματη Οργάνωση Μουσικών Βιβλιοθηκών

Τις περισσότερες φορές, οι προσωπικές μουσικές βιβλιοθήκες έχουν αρκετά κενά στα μεταδεδομένα των μουσικών κομματιών. Υπάρχουν αρκετές εφαρμογές που αναγνωρίζουν τους μουσικούς τίτλους και συμπληρώνουν αυτόματα τα κενά.

2.4 Γενική Δομή Συστήματος

Το πεδίο ανάπτυξης εφαρμογών που βασίζονται στην χρήση ακουστικού αποτυπώματος είναι αρκετά ενεργό τα τελευταία χρόνια, και λόγω αυτού έχει αναπτυχθεί αρκετή έρευνα για την επίλυση του προβλήματος, ενώ ταυτόχρονα έχει τραβήξει το ενδιαφέρον και σε μεγάλες βιομηχανίες (Shazam[7], Philips[10], Google[9]).

Αν και υπάρχουν αρκετά διαφορετικές προσεγγίσεις, είναι δυνατόν να ορισθεί μια κοινή βάση των αλγορίθμων που έχουν προταθεί[3]. Όπως προαναφέρθηκε, ένα σύστημα ακουστικού αποτυπώματος αποτελείται από δύο επιμέρους διαδικασίες. Αρχικά, ξεκινά με την διαδικασία μάθησης μιας βάσης δεδομένων με τα κομμάτια αναφοράς σύμφωνα με το μοντέλο που επιλέχθηκε.

Έπειτα, χρειάζεται ο αλγόριθμος που να είναι σε θέση να αναγνωρίζει άγνωστα αποσπάσματα με το ίδιο μοντέλο. Ο υπολογισμός του ακουστικού αποτυπώματος από τα αρχικά αρχεία ήχου μπορεί να αναλυθεί στην εξής μεθοδολογία:

Προ-επεξεργασία

Αρχικά, τα σήματα περνάνε από μια διαδικασία προ-επεξεργασίας, ώστε να βρίσκονται στην ίδια μορφή. Σημεία της προ-επεξεργασίας είναι η εξασφάλιση της ομοιόμορφης κωδικοποίησης των αρχείων, η μετατροπή από stereo σε mono, η μετατροπή του ρυθμού διειγματοληψίας, ενώ σε ορισμένες εφαρμογές το σήμα μετατρέπεται έτσι ώστε να εξομοιώνει τον τύπο του καναλιού μετάδοσης ή την κωδικοποίηση που ακολουθείται (περιορισμός του εύρους ζώνης του σήματος, GSM κωδικοποίηση).

Πλαισίωση και επικάλυψη

Το επόμενο βήμα είναι ο καθορισμός του ρυθμού πλαισίωσης του σήματος, δηλαδή η διαίρεση του σε σταθερού μικρού μεγέθους τμήματα ώστε το σήμα να μπορεί να θεωρηθεί στάσιμο, και η χρήση κατάλληλης συνάρτησης παραθύρωσης για να

εξομαλυνθούν οι ασυνέχειες στην αρχή και το τέλος του κάθε μπλοκ. Σημαντική είναι και η ύπαρξη ενός ποσοστού επικάλυψης μεταξύ των διαδοχικών μπλοκς του σήματος, για να αυξηθεί η ευελιξία σε πιθανές μικρές χρονικές μετατοπίσεις του σήματος.

Μετασχηματισμός της αναπαράστασης του σήματος

Οι περισσότερες τεχνικές αποφεύγουν την απευθείας χρήση της κυματομορφής για την εξαγωγή χαρακτηριστικών, καθώς η μορφή της είναι πάρα πολύ ευαίσθητη όταν το σήμα παραμορφώνεται. Για αυτόν τον λόγο, προτιμούνται αναπαραστάσεις όπως η ενέργεια του σήματος, ο Ταχύς Μετασχηματισμός Fourier, (FFT), ο διακριτός Μετασχηματισμός Συνημιτόνου DCT, μετασχηματισμοί κυματίων (wavelets), κ.α..

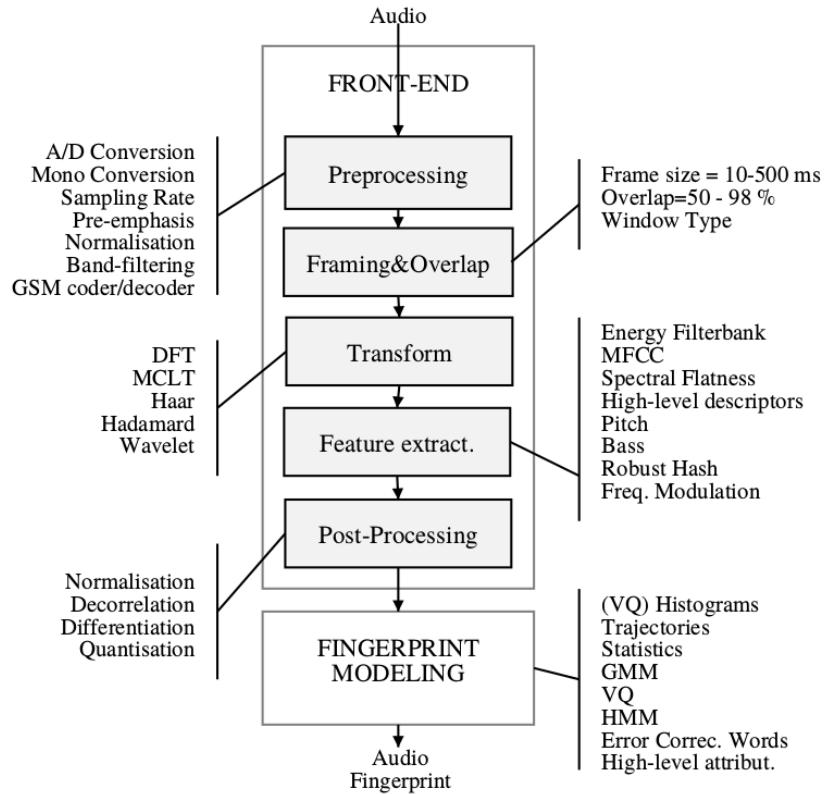
Εξαγωγή Χαρακτηριστικών

Σε αυτό το σημείο, σκοπός είναι η εξαγωγή των χαρακτηριστικών από την νέα αναπαράσταση του σήματος, με στόχο την μείωση των διαστάσεων της αναπαράστασης και ταυτόχρονα την αύξηση της ευελιξίας της στην περίπτωση ύπαρξης παραμορφώσεων. Το σύνολο των χαρακτηριστικών συγκεντρώνεται σε μια συμπαγή και συμπυκνωμένη μορφή και αποτελεί το τελικό ακουστικό αποτύπωμα, το όποιο θα πρέπει και να είναι επαρκές για την αναγνώριση του σήματος.

Αναζήτηση

Το επόμενο κομβικό σημείο του συστήματος είναι ένας αποδοτικός αλγόριθμος αναζήτησης στην βάση των ακουστικών αποτυπώμάτων κατά την διαδικασία αναγνώρισης ενός άγνωστου αποσπάσματος. Όπως ειπώθηκε, υπάρχουν δύο μεγάλες κατηγορίες που βασίζονται στην παραδοχή ή όχι της υπόθεσης πως τα χαρακτηριστικά του αποτυπώματος είναι τόσο ευέλικτα ώστε να διατηρούνται στο απόλυτο υπό την παρουσία παραμορφώσεων.

- Στην πρώτη περίπτωση, η αναζήτηση είναι αρκετά απλή, καθώς αρκεί η αναζήτηση σε μια βάση δεδομένων με ζευγάρια δεικτών-τιμών με μοναδική απαίτηση την ακριβής ταύτιση των δεικτών των καταχωρήσεων (hash-table, B-tree). Για να ισχύει όμως η παραπάνω υπόθεση, οι αλγόριθμοι αυτοί χρησιμοποιούν τοπικά χαρακτηριστικά σε μικρή χρονική έκταση του σήματος, καθώς όσο πιο διευρυμένο στον χρόνο είναι ένα χαρακτηριστικό, τόσο πιο πιθανό είναι να επηρεάζεται από αλλαγές λόγω παραμορφώσεων.
- Στην δεύτερη περίπτωση, οι αλγόριθμοι αναζήτησης ψεωρούν πως τα χαρακτηριστικά μπορούν να είναι παραμορφωμένα, και για αυτό χρησιμοποιούν στρατηγικές αναζήτησης ταύτισης κατά προσέγγισης (προσεγγιστικό ταίριασμα αλφαριθμητικών, αλγόριθμος K-κοντινότερων γειτόνων, κ.α.). Όμως, η αναζήτηση σε αυτού του τύπου συστήματα είναι αρκετά πιο περίπλοκη. Το πλεονέκτημα είναι πως μπορούν να επιλέγονται χαρακτηριστικά μεγαλύτερης χρονικής εμβέλειας του σήματος, γεγονός που επιτρέπει την περιγραφή του σήματος σε υψηλότερο επίπεδο, καθώς τα χαρακτηριστικά μπορούν να



Σχήμα 2.2: Τυπική Δομή Συστήματος Απόσπασης του Ακουστικού Αποτυπώματος [3].

μεταφράζονται σε πιο μουσικά χαρακτηριστικά (ακολουθία νοτών, ρυθμός, κ.α.), ενώ οι τοπικές παραμορφώσεις απορροφούνται.

2.5 Εποπτική Περιγραφή Προσεγγίσεων

Οι προσεγγίσεις λύσης του προβλήματος της ηχητικής αναγνώρισης μέσω ηχητικού αποτυπώματος μπορούν να χωριστούν και σε 4 κατηγορίες με κριτήριο την τεχνική που χρησιμοποιούν για την δημιουργία του αποτυπώματος:

- Το ηχητικό αποτύπωμα εξάγεται κατευθείαν από την κυματομορφή του σήματος, δηλαδή την χρονική του αναπαράσταση. Στο [17], αρχικά υπολογίζεται μια εκτίμηση της περιοδικότητας του σήματος, και η πορεία που ακολουθεί στον χρόνο ορίζεται ως το τελικό αποτύπωμα.
- Ως βάση του αποτυπώματος χρησιμοποιείται το *STFT* φασματογράφημα του σήματος, το οποίο στη συνέχεια φιλτράρεται με διάφορες τεχνικές ώστε να καταλήξει σε μια συμπαγή και συμπιεσμένη μορφή που να περιέχει τοπικά χρονικά σύντομης διάρκειας

χαρακτηριστικά του σήματος [7][8][10]. Συνήθως, έπειτα χρησιμοποιούνται τεχνικές αναζήτησης που να χρησιμοποιούν ένα σύστημα δεικτών, περιορίζοντας έτσι τους αλγορίθμους αυτούς στην κατηγορία του ακριβές ταιριάσματος.

- Χρησιμοποιούνται μέθοδοι ώστε να εξαχθούν χαρακτηριστικά μεγάλης χρονικής διάρκειας από το φασματογράφημα. Στα [15][16], υπολογίζεται ο μετασχηματισμός Φουριέ για κάθε συχνοτική μπάντα της κλίμακας *Bark*, και το αποτύπωμα αποτελείται από τους συντελεστές για κάθε συχνοτική μπάντα. Για την αναζήτηση σε αυτή την κατηγορία, χρησιμοποιούνται τεχνικές ταιριάσματος κατά προσέγγιση. Τέλος, αξίζει να αναφερθεί, πως λόγω της μεγάλης χρονικής διάρκειας των χαρακτηριστικών, το μέγεθος του τελικού αποτυπώματος είναι αρκετά συμπαγές με αποτέλεσμα η διαδικασία της αναζήτησης να είναι χαμηλού επιπέδου πολυπλοκότητας.
- Εδώ το ηχητικό σήμα μετατρέπεται σε μια ακολουθία πεπερασμένων χαρακτήρων (αλφάριθμο), που μπορεί είτε να αντιπροσωπεύει μουσικά χαρακτηριστικά του σήματος (μελωδία, ρυθμός)[14], είτε μέσω της μοντελοποίησης με την τεχνική των AudioGenes[12][13]. Έτσι, με την τελική μορφή του αποτυπώματος να είναι μια ακολουθία χαρακτήρων, η διαδικασία της αναζήτησης συνήθως γίνεται με τεχνικές ταιριάσματος κατά προσέγγιση.

2.6 Σύστημα Δεικτών

Ως μοντέλο δεικτών, ορίζεται η δημιουργία μιας λίστας ορισμένων χαρακτηριστικών μιας συλλογής αντικειμένων. Κάθε χαρακτηριστικό συσχετίζεται με μια λίστα των αντικειμένων που το περιέχουν. Ονομάζουμε τα χαρακτηριστικά ‘κλειδιά’, και για την συσχέτιση μεταξύ κλειδιών και αντικειμένων λέμε, ‘το κλειδί δείχνει προς το αντικείμενο’.

Για την ίδια συλλογή αντικειμένων, μπορούν να εφαρμοστούν πολλά μοντέλα δεικτών. Ανάλογα με την επιλεγμένη στρατηγική, η αποδοτικότητα μιας αναζήτησης στην λίστα των αντικειμένων μπορεί να βελτιωθεί δραματικά, και για αυτό τον λόγο μπορούμε να μιλήσουμε για ποιότητα του δείκτη. Το κέρδος σε ταχύτητα αναζήτησης έρχεται ταυτόχρονα με το κόστος του χρόνου κατασκευής των δεικτών καθώς και τον απαιτούμενο χώρο που χρειάζεται η αποθήκευση τους. Παρόλα αυτά, η κατασκευή του μοντέλου δεικτών γίνεται πριν την αναζήτηση.

Τα πλεονεκτήματα του παραπάνω μοντέλου είναι πολύ σημαντικά για το πρόβλημα της αναγνώρισης ήχου με ακουστικά αποτυπώματα, καθώς μας επιτρέπει να κάνουμε γρήγορες αναζητήσεις και συγχρίσεις στον μεγάλο όγκο δεδομένων της βάσης του συστήματος. Η δυσκολία που έγκειται πως ως το μοντέλο των κλειδιών είναι ένα σύστημα ακριβούς σύγχρισης. Αν τα κλειδιά διαφέρουν μεταξύ τους στο ελάχιστο, τότε δεν γίνεται ταυτοποίηση. Για αυτό η αναζήτηση στην απάντηση στο ποια ποσότητα μπορεί να οριστεί ως κλειδί στο μοντέλο του ακουστικού αποτυπώματος είναι κεντρικής σημασίας.

2.7 Κλασσικές Παραμορφώσεις

Όπως ειπώθηκε το ακουστικό αποτύπωμα πρέπει να είναι ευέλικτο στις παραμορφώσεις που μπορεί να υποστεί ένα ηχητικό σήμα σε ένα κλασσικό κανάλι μετάδοσης. Αυτές περιλαμβάνουν την επεξεργασία του ήχου κατά την αναπαραγωγή της μουσικής που μπορεί να προσθέτει το ραδιοφωνικό κανάλι ή ο καλλιτέχνης, παραμορφώσεις του μέσου μετάδοσης, αλλά και τις παραμορφώσεις που προκύπτουν όταν ο χρήστης του συστήματος ηχογραφεί το απόσπασμα.

Θα προσπαθήσουμε να δούμε τις κύριες παραμορφώσεις που χρησιμοποιούνται από τους ραδιοφωνικούς σταθμούς καθώς είναι αρκετά αντιπροσωπευτικές για τις περισσότερες χρήσεις του συστήματος, και το ηχητικό απόσπασμα πρέπει να βρίσκεται σε θέση να τις αντιμετωπίζει ικανοποιητικά.

Θεωρούμε πως το ηχητικό σήμα είναι μια συνάρτηση S δύο μεταβλητών t, f , χρόνου και συχνότητας αντίστοιχα. Τα ηχητικά σήματα μπορούν να εκφραστούν τοπικά στο χρόνο, ως ένα άνθροισμα αρμονικών συναρτήσεων για τις συχνότητες που ανήκουν στο εύρος $0, f_s/2$, όπου f_s η συχνότητα δειγματοληψίας του σήματος. Η $S(t_0, f_0)$ αναπαριστά δηλαδή, το βάρος της αρμονικής συνάρτησης στην συχνότητα f_0 , όταν το σήμα μελετάται κοντά στον χρόνο t_0 , ενώ L είναι το μήκος του σήματος στον χρόνο.

$$S : \begin{array}{ccc} [0; L] \times [0; f_s/2] & \longrightarrow & C \\ (t, f) & \longmapsto & S(t, f) \end{array}$$

- Στις περισσότερες εφαρμογές της αναγνώρισης ήχου και ειδικά μουσικής, απαιτείται η ηχογράφηση ενός αποσπάσματος ως είσοδος του συστήματος. Αυτό έχει ως συνέπεια πως τα χρονικά όρια του αποσπάσματος δεν θα ταιριάζουν με τα όρια του σήματος αναφοράς με το οποίο θα πρέπει να πετύχει η ταυτοποίηση του αποσπάσματος. Έτσι, το μοντέλο του αποτυπώματος θα πρέπει να είναι ευέλικτο στην περικοπή του σήματος. Στον παρακάτω πίνακα δίνεται η έκφραση του αποσπάσματος του σήματος μήκους M , ($M < L$) που ξεκινάει από το t_0 του αρχικού σήματος.
- Επιπλέον, πριν την αναπαραγωγή των μουσικών κομματών, το σήμα περνάει από μια διαδικασία ισοστάθμισης (equalisation), η οποία δίνει αυξημένη ή μειωμένη ένταση στον ήχο σε συγκεκριμένες συχνοτικές ζώνες. Το τυπικό παράδειγμα για τα μοντέρνα είδη μουσικής αποτελείται από αύξηση έντασης στις χαμηλές συχνότητες ώστε να δοθεί έμφαση στα ρυθμικά σημεία της μουσικής (χρουστά και μπάσο).
- Είναι διαδεδομένη η χρήση συμπίεσης του δυναμικού εύρους του μουσικού τίτλου. Η μείωση, δηλαδή, της έντασης του ήχου όταν ξεπερνά ένα άνω όριο decibel που τίθεται. Οι compressors είναι ιδιαίτερα χρήσιμοι για τους ραδιοφωνικούς σταθμούς κατά την αναπαραγωγή σε θορυβώδη περιβάλλοντα, όπως στο αυτοκίνητο.
- Συνηθισμένο φαινόμενο, επίσης, αποτελεί η εφαρμογή του εφέ τονικής μετατόπισης(pitch-shifting), η οποία δίνει την αίσθηση στον ακροατή πως η μουσική

είναι πιο ζωντανή και γρήγορη. Μπορούμε να ορίσουμε το pitch-shifting ως μια επέκταση των συχνοτήτων. Συνήθως, υλοποιείται μέσω αλλαγής του ρυθμού δειγματοληψίας του σήματος, και όταν γίνεται έτσι ταυτόχρονα συνδυάζεται με επέκταση ή συμπίεση στον χρόνο. Ο παράγοντας επέκτασης στον χρόνο είναι αντιστρόφως ανάλογος με τον παράγοντα επέκτασης στις συχνότητες ($K = 1/K'$).

- Τέλος, αλασσική παραμόρφωση του ήχου αποτελεί ο θόρυβος, ο οποίος μπορεί να χωριστεί σε εσωτερικό θόρυβο που προκύπτει από το σύστημα μετάδοσης του ήχου, ή θόρυβο που προστίθεται στο σήμα προς αναγνώριση κατά την αναπαραγωγή του στον δέκτη και κατά την στιγμή ηχογράφησης του αποσπάσματος (θόρυβος παρασκηνίου). Και οι δύο κατηγορίες μπορούν να μοντελοποιηθούν ως πρόσθεση θορύβου στο καθαρό αρχικό σήμα.

Η λίστα αυτή περιλαμβάνει τις πιο βασικές παραμορφώσεις που εφαρμόζονται από ραδιοφωνικούς σταθμούς κατά την αναπαραγωγή των κομματιών, υπάρχουν όμως και άλλα μοντέλα επεξεργασίας ήχου που χρησιμοποιούνται (stereo enhancers, limeters, enhancers κ.α.) [2][11].

Παραμόρφωση	Μοντέλο Συναρτήσεων		
Περικοπή	$S :$	$[0; M] \times [0; f_s/2] \longrightarrow C$	
		$(t, f) \longrightarrow S(t + t_0, f)$	
Ισοστάθμιση	$\tilde{S}(t, f) = S(t, f)h(f)$ όπου $h : [0; f_s/2] \rightarrow [0; 1]$		
Συμπίεση	$\tilde{S}(t, f) = S(t, f)g(t)$ όπου $g : [0; L] \rightarrow [0; 1]$		
Τονική Μετατόπιση	$\tilde{S}(t, f) = S(t, Kf)$		
Χρονικη Παραμόρφωση	$\tilde{S}(t, f) = S(K't, f)$ όπου $K' = 1/K$		
Προσθήκη θορύβου	$\tilde{S}(t, f) = S(t, f) + n(t, f)$		

Πίνακας 2.1: Μαθηματική αναπαράσταση των συνηθέστερων ηχητικών παραμορφώσεων.

3 Μαθηματικά Εργαλεία Ψηφιακής Επεξεργασίας Σήματος

3.1 Μετασχηματισμός Φουριέ Βραχύ Χρόνου

Η κύρια ιδέα πίσω από τον μετασχηματισμό Fourier σύντομου χρόνου είναι η έκφραση του σήματος σαν ένας γραμμικός συνδυασμός στοιχειώδων σημάτων τα οποία είναι πιο εύκολο να κατανοηθούν αλλά και να υποστούν επεξεργασία[4]. Η προκύπτουσα αναπαράσταση περιέχει πληροφορία για το πως είναι κατανεμημένη η ενέργεια τόσο στον χρόνο όσο και στη συχνότητα.

Ο STFT προκύπτει από τον διακριτό μετασχηματισμό Fourier, ο οποίος είναι ουσιαστικά η ομοιόμορφη δειγματοληψία του Μετασχηματισμού Φουριέ σε διακριτές συχνότητες. Είναι δυνατό να υπολογίσουμε τον DFT ενός ολόκληρου ηχητικού δείγματος και να βρούμε πως η ενέργεια του σήματος κατανέμεται ανάμεσα στις διαφορετικές συχνότητες. Παρόλα αυτά μια τέτοια ανάλυση δεν δίνει καμία πληροφορία για το πώς αυτές οι συχνότητες μεταβάλλονται στον χρόνο. Ο Μετασχηματισμός Φουριέ Βραχύ Χρόνου (STFT) προκύπτει από την επεξεργασία μικρών τμημάτων του σήματος, που λέγονται πλαίσια, με την εφαρμογή DFT σε κάθε ένα από αυτά. Ο συνηθισμένος τρόπος ορισμού του διακριτού μετασχηματισμού Fourier βραχύ χρόνου για το πλαίσιο m είναι ο ακόλουθος:

3.1.1 Μαθηματικός Ορισμός Μετασχηματισμού Φουριέ Βραχύ Χρόνου

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)w(n - mR)e^{-j\omega n} = DTFT_{\omega}(xSHIFT_{mR}(w))$$

όπου

- $x(n)$: Σήμα εισόδου την χρονική στιγμή n
- $w(n)$: Ακολουθία παραθύρου μήκους M
- $X_m(\omega)$: DTFT των παραθυρωμένων δεδομένων τη χρονική στιγμή mR
- R : Μέγεθος βήματος σε δείγματα, μεταξύ διαδοχικών DTFT

Αν η ακολουθία παραθύρωσης $w(n)$ έχει την ιδιότητα συνεχούς επικαλυπτόμενης άνθροισης στο βήμα R , δηλαδή αν:

$$\sum_{m=-\infty}^{\infty} w(n - mR) = 1, \forall n \in \mathbb{Z}$$

τότε, το άνθροισμα των διαδοχικών DTFT στον χρόνο ισούται με τον DTFT του συνολικού σήματος X :

$$\sum_{m=-\infty}^{\infty} X_m(\omega) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x(n)w(n - mR)e^{-j\omega n}$$

$$\begin{aligned}
&= \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \sum_{m=-\infty}^{\infty} w(n-mR) \\
&= \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} = DTFT_{\omega}(x) = X(\omega)
\end{aligned}$$

3.1.2 Πρακτικός Υπολογισμός Μετασχηματισμού Φουριέ Βραχύ Χρόνου

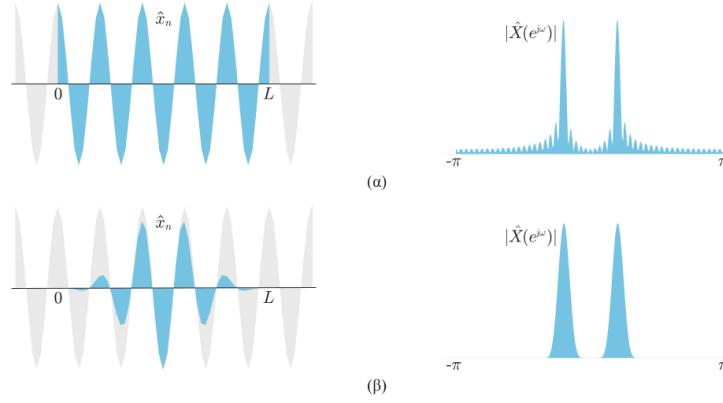
Στην πράξη ο υπολογισμός του Μετασχηματισμού Φουριέ Βραχύ Χρόνου γίνεται με τον διαδοχικό υπολογισμό του Διακριτού Μετ. Φουριέ, με τον αλγόριθμο Ταχύ Μετ. Φουριέ, σε παραθυρωμένα πλαίσια της ακολουθίας εισόδου (με κατάλληλο μήκος πλαισίων), με το παράθυρο να μετατοπίζεται προς τα μπροστά στον χρόνο. Η υλοποίηση αυτή μπορεί να προκύψει από τον θεωρητικό ορισμό που δόθηκε παραπάνω:

Αν προσθέσουμε mR στο n , προκύπτει:

$$\begin{aligned}
X_m(\omega) &= \sum_{n=-\infty}^{\infty} x(n+mR)w(n)e^{-j\omega(n+mR)} \\
&= \sum_{n=-\infty}^{\infty} x(n+mR)w(n)e^{-j\omega n}e^{-j\omega mR} \\
&= e^{-j\omega mR} \sum_{n=-\infty}^{\infty} x(n+mR)w(n)e^{-j\omega n} \\
&= e^{-j\omega mR} DTFT_{\omega}(SHIFT_{-mR}(x)w)
\end{aligned}$$

Σε αυτή την μορφή, τα δεδομένα που είναι κεντραρισμένα στην χρονική στιγμή mR μεταφράζονται στην χρονική στιγμή 0, πολλαπλασιάζονται με την ακολουθία παραθυρωσης w και έπειτα εκτελείται ο DTFT. Επειδή η μη-μηδενική ποσότητα των παραθυρωμένων δεδομένων βρίσκεται είναι κεντραρισμένη στην χρονική στιγμή 0, ο DTFT μπορεί να αντικατασταθεί με τον DFT, να γίνει, δηλαδή, δειγματοληφία του στην συχνότητα. Η δειγματοληφία του άξονα της συχνότητας διατηρεί την πληροφορία όταν το σήμα είναι χρονικά περιορισμένο με κατάλληλο τρόπο. Έστω M το μήκος του παραθύρου και $N \geq M$ το μήκος του DFT, που συνήθως είναι δύναμη του 2. Δειγματοληπτώντας στα σημεία $\omega = \omega_k = 2\pi k/N, k = 0, 1, 2, \dots, N - 1$, και χρησιμοποιώντας το γεγονός ότι το παράθυρο $w(n)$ είναι χρονικά περιορισμένο σε λιγότερα από N δείγματα γύρω από το 0, πάρουμε:

$$X_m(\omega_k) = e^{-j\omega_k mR} \sum_{n=-N/2}^{N/2-1} x(n+mR)w(n)e^{-j\omega_k n}$$



Σχήμα 3.1: Εκτίμηση συχνοτικού περιεχομένου καθαρού ημιτονικού σήματος, (α) ορθογώνια παραθύρωση, (β) παραθύρωση με παράθυρο Kaiser[6].

$$= e^{-j\omega_k mR} DFT_{N,\omega_k}(SHIFT_{-mR}(x)w)$$

Αφού δεικτοδοτούμε τον *DFT* βάση του *moduloN*, το άθροισμα στο n μπορεί να μετατραπεί σε ένα άθροισμα από το 0 μέχρι το $N - 1$ όπως είναι συνηθισμένο να γίνεται στην περίπτωση του *DFT*. Στην πράξη αυτό σημαίνει ότι το δεξί μισό του παραθυρωμένου πλαισίου, πάει στο τέλος με συμπλήρωση μηδενικών στην μέση.

3.1.3 Σημασία της Ακολουθίας Παραθύρωσης

Ο ορισμός που δόθηκε για τον STFT δείχνει ότι ο μετασχηματισμός εξαρτάται τόσο από το σήμα όσο και από την ακολουθία παραθύρωσης, παρότι συνήθως ενδιαφερόμαστε μόνο για τις ιδιότητες του σήματος. Η ακολουθία παραθύρωσης w_n συνήθως επιλέγεται με τρόπο τέτοιο ώστε να έχει μηδενική τιμή έξω από κάποια συγκεκριμένη περιοχή. Έτσι, όταν πολλαπλασιάζεται με το σήμα, το γινόμενο να είναι επίσης μηδενικό έξω από αυτή την περιοχή. Ένα ορθογώνιο παράθυρο w_n , μήκους L είναι η απλούστερη μορφή παραθύρου και ορίζεται ως:

$$w_n = \begin{cases} 1 & , 0 \leq n \leq L - 1 \\ 0 & , \text{αλλού} \end{cases}$$

Όπως φαίνεται και στο Σχήμα 3.2(α), η χρήση του ορθογώνιου παραθύρου μπορεί να προκαλέσει ασυνέχειες στα όρια της παραθυρωμένης περιοχής και τέτοιες απότομες αλλαγές μπορεί να προκαλέσουν παρασιτικές τιμές σε όλο το συχνοτικό φάσμα. Οι τιμές αυτές προέρχονται από ιδιότητες του ορθογώνιου παραθύρου και όχι του αρχικού σήματος.

Η εκτίμηση του συχνοτικού περιεχομένου μπορεί να βελτιωθεί, εάν χρησιμοποιήσουμε

εναλλακτικά παράθυρα στη θέση του ορθογώνιου, τα οποία επιχειρούν να εξομαλύνουν τις απότομες αλλαγές που εμφανίζονται στην ορθογώνια παραθύρωση, τοποθετώντας μικρό βάρος στα δείγματα που είναι κοντά στις χρονικές στιγμές 0 και L . Αυτό φαίνεται και στο Σχήμα 3.1(β), όπου το παράθυρο που χρησιμοποιείται είναι παράθυρο Kaiser, με αποτέλεσμα οι παρασιτικές συχνότητες να εξαλείφονται από την εκτίμηση του συχνοτικού περιεχομένου.

3.2 Φασματογράφημα Μετασχηματισμού Φουριέ Βραχύ Χρόνου

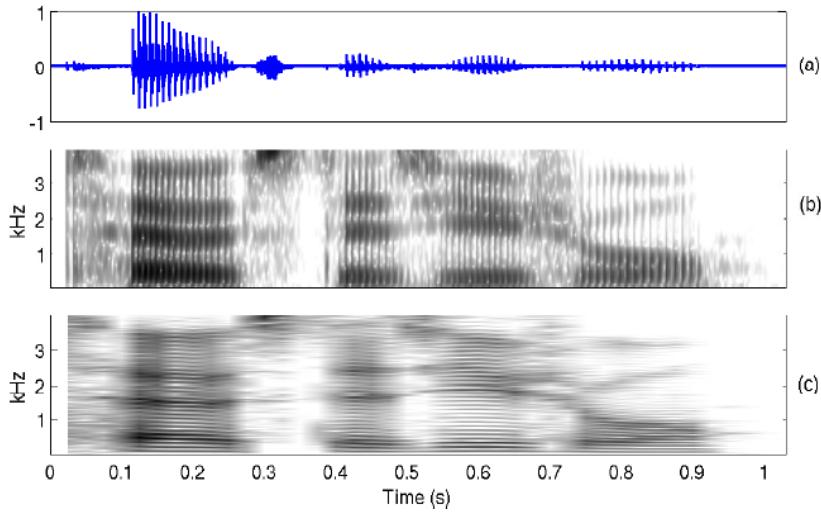
Ο Μετασχηματισμός Φουριέ Βραχύ Χρόνου ενός σήματος x για κάθε σημείο στον χρόνο t και στην συχνότητα ω , επιστρέφει έναν μιγαδικό αριθμό $X(m, \omega)$. Αυτή η πληροφορία αναπαριστάται συνήθως με τη βοήθεια του φασματογραφήματος το οποίο είναι μια αναπαράσταση δύο διαστάσεων του τετραγώνου του πλάτους:

$$Spec(m, \omega) = |X(m, \omega)|^2$$

Όταν δημιουργείται μια απεικόνιση ενός φασματογραφήματος, ο οριζόντιος άξονας αναπαριστά τον χρόνο, ο κάθετος άξονας αναπαριστά την συχνότητα και η διάσταση που δίνει την τιμή του φασματογραφήματος σε μια συγκεκριμένη χρονική στιγμή αναπαριστάται με την ένταση του χρώματος στην εικόνα. Για να τονιστούν σχέσεις που αφορούν μουσική ή τονικά χαρακτηριστικά ο άξονας των συχνοτήτων συχνά απεικονίζεται με λογαριθμικό τρόπο, που είναι γνωστή ως αναπαράσταση της λογαριθμικής συχνότητας. Ένας άξονας συχνοτήτων σε λογαριθμική κλίμακα σχετίζεται επίσης με το ότι η ακουστική αντίληψη του ανθρώπου στον τόνο είναι λογαριθμικής φύσεως. Έτσι, μικρές τιμές που ωστόσο είναι αντιληπτές στο αυτί, γίνονται ορατές και στην εικόνα.

Η τελική μορφή του φασματογραφήματος εξαρτάται σημαντικά από το μήκος L του παραθύρου που χρησιμοποιείται κατά τον υπολογισμό του Μετασχηματισμού Φουριέ Βραχύ Χρόνου, καθώς η διακριτική ικανότητα του, στο πεδίο της συχνότητας, βελτιώνεται για αυξανόμενο L .

Καθώς το παράθυρο ολισθαίνει προς τα δεξιά, για να υπολογιστεί το συχνοτικό περιεχόμενο για κάθε χρονική στιγμή, οι συντελεστές του w_n συμπεριφέρονται σαν ένα FIR κατωπερατό φίλτρο. Για αυτό τον λόγο, το παράθυρο εμφανίζει μεταβατικά φαινόμενα που διαρκούν όσο είναι και το μήκος του φίλτρου. Τα μεταβατικά φαινόμενα αυτά είναι η αιτία του γεγονότος πως όσο πιο μεγάλο είναι το μήκος του παραθύρου, τόσο μικρότερη διακριτική ικανότητα υπάρχει στο πεδίο του χρόνου. Η ανταγωνιστική αυτή σχέση μεταξύ της διακριτικής ικανότητας στον χρόνο και την συχνότητα στο φασματογράφημα είναι γνωστή ως αρχή της αβεβαιότητας συχνότητας - χρόνου, και μοιάζει σε υφή με την αρχή αβεβαιότητας του Heisenberg στο πεδίο της Κβαντομηχανικής Φυσικής [6], και ένα παράδειγμα της φαίνεται και στο παρακάτω σχήμα.

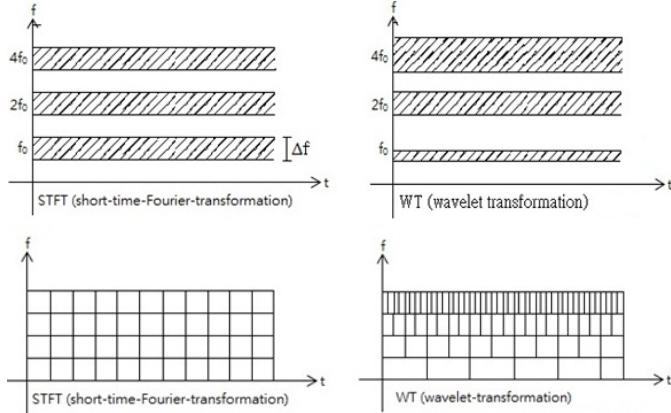


Σχήμα 3.2: (a) Κυματομορφή Ομιλίας. (b) Φασματογράφημα με πλατιές φασματικές μπάντες. (c) Φασματογράφημα με στενές φασματικές μπάντες. [5]

3.3 Μετασχηματισμός Σταθερού-Q

Ο μετασχηματισμός σταθερού Q [19] είναι ένα ακόμη μαθηματικό εργαλείο που εκφράζει ένα σήμα ως μια αναπαράσταση που περιέχει πληροφορία για την κατανομή της ενέργειας του τόσο στον χρόνο όσο και στις συχνότητες. Η ιδιαιτερότητα που παρουσιάζει είναι πως οι κεντρικές συχνότητες των συχνοτικών ζωνών ανάλυσης ακολουθούν γεωμετρική κατανομή και ταυτόχρονα οι συντελεστές Q , που ορίζονται ως ο λόγος τις κεντρικής συχνότητας προς το εύρος ζώνης των ζωνοπεριορισμένων φίλτρων, είναι σταθεροί για όλη την τράπεζα φίλτρων που χρησιμοποιείται. Αυτή η τεχνική έχει ως αποτέλεσμα πως η συχνοτική διαχριτική ικανότητα του μετασχηματισμού είναι καλύτερη για τις χαμηλές συχνότητες από τις υψηλές.

Στην ουσία, ο μετασχηματισμός σταθερού Q (CQT) είναι ένας μετασχηματισμός κυμάτων (wavelet), αλλά η ονομασία του τονίζει πως κλασσικές τεχνικές όπως αυτές που βασίζονται στις επαναλαμβανόμενες τράπεζες φίλτρων είναι ανεπαρκείς για αρκετά υψηλές τιμές του συντελεστή Q (όπως και στην περίπτωση αυτή) που αντιστοιχούν σε 12-96 συχνοτικές ζώνες ανά μουσική οκτάβα.



Σχήμα 3.3: Γραφικές Διαφορές STFT με ένα τυπικό μετασχηματισμό χυμάτιων.

Ο CQT έχει πλεονεκτήματα και από την πλευρά της μουσικής αλλά και από της ανθρώπινης ακουστικής αντίληψης. Οι θεμελιώδεις συχνότητες $F0$ στους τόνους της Δυτικής μουσικής θεωρίας είναι κατανεμημένοι γεωμετρικά. Για παράδειγμα, στο κλασσικό κούρδισμα των οργάνων σε 12 τόνους, για τις θεμελιώδεις συχνότητες ισχύει $F0 = 440 \times 2^{k/12} Hz$, όπου το $k \in [-50, 40]$ είναι ακέραιος. Από ακουστικής πλευράς, η διακριτική ικανότητα στις συχνότητες της ανθρώπινης ακοής έχει κατά προσέγγιση σταυρερό συντελεστή ποιότητας (Q) από τα $20kHz$ μέχρι περίπου τα $500Hz$, ενώ κάτω από αυτό το όριο οι τιμές του Q σταδιακά μειώνονται [18]. Είναι λοιπόν εμφανές πως ο κλασικός Διακριτός Μετασχηματισμός Φουριέ δεν είναι σε θέση να ικανοποιήσει την συνθήκη της χρονικής και συχνοτικής διακριτικής ικανότητας έτσι ώστε αυτές να έχουν διαφορετικές τιμές στο φάσμα των ακουστικών συχνοτήτων, καθώς χρησιμοποιεί γραμμικά κατανεμημένες συχνοτικές ζώνες.

	C	C#	D	E♭	E	F	F#	G	G#	A	B♭	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

Σχήμα 3.4: Αντιστοιχία συχνοτήτων - μουσικών νοτών ανά οκτάβα (Hz).

Τηπάρχουν όμως τρεις βασικοί λόγοι που ο μετασχηματισμός αυτός δεν έχει αντικαταστήσει σε μεγάλο βαθμό τον DFT στην επεξεργασία ήχου. Πρώτον, είναι

υπολογιστικά ακριβότερος. Δεύτερον, υπάρχει έλλειψη αντίστροφου μετασχηματισμού τέλειας ανακατασκευής του αρχικού σήματος. Τρίτον, η έξοδος του μετασχηματισμού είναι μια δομή δεδομένων που είναι πιο περίπλοκη από το κλασσικό φασματογράφημα που προκύπτει από τον Μετασχηματισμό Φουριέ Σύντομου Χρόνου. Στο [20], το οποίο χρησιμοποιείται και στην συνέχεια, προτείνονται λύσεις και για τα τρία αυτά ζητήματα.

3.3.1 Μαθηματική Μοντελοποίηση Σήματος

Ο μετασχηματισμός Σταθερού- Q $X^{CQ}(k, n)$ ενός διακριτού σήματος $x(n)$ ορίζεται ως:

$$X^{CQ}(k, n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j)a_k^*(j - n + N_k/2)$$

όπου $k = 1, 2, \dots, K$ είναι οι δείκτες των συχνοτικών ζώνων του μετασχηματισμού, και $a_k^*(n)$ είναι ο συζυγής μιγαδικός του $a_k(n)$. Οι συναρτήσεις βάσης $a_k(n)$ είναι μιγαδικές κυματομορφές, που ονομάζονται και άτομα χρόνου-συχνότητας και ορίζονται ως:

$$a_k(n) = \frac{1}{N_k} w\left(\frac{n}{N_k}\right)^{[-i2\pi n \frac{f_k}{f_s}]}$$

όπου f_k η κεντρική συχνότητα της ζώνης k , f_s η συχνότητα δειγματοληψίας του σήματος και $w(n)$ μια συνεχής συνάρτηση παραθύρωσης (πχ. Hann, Blackman) δειγματοληπτημένο σε σημεία που καθορίζονται από το t . Η $w(n)$ είναι μηδέν εκτός του $t \in [0, 1]$.

Τα μήκη των παραθύρων $N_k \in \mathbb{R}$ είναι αντιστρόφως ανάλογα των κεντρικών συχνοτήτων f_k έτσι ώστε ο συντελεστής- Q να είναι σταθερός για κάθε ζώνη k .

Επιπλέον, οι κεντρικές συχνότητες ακολουθούν τον τύπο:

$$f_k = f_1 2^{\frac{k-1}{B}}$$

όπου f_1 είναι η κεντρική συχνότητα της χαμηλότερης συχνοτικής ζώνης και το B καθορίζει το πλήθος των συχνοτικών ζώνων ανά οκτάβα. Η παράμετρος B είναι στην πραγματικότητα η σημαντικότερη παράμετρος όταν χρησιμοποιείται ο μετασχηματισμός διότι καθορίζει την αντίστροφη σχέση της διακριτικής ικανότητας χρόνου-συχνότητας.

Ο συντελεστής ποιότητας Q του φίλτρου k δίνεται από τον τύπο:

$$Q_k = \frac{f_k}{\Delta f_k} = \frac{N_k}{\Delta \omega f_s}$$

όπου Δf_k είναι το $-3dB$ εύρος ζώνης της απόκρισης συχνότητας της συνάρτησης βάσης $a_k(n)$ και $\Delta \omega$ είναι το $-3dB$ εύρος ζώνης του κεντρικού λοβού του φάσματος της

συνάρτησης παραθύρωσης. Αφού οι συντελεστές- Q Q_k είναι σταθεροί για όλα τα φίλτρα εξ ορισμού, ο δείκτης k θα παραλείπεται.

Είναι θεμιτό ο συντελεστής- Q να έχει μεγάλες τιμές, έτσι ώστε το εύρος ζώνης κάθε φίλτρου Δf_k να είναι όσο μικρότερο γίνεται. Έτσι, διασφαλίζεται η ελαχιστοποίηση της συχνοτικής επικάλυψης μεταξύ των διαδοχικών φίλτρων. Από την άλλη πλευρά όμως, ο Q δεν μπορεί να παίρνει αυθαίρετα μεγάλες τιμές, διότι κάποια σημεία του συχνοτικού φάσματος μπορεί να ξεφεύγουν από το εύρος ζώνης του κοντινότερου φίλτρου. Ο τύπος που διασφαλίζει την βέλτιστη τιμή του συντελεστή είναι:

$$Q = \frac{q}{\Delta\omega(2^{\frac{1}{B}} - 1)}$$

όπου $0 < q \leq 1$ είναι παράγοντας χλιμάκωσης και συνήθως $q = 1$. Όταν το q παίρνει τιμές μικρότερες του 1, η χρονική διακριτική ικανότητα του μετασχηματισμού βελτιώνεται σε βάρος της συχνοτικής. Είναι ενδιαφέρον πως με τιμές $q < 1$ γίνεται υπερδειγματοληψία στο πεδίο της συχνότητας και κατά μια έννοια λειτουργούν ανάλογα με την συμπλήρωση με μηδενικά όταν υπολογίζεται ο DFT . Για παράδειγμα $q = 0.5$ αντιστοιχεί σε υπερδειγματοληψία παράγοντα 2.

Αντικαθιστώντας από τους προηγούμενους τύπους, προκύπτει πως:

$$N_k = \frac{qf_s}{f_k(2^{\frac{1}{B}} - 1)}$$

όπου η εξάρτηση από τον παράγοντα $\Delta\omega$ εξαλείφεται.

Καθώς το υπολογιστικό κόστος για τον υπολογισμό των συντελεστών $X^{CQ}(k, n)$ για κάθε δείγμα n είναι στην ουσία απαγορευτικό, οι διαδοχικές συναρτήσεις βάσης $a_k(n)$ υπολογίζονται ανά H_k δείγματα (hop size). Για να γίνει σωστά η διαδικασία της ανάλυσης του σήματος αλλά και της ανακατασκευής του σήματος, προτείνεται το βήμα να είναι στο εύρος $0 < H_k < \frac{1}{2}N_k$.

3.3.2 Αποδοτικός Υπολογισμός Μετασχηματισμού Σταθερού- Q

Ο αποδοτικός αλγόριθμος υπολογισμού βρίσκεται στο [20] και η κύρια λογική του βασίζεται στο [21] και γίνεται μέσω 4 ιδεών:

Μετακίνηση από το πεδίο του χρόνου στο πεδίο της συχνότητας μέσω της ιδιότητας:

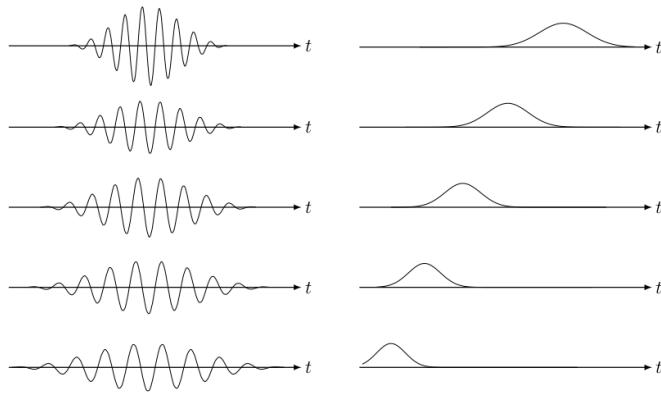
$$\sum_{n=0}^{N-1} x(n)a^*(n) = \sum_{i=0}^{N-1} N - 1 X(i)A^*(i)$$

όπου $X(i)$ είναι ο DFT του $x(n)$ και $A(i)$ ο DFT του $a(n)$. Η ιδιότητα αυτή προκύπτει από το θεώρημα του Parseval, ισχύει για κάθε διαχριτό σήμα και με την χρήση της, ο

ορισμός του μετασχηματισμού μπορεί να ξαναγραφτεί ως:

$$X^{CQ}(k, \frac{N}{2}) = \sum_{i=0}^{N-1} N - 1 X(i) A_k^*(i)$$

όπου $A_k(j)$ είναι ο *DFT* N δειγμάτων των βάσεων $a_k(n)$ έτσι ώστε οι βάσεις να είναι κεντραρισμένες στο δείγμα $N/2$ για κάθε πλαίσιο του μετασχηματισμού. Όπως αναφέρεται στο [21], οι $A_k(j)$ ονομάζονται φασματικοί πυρήνες και οι $a_k(n)$ χρονικοί πυρήνες, και έχουν την μορφή που εμφανίζεται στο παρακάτω σχήμα.



Σχήμα 3.5: Κάποιοι από τους χρονικούς πυρήνες με τους αντίστοιχους φασματικούς τους, ξεκινώντας από υψηλότερη προς χαμηλότερη συχνότητα. Είναι φανερό πως οι φασματικοί πυρήνες δεν έχουν το ίδιο μήκος.

Καθώς οι χρονικοί πυρήνες $a_k(n)$ είναι ημιτονοειδή σήματα διαμορφωμένα κατά πλάτος, οι μετασχηματισμοί Φουριέ τους $A_k(k)$ είναι αραιοί. Οι περισσότερες τιμές τους είναι πολύ μικρές σε σχέση με το μέγιστο της συνάρτησης, και υπάρχει μόνο μια κορυφή στο συχνοτικό φάσμα.

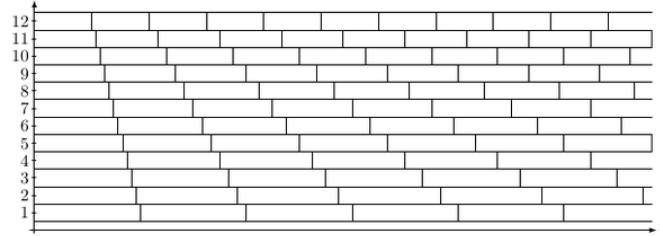
Οι συχνοτικές πυρήνες τώρα μπορούν να αποθηκευτούν σε μορφή μητρώου και έτσι ο ορισμός του X^{CQ} μπορεί να γραφτεί ως:

$$X^{CQ} = A^* X$$

όπου Q είναι διάνυσμα που περιέχει τις τιμές του $Q(i)$. Λόγω της δομής των $A_k(k)$, στον πολλαπλασιασμό διανυσμάτων χρησιμοποιούνται αλγόριθμοι γρήγορου υπολογισμού που παραλείπουν τους πολλαπλασιασμούς με πολύ μικρές τιμές του A^* .

Πλέον, είναι δυνατόν ο A να υπολογιστεί μια φορά και να υπολογίζεται ο X^{CQ} για όλες τις συχνοτικές ζώνες για κάθε δείγμα, αφού ο A παραμένει σταθερός για κάθε πλαίσιο. Όμως ο υπολογισμός του X^{CQ} δεν χρειάζεται να γίνεται ανά διαδοχικά δείγματα, καθώς λόγω της σχέσης συχνότητας-χρόνου οι τιμές του δεν αλλάζουν ανά δείγμα. Αυτό

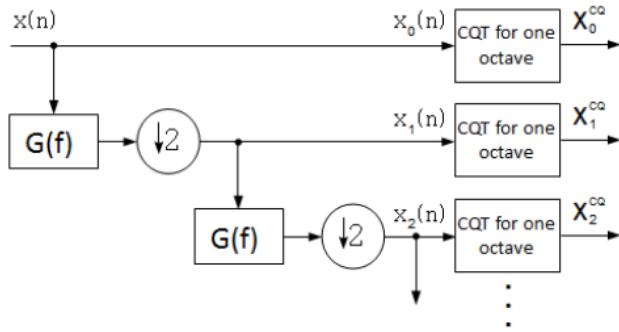
συμβαίνει διότι επιλέγεται επικάλυψη 50% του μήκους του πλαισίου N_k . Στην συγκεκριμένη περίπτωση, λόγω της ιδιαιτερότητας του μετασχηματισμού, τα δείγματα επικάλυψης και το μήκος του πλαισίου εξαρτώνται από την συχνότητα, και έτσι είναι διαφορετικά για κάθε συχνοτική ζώνη k .



Σχήμα 3.6: Τα σχετικά μήκη των ακολουθιών παραθύρωσης N_k του μετασχηματισμού.

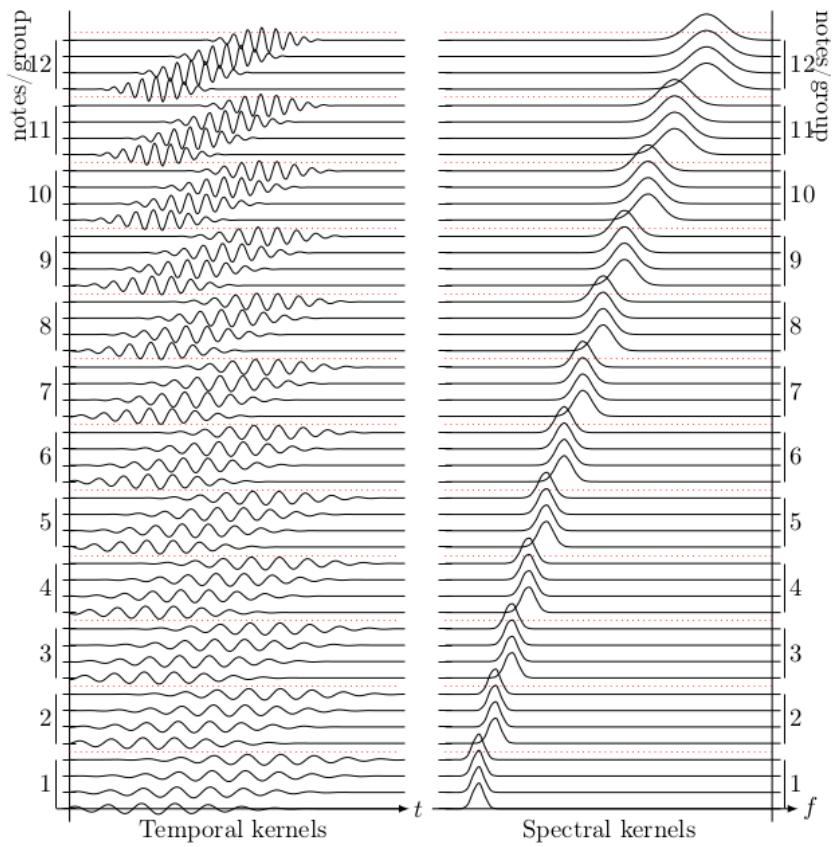
Αν επιλεγεί το ίδιο ποσοστό επικάλυψης για κάθε συχνοτική ζώνη, τότε η πράξη στην μορφή μητρώων δεν θα μπορούσε να υλοποιηθεί, το αποτέλεσμα δεν θα ήταν σε μορφή παραλληλεπίπεδου, αλλά τραπέζιο. Για αυτό επιλέγεται σταυθερός αριθμός επικάλυψης σε δείγματα για κάθε ζώνη. Για να μην χαθεί πληροφορία, το μήκος της επικάλυψης πρέπει να είναι τουλάχιστον το 50% του N_k^{min} , δηλαδή το μισό του μήκους του μικρότερου φασματικού πυρήνα (αυτού στην χαμηλότερη συχνότητα). Με αυτόν τον τρόπο εισάγονται περιττές πράξεις στον υπολογισμό, από την άλλη πλευρά όμως, ο χειρισμός των αποτελεσμάτων γίνεται ευκολότερος.

Μια ακόμα μέθοδος που εφαρμόζεται για την επιτάχυνση της διαδικασίας, είναι ο διαχωρισμός της επεξεργασίας και υπολογισμού ανά οκτάβα. Αρχικά, χρησιμοποιώντας το μητρώο φασματικών πυρήνων A που χρειάζεται, υπολογίζουμε τον CQT της υψηλότερης οκτάβας για όλη την διάρκεια του σήματος. Στην συνέχεια, το σήμα περνάει από ένα χαμηλοπερατό φίλτρο, υποδειγματοληπτείται κατά παράγοντα 2, και η ίδια διαδικασία επαναλαμβάνεται για να υπολογιστεί ο CQT της αμέσως χαμηλότερης οκτάβας. Με αυτή την προσέγγιση το μήκος του DFT είναι το μίσο ανά οκτάβα και ο A παραμένει αραιός.



Σχήμα 3.7: Επισκόπηση του συστήματος υπολογισμού του *CQT* ανά οκτάβα. Το $G(f)$ συμβολίζει το χαμηλοπερατό φίλτρο, ενώ το $\downarrow 2$ συμβολίζει την υποδειγματοληψία ανά παράγοντα 2.

Τέλος, γίνεται επιπλέον μείωση των απαιτούμενων πράξεων κατά τον υπολογισμό του *CQT*, **χρησιμοποιώντας χρονικές μετατοπίσεις** του ίδιου χρονικού πυρήνα $a(n)$ μέσα στα πλαίσια του ίδιου φασματικού πυρήνα $A(k)$. Με αυτόν τον τρόπο, χρειάζονται λιγότεροι υπολογισμοί των $X(i)$. Με άλλα λόγια, πλέον, διαδοχικές στήλες του A θα περιέχουν τους *DFT* των $a_k(n)$ που έχουν μετατοπισθεί χρονικά. Οι χρονικοί πυρήνες πρέπει να καλύπτουν όλη την διάρκεια του σήματος, αλλιώς επιμέρους τμήματα του σήματος παραλείπονται και έτσι χάνεται πληροφορία. Αν υπάρχουν P διαδοχικοί χρονικοί πυρήνες μέσα στον ίδιο φασματικό πυρήνα, οι *DFT* μπορούν να υπολογιστούν $P - 1$ φορές λιγότερες.



Σχήμα 3.8: Οι μετατοπισμένοι πυρήνες με τους αντίστοιχους φασματικούς πυρήνες τους.
Στο διάγραμμα εμφανίζεται μόνο το πραγματικό μέρος τους.

4 Η μέθοδος του Shazam

4.1 Τι είναι το Shazam

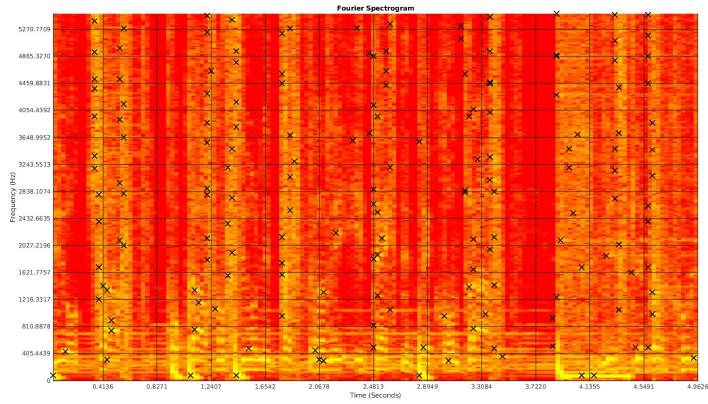
Μια από τις πιο διάσημες εφαρμογές που βασίζεται στο ακουστικό αποτύπωμα αναπτύχθηκε από την Shazam. Η εταιρία αυτή ήταν η πρώτη που παρουσίασε μια επιτυχημένη υπηρεσία αναγνώρισης μουσικών κομματιών μέσω αναζήτησης από κινητό τηλέφωνο. Καθώς η υπηρεσία αυτή έβαλε την τεχνική του ακουστικού αποτυπώματος στο προσκήνιο, η δουλεία τους αναφέρεται πολύ συχνά σε επιστημονικά συγγράμματα αυτού του τομέα. Συγκεκριμένα, θα περιγραφεί η προσέγγιση που ακολούθησε ο Avery Wang στο [7], που είναι και το μοναδικό που έχει δημοσιευθεί, και αφορά την τεχνική που χρησιμοποιούν. Φυσικά, επειδή αναφερόμαστε για μια εταιρία και επιπλέον έχει περάσει αρκετός καιρός από την δημοσίευση του, ενδέχεται να υπάρχει αρκετά μεγάλη διαφορά με τον αλγόριθμο που χρησιμοποιείται στην πραγματικότητα.

4.2 Αναπαράσταση Σήματος

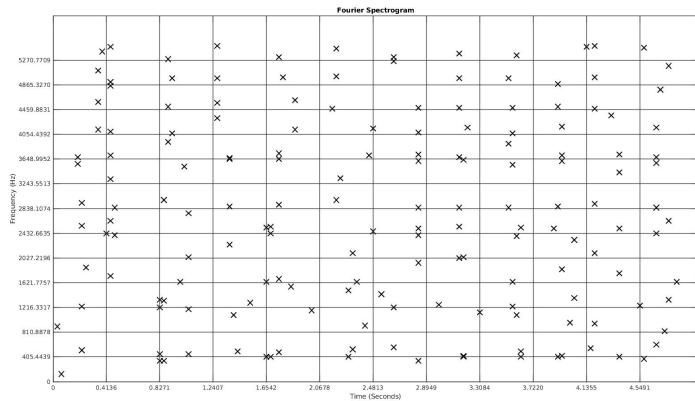
Όπως στις περισσότερες τεχνικές με ακουστικό αποτύπωμα, η αναπαράσταση που χρησιμοποιείται στο Shazam βασίζεται στο φασματόγραμμα του σήματος. Βέβαια, το φασματόγραμμα που προκύπτει δεν είναι καλή αναπαράσταση για αποτύπωμα από μόνο του, καθώς δεν είναι ευέλικτο γύρω από τις κλασσικές παραμορφώσεις που εξετάσαμε. Όπως και στις περισσότερες μεθόδους, η ιδέα εδώ βρίσκεται στην απλοποίηση της πληροφορίας που προσφέρει το φασματόγραμμα ώστε η πληροφορία να είναι πιο σταθερή και να βασίζεται στα κεντρικά χαρακτηριστικά του σήματος. Έτσι, το φασματογράφημα χωρίζεται σε κελιά συγκεκριμένου εύρους συχνοτήτων και χρόνου, και κάθε κελί κρατάει μόνο τις συντεταγμένες του τοπικού μεγίστου, αδιαφορώντας για όλες τις άλλες τιμές.

Με αυτή την τροποποίηση η αναπαράσταση του σήματος γίνεται πιο χαρακτηριστική. Μειώνοντας την πληροφορία του φασματογράφηματος με αυτόν τον τρόπο, αφαιρείται η πληροφορία του πλάτους καθώς και αυξάνεται η ευελιξία ως προς τις αλλαγές του σήματος στις κλασσικές παραμορφώσεις. Αυτό γίνεται διότι κρατώντας μόνο την περισσότερη επικρατέστερη πληροφορία, η αναπαράσταση γίνεται πιο ευέλικτη στην προσθήκη θορύβου, η οποία προσθέτει μικρά ποσά ενέργειας σε διάφορα σημεία του φασματογράμματος, αλλά και στις παραμορφώσεις που αλλάζουν την κατανομή ενέργειας του σήματος χωρίς όμως να επηρεάζεται η κατανομή των τοπικών μεγίστων που επιλέγονται (ισοστάθμιση, δυναμική συμπίεση).

Βέβαια, η αναπαράσταση που προκύπτει είναι ακόμα ευαίσθητη στις παραμορφώσεις του σήματος που μπορούν να προσθέσουν ή να αφαιρέσουν τα τοπικά μέγιστα. Για αυτό, κατά το βήμα της αναζήτησης δεν απαιτείται η πλήρης ταύτιση των μεγίστων, αλλά αναζητείται το απλοποιημένο φασματογράφημα που περιέχει το υψηλότερο αριθμό κοινών τοπικών μεγίστων με το query που γίνεται στη βάση δεδομένων. Χάριν επεκτασιμότητας, στο



(α') Τοπικά μέγιστα αποσπάσματος



(β') Φασματογράφημα Αποσπάσματος με τα τοπικά μέγιστα

Σχήμα 4.1: Αναπαράσταση σήματος 5 δευτερολέπτων με το φασματογράφημα, και τα τοπικά μέγιστα που προκύπτουν

Shazam προτείνεται η χρήση ενός συστήματος δεικτών αντί τις γραμμικής σύγχρισης κάθε αντικειμένου της βάσης.

4.3 Δημιουργία Κλειδιών

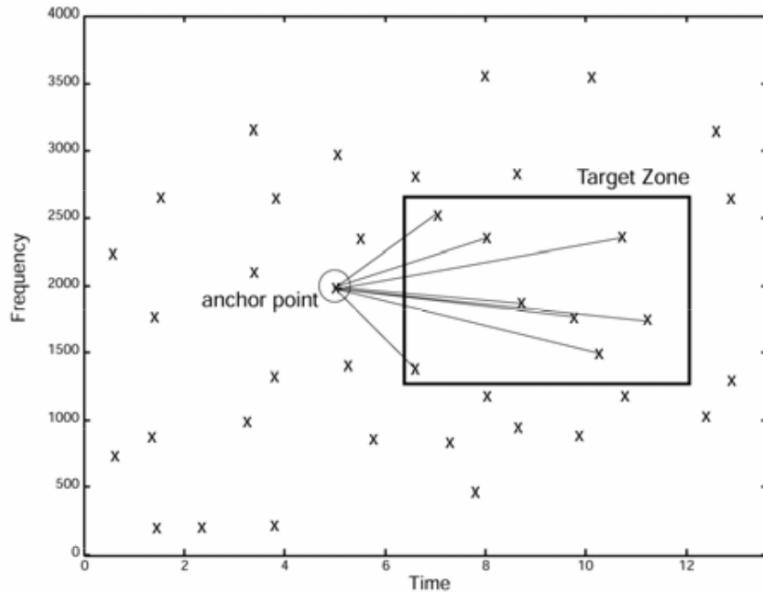
Προκύπτει λοιπόν, το ζήτημα της κατασκευής ενός κλειδιού που να βασίζεται στην παραπάνω τροποποίηση του φασματογραφήματος, το οποίο να παρέχει αρκετή πληροφορία για τα χαρακτηριστικά του σήματος. Η ανάλυση του Wang αναφέρει πως η χρήση των τοπικών μεγίστων καθ' αυτών ως κλειδιά για το αποτύπωμα του σήματος δεν αρκεί. Στο τροποποιημένο φασματογράφημα, ένα σημείο περιγράφεται πλήρως από τις συντεταγμένες του: την χρονική τοποθεσία t_0 και την συχνότητα f_0 . Καθώς, το σύστημα πρέπει να είναι ευέλικτο στην περικοπή του σήματος, δεν είναι δυνατόν να χρησιμοποιείται η απόλυτη χρονική πληροφορία από ένα μόνο σημείο, γιατί με την περικοπή του σήματος σε αυθαίρετα σημεία, η πληροφορία αυτή αλλάζει. Οπότε, από μεμονωμένα σημεία, μπορεί να αξιοποιηθεί μόνο η συχνοτική θέση f_0 . Αυτό θα οδηγούσε τα ερωτήματα στην βάση δεδομένων να είναι της μορφής “Επέστρεψε όλους του τίτλους, που περιέχουν στο τροποποιημένο φασματόγραφμα τους, τουλάχιστον ένα σημείο που έχει συχνότητα f_0 ”. Φυσικά, οι απαντήσεις αυτού του ερωτήματος δεν μπορούν να είναι χρήσιμες. Θεωρητικά, όλα τα κομμάτια της βάσης θα μπορούσαν να περιέχουν κάποιο σημείο.

Για να ξεπεραστεί αυτό το πρόβλημα, ο Wang προτείνει την χρήση ζευγαριών σημείων, αντί μεμονωμένων. Έστω, δύο τοπικά μέγιστα με συντεταγμένες (t_1, f_1) και (t_2, f_2) . Η παραπάνω ανάλυση μας οδηγεί πως ένα κλειδί θα είχε αρκετή πληροφορία αν χρησιμοποιεί τις συχνότητες των μεμονωμένων σημείων, και έκανε την χρονική πληροφορία σχετική ως προς τα δύο σημεία. Ο Wang προτείνει ένα κλειδί που να περιέχει τρία κομμάτια πληροφορίας: $f_1, f_2, t_2 \Gamma t_1$. Με αυτόν τον τρόπο οι απαντήσεις των ερωτημάτων είναι πολύ πιο συγκεκριμένες, καθώς τα κλειδιά περιέχουν αρκετή πληροφορία και ταυτόχρονα λύνεται το ζήτημα της περικοπής του σήματος.

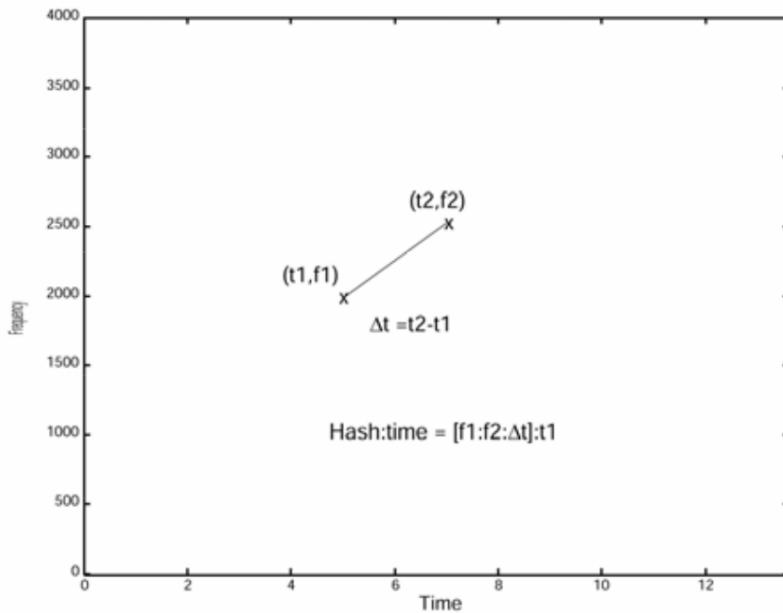
Από την άλλη πλευρά, επειδή τα κλειδιά περιέχουν αρκετή πληροφορία, είναι πιο πιθανό να αλλάζουν όταν το σήμα υπόκειται παραμόρφωση. Επίσης, με την παραπάνω προσέγγιση, οδηγούμαστε σε μεγαλύτερο αριθμό κλειδιών σε κάθε ηχητικό σήμα. Αντί να υπολογίζονται N κλειδιά, που να αντιστοιχίζονται στα N τοπικά μέγιστα, υπολογίζονται N^2 κλειδιά. Για να μειωθεί η πολυπλοκότητα αυτού του υπολογισμού, εφαρμόζεται ένας περιορισμός για τον αριθμό των ζευγαριών για κάθε τοπικό μέγιστο.

4.4 Μηχανισμός Αναζήτησης

Όταν αναγνωρίζεται ένα άγνωστο απόσπασμα ήχου, σκοπός είναι να βρεθεί το κομμάτι αναφοράς με το υψηλότερο αριθμό όμοιων κλειδιών. Όντως, αν το άγνωστο απόσπασμα u είναι απόσπασμα του κομματιού αναφοράς r_0 με χρόνο εκκίνησης d , τότε όλα τα κλειδιά



(α') Τοπικά μέγιστα αποσπάσματος.

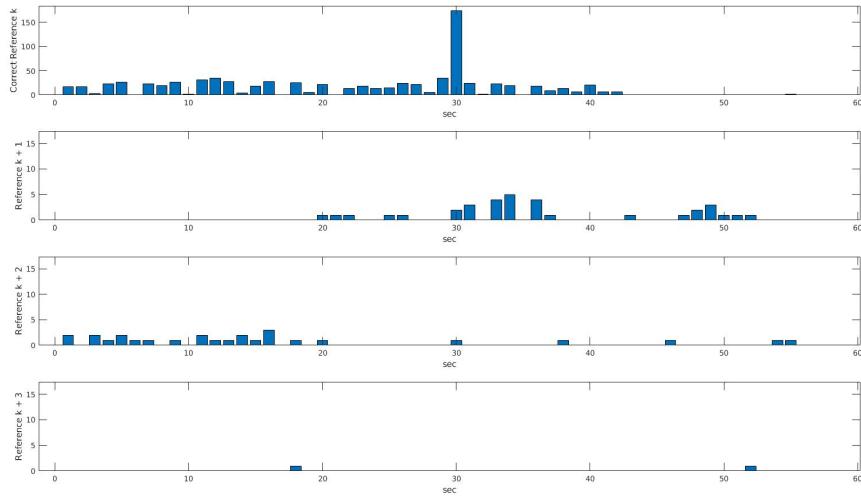


(β') Φασματογράφημα Αποσπάσματος με τα τοπικά μέγιστα.

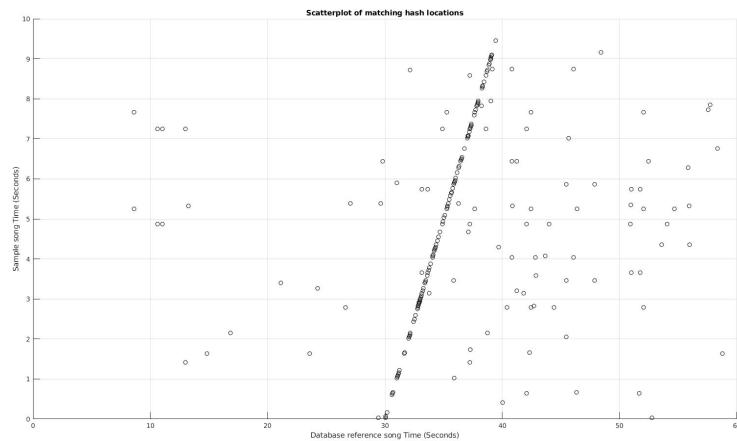
Σχήμα 4.2: Μηχανισμός δημιουργίας κλειδιών.

που θα εμφανίζονται στο u θα πρέπει να βρεθούν στο r_0 . Πιο συγκεκριμένα, το κλειδί k με χρόνο εμφάνισης $t_{k,u}$ στο u θα πρέπει να βρεθεί στο r_0 στον χρόνο $t_{k,r_0} = t_{k,u} + d$.

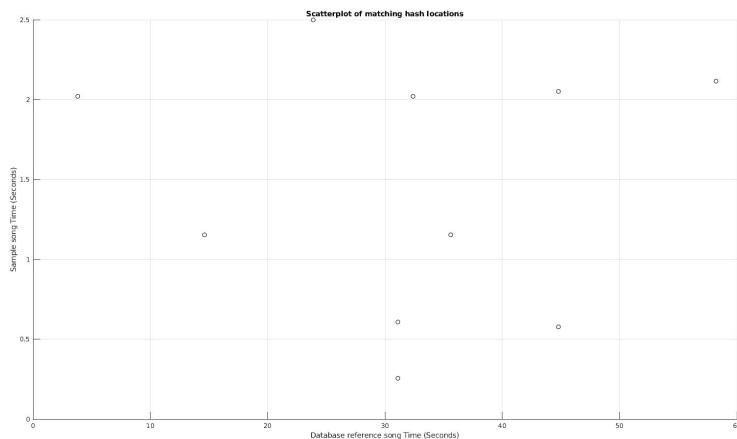
Αν μελετηθούν οι τιμές $\{t_{k,r_0} - t_{k,u}\}$ για κάθε κλειδί k του άγνωστου απόσπασμας, θα πρέπει να υπάρχει μέγιστη συσσώρευση γύρω από την τιμή d . Προτείνεται από τον Wang πως για κάθε κομμάτι αναφοράς r_i , να αποθηκεύονται οι τιμές $\{t_{k,r_0} - t_{k,u}|k \text{ κλειδί από το άγνωστο απόσπασμα}\}$ σε ένα ιστόγραμμα. Άρα, το ιστόγραμμα με το υψηλότερο μέγιστο αντιστοιχεί στο καλύτερο ταίριασμα με το άγνωστο απόσπασμα. Επίσης, σύμφωνα με τα παραπάνω, το απόσπασμα θα ζεκινάει από το σημείο που βρίσκεται το μέγιστο του ιστογράμματος.



Σχήμα 4.3: Ιστογράμματα σωστού και λάθος κομματιών.



(α') Σωστό κομμάτι αναφοράς - Με διαγώνιο.



(β') Λάθος κομμάτι αναφοράς - Χωρίς διαγώνιο.

Σχήμα 4.4: Διαγράμματα χρονικής αντιστοίχησης κλειδιών.

5 Υλοποίηση

5.1 Υλοποίηση του αλγορίθμου Shazam

5.1.1 Εξαγωγή τοπικών μεγίστων από το Φασματογράφημα

Για να παραχθεί το φασματογράφημα, χρησιμοποιούνται συνεχόμενοι Διακριτοί Μετασχηματισμοί Fourier Σύντομου Χρόνου, σε πλαίσια μήκους $64ms$ με συνάρτηση παραθύρωσης τύπου Hamming και μέγεθος βήματος $32ms$, η σειριακή σύνδεση των οποίων δίνει το φασματογράφημα. Το επόμενο βήμα αποτελείται από την εξαγωγή των τοπικών μεγίστων, ικανοποιώντας παράλληλα ένα κριτήριο πυκνότητας, το οποίο ορίζει τις περιοχές στις οποίες θα γίνει η αναζήτηση του μεγίστου.

Χωρίζουμε το φασματογράφημα σε ορθογώνια πλάτους ΔT_{tile} δευτερολέπτων και ύψους ΔF_{tile} Hertz. Για κάθε ορθογώνια περιοχή, υπολογίζονται και κρατούνται οι συντεταγμένες του τοπικού μέγιστου σημείου. Με αυτό τον τρόπο έχουμε μια ομοιόμορφη αναπαράσταση των μεγίστων σημείων όλου του φασματογράμματος, ενώ ταυτόχρονα, η ρύθμιση των τιμών ΔT_{tile} , ΔF_{tile} επιτρέπει τον καθορισμό της πυκνότητας των επιλεγμένων κορυφών στον άξονα του χρόνου και της συχνότητας αντίστοιχα.

5.1.2 Εξαγωγή και κωδικοποίηση ζευγαριών κορυφών

Το επόμενο βήμα είναι ο υπολογισμός όλων των πιθανών ζευγαριών των κορυφών (τοπικών μεγίστων) του σήματος που υπολογίστηκαν. Για δύο κορυφές (t_1, f_1) και (t_2, f_2) , προτείνεται η κωδικοποίηση της μορφής $[f_1, f_2, t_2 - t_1]$. Για να αποφευχθεί η ραγδαία αύξηση του αριθμού των ζευγαριών, εφαρμόζεται ένας περιορισμός στην επιλογή των κορυφών που γίνονται ζευγάρια. Συγκεκριμένα, θεωρείται πως ζευγάρια γίνονται οι κορυφές εκείνες που το εύρος των συχνοτήτων τους $f_2 - f_1$ δεν ξεπερνά το άνω όριο ΔF_{max} και η χρονική διαφορά τους $t_2 - t_1$ δεν ξεπερνά το άνω όριο ΔT_{max} .

Σειρά έχει η κωδικοποίηση του κλειδιού $[f_1, f_2, t_2 - t_1]$ που μπορεί να επιτευχθεί σε 32 δυαδικά ψηφία. Συγκεκριμένα, χρησιμοποιούνται n_1 ψηφία για την κωδικοποίηση της f_1 , αντίστοιχα n_2 ψηφία για την f_2 , και n_3 ψηφία για την χρονική διαφορά Δt , έτσι ώστε να ικανοποιείται η συνθήκη $n_1 + n_2 + n_3 = 32$.

Οι τιμές του κλειδιού κανονικοποιούνται με τον εξής τύπο, με σκοπό την ομοιόμορφη αναπαράσταση τους σύμφωνα με το εύρος τιμών τους και τον πλήθος των ψηφίων της δυαδικής τους αναπαράστασης, όπου f_{max} η μέγιστη συχνότητα του φασματογράμματος ($f_{max} = f_s/2$):

$$\begin{aligned}\tilde{f}_1 &= \left\lfloor \frac{f_1}{f_{max}} (2^{n_1} - 1) \right\rfloor \\ \tilde{f}_2 &= \left\lfloor \frac{f_2}{f_{max}} (2^{n_2} - 1) \right\rfloor\end{aligned}$$

Λόγω της τεχνικής περιορισμού δημιουργίας ζευγαριών, γνωρίζουμε και το ΔT_{max} . Έτσι, εξασφαλίζεται πως δεν θα υπάρχει υπερχείλιση στα επιμέρους τμήματα του κλειδιού, με διακριτική ικανότητα ανά $\frac{\Delta t_{max}}{2^{n_3} - 1} sec$:

$$\widetilde{\Delta t} = \left\lfloor \frac{\Delta t}{\Delta t_{max}} (2^{n_3} - 1) \right\rfloor$$

Έχει σημασία να δοθεί προσοχή στην διακριτικότητα στις κωδικοποιημένες χρονικές διαφορές, διότι υπάρχει ένα θεωρητικό όριο που προκύπτει από το φασματογράφημα. Η χρονική διακριτικότητα του φασματογράφηματος είναι t_{hop} , αρά για την διαφορά Δt είναι $2t_{hop}$. Πρέπει επομένως να εξασφαλισθεί πως η κωδικοποιημένη διακριτικότητα των στοιχείων δεν ξεπερνά το θεωρητικό αυτό όριο, καθώς λόγω παραμορφώσεων (πχ, ο μετασχηματισμός Φουριέ σε αποκομένο απόσπασμα να μην συγχρονίζεται πλήρως με τον αντίστοιχο του κομματιού αναφοράς στην βάση δεδομένων). Άρα πρέπει να ισχύει:

$$\frac{\Delta t_{max}}{2^{n_3} - 1} < 2t_{hop}$$

Η τελική μορφή της δυαδικής αριθμητικής αναπαράστασης του κλειδιού προκύπτει ενώνοντας τα τρία στοιχεία. Η ένωση είναι ο αριθμός: $k = \tilde{f}_1 \times 2^{n_2+n_3} + \tilde{f}_2 \times 2^{n_3} + \widetilde{\Delta t}$, η οποία χωράει σε 32 δυαδικά ψηφία.

5.1.3 Αποθήκευση των κλειδιών στην βάση δεδομένων

Το στάδιο μάθησης όπως προτείνεται από το Wang είναι το εξής. Για κάθε κομμάτι αναφοράς, εξάγονται όλα τα κλειδιά του και κωδικοποιούνται στις δυαδικές μορφές τους. Έπειτα, κάθε κλειδί συνδέεται με την τιμή του, και καταγράφεται στην μηχανή δεικτών του συστήματος. Συγχεριμένα, η τιμή του κλειδιού περιέχει δύο κομμάτια πληροφορίας: τον χρόνο εμφάνισης του t_1 στο κομμάτι αναφοράς, και τον δείκτη του κομματιού αναφοράς, που πρακτικά μπορεί να είναι ένας μοναδικός ακέραιος αριθμός (σε αύξουσα σειρά στην λίστα της βάσης κομματιών). Επιπλέον, προτείνεται η κωδικοποίηση της τιμής του κλειδιού σε 32 δυαδικά ψηφία, με τον τρόπο που ακολουθήθηκε και στην κωδικοποίηση του κλειδιού.

Με την ολοκλήρωση της παραπάνω διαδικασίας, καταλήγουμε σε μια λίστα δεικτών που μπορεί να χειριστεί ερωτήματα των 32 δυαδικών ψηφίων που αποτελούνται από υποψήφια

κλειδιά και να επιστρέψει τις τιμές που αντιστοιχούν σε αυτά τα κλειδιά και υπάρχουν στην βάση. Σημαντικό είναι το γεγονός πως επειδή δεν αποκλείεται η εμφάνιση ενός κλειδιού σε παραπάνω από μία χρονικές στιγμές αλλά και κομμάτια αναφοράς, ο μηχανισμός θα πρέπει να είναι σε θέση να επιστρέψει πολλές τιμές όταν γίνεται αναζήτηση για ένα κλειδί. Επιπλέον, υπάρχει και το ενδεχόμενο ο μηχανισμός να αναζητεί και ένα κλειδί το οποίο δεν περιέχει και σε αυτήν την περίπτωση πρέπει να είναι σε θέση να επιστρέψει κενό σύνολο ως απάντηση. Καθώς το σύνολο των δεδομένων που είναι να αναγκαίο να αποθηκευτούν σε αυτόν τον μηχανισμό μπορεί να είναι αρκετά μεγάλο ανάλογα με την εφαρμογή, η επιλογή του πρέπει να γίνει με μεγάλη προσοχή ώστε το σύστημα να εύκολα επεκτάσιμο.

Η βάση δεδομένων που χρησιμοποιείται κατά την υλοποίηση είναι η Lightning Memory-mapped Database (LMDB) [22], η οποία είναι μια βάση δεδομένων ζευγαριών κλειδιών-τιμών και χρησιμοποιεί την δομή B-Tree. Έχει συνδεθεί στην Matlab μέσω του Mex API [23].

5.1.4 Διαχείριση απαντήσεων των ερωτημάτων της βάσης δεδομένων

Στην φάση αναγνώρισης, όλα τα ζεύγη κορυφών εξάγονται από το άγνωστο απόσπασμα, με την μεθοδολογία που προτάθηκε. Έπειτα, μετατρέπονται σε κλειδιά που χρησιμοποιούνται ως ερωτήματα για την αναζήτηση στη βάση δεδομένων. Όπως ειπώθηκε, οι έξοδοι της βάσης τοποθετούνται σε μια λίστα ιστογραμμάτων. Πρακτικά, για ένα κλειδί k που εξάγεται στον χρόνο t_u στο άγνωστο απόσπασμα, η βάση δεδομένων επιστέφει την τιμή v_i η οποία διαχωρίζεται στα δύο συστατικά της στοιχεία: τον δείκτη αναγνώρισης του μουσικού κομματιού r_i που περιέχει το κλειδί k , και την χρονική στιγμή t_{r_i} που βρέθηκε το κλειδί στο κομμάτι αναφοράς. Έπειτα, η τιμή $t_{r_i} - t_u$ αποθηκεύεται στο ιστόγραμμα που αντιστοιχεί στο r_i .

Τα ιστογράμματα χρησιμοποιούν την χρονική ανάλυση δt . Γνωρίζοντας πως τα κομμάτια αναφοράς έχουν ορισμένη χρονική διάρκεια L_{ref} δευτερολέπτων, προκύπτει πως τα ιστογράμματα αποτελούνται από $L_{ref} / \delta t$ θέσεις bins. Εν τέλει, το τελικό σετ ιστογραμμάτων είναι ένας στατικός πίνακας μεγέθους $L_{ref} \times \delta t$, ενώ, κάθε ταυτοποίηση κλειδιού αυξάνει κατά ένα, ένα κελί του πίνακα ιστογραμμάτων. Διαδικασία που επαναλαμβάνεται αρκετές φορές κατά την αναζήτηση στη βάση.

Τέλος, για να πραγματοποιήσουμε την αναγνώριση, γίνεται αναζήτηση του κομματιού αναφοράς, το ιστόγραμμα του οποίου, περιέχει το υψηλότερο μέγιστο σε όλο τον πίνακα. Το κομμάτι αυτό θεωρείται πως είναι το ταίριασμα του άγνωστου αποσπάσματος, και η θέση που βρίσκεται το μέγιστο στο ιστόγραμμα το σημείο εκκίνησης του αποσπάσματος.

5.1.5 Σύντηξη των τοπικών αποφάσεων

Όποιο και αν είναι το άγνωστο απόσπασμα, το προηγούμενο βήμα επιστρέφει το καλύτερο ταίριασμα με τον χρόνο εκκίνησης του. Αυτό σημαίνει πως η περίπτωση στην οποία σε μια αναζήτηση δεν υπάρχει καμία αντιστοιχία κλειδιών δεν έχει ληφθεί ακόμα υπόψιν.

Η τεχνική που προτείνεται από τον Wang είναι ο καθορισμός ενός μηχανισμού κάτω ορίου. Συγκεκριμένα, να τεθεί ένα κάτω όριο στον αριθμό των κλειδιών που ταυτίζονται μεταξύ του απόσπασματος και του βέλτιστου κομματιού αναφοράς, όπως αυτό ορίστηκε παραπάνω. Η σύγκριση στην ουσία γίνεται μεταξύ του μέγιστου σημείου στο ιστόγραμμα του κομματιού και του κάτω ορίου, και αν η συνθήκη ικανοποιείται, τότε το αποτέλεσμα θεωρείται αποδεκτό, ενώ όχι στην αντίθετη περίπτωση, με την προβολή ανάλογης εξόδου από το σύστημα. Όμως, ο καθορισμός του εν λόγω κάτω ορίου είναι πολύ δύσκολος, καθώς ο αριθμός των ταιριασμάτων κλειδιών είναι σχεδόν αδύνατο να προβλεφθεί όταν το απόσπασμα που αναζητείται έχει υποστεί διάφορες παραμορφώσεις. Αυτό συμβαίνει διότι, λόγω των παραμορφώσεων, ο αριθμός των ταιριασμάτων είναι αρκετά μικρός, ακόμα και αν η αναγνώριση είναι επιτυχημένη, το οποίο και κατ' επέκταση σημαίνει πως η κατανομή των εσφαλμένων αναγνωρίσεων έχει μεγάλη επικάλυψη με την κατανομή των σωστών, όσον αφορά τον αριθμό των ζευγαριών ταυτοποίησης. Τέλος, στην επιλογή του κάτω ορίου πρέπει να ληφθεί υπόψιν και το κανάλι μετάδοσης, γεγονός που σημαίνει πως πρέπει να γίνεται ξεχωριστή ρύθμιση για κάθε διαφορετική εφαρμογή.

Για αυτό, και προτείνεται μια διαφορετική προσέγγιση[8]. Σχοπός είναι να ικανοποιηθεί το κενό που υπάρχει στην τεχνική του Wang, και ταυτόχρονα να υπάρχει μεγαλύτερη ευελιξία στην διαδικασία λήψης της τελικής εξόδου του συστήματος, που επιτυγχάνεται με την ακόλουθη τεχνική. Το άγνωστο απόσπασμα u χωρίζεται σε υποσήματα u_i^{sub} , με μήκος l^{sub} και ποσοστό επικάλυψης o^{sub} . Ορίζονται κατ' αυτόν τον τρόπο P συνεχόμενα υποσήματα $o_{j=1 \dots P}^{sub}$, και θένεται από τα οποία περνάει από όλη την διαδικασία αναγνώρισης όπως παρουσιάστηκε παραπάνω, με έξοδο της μορφής (r_i, s_i) από το σύστημα, όπου r_i το βέλτιστο υποψήφιο κομμάτι αναγνώρισης και s_i τον χρόνο εκκίνησης του απόσπασματος στο r_i . Για να θεωρηθεί σωστή η αναγνώριση του απόσπασματος, πρέπει η αναγνώριση να έχει συνοχή στα αποτελέσματα για πάνω από T_{vote} υποσήματα, ενώ στην αντίθετη περίπτωση θεωρείται πως το απόσπασμα δεν βρέθηκε. Η συνοχή των αποτελεσμάτων δύο υποσημάτων (r_i, s_i) και (r_j, s_j) ορίζεται ως εξής:

$$\left| \begin{array}{l} r_i = r_j \\ s_i - iL^{sub}(1 - o^{sub}) = s_j - jL^{sub}(1 - o^{sub}) \end{array} \right.$$

Οι συνοχή, δηλαδή, ορίζεται η ταύτιση του βέλτιστου υποψήφιου κομματιού αναφοράς προς ταυτοποίηση και η σωστή αντιστοίχηση στους χρόνους εκκίνησης των υποσημάτων μέσα σε αυτό.

Η παρόμετρος T_{vote} ορίζει και το επίπεδο ευαισθησίας του συστήματος. Μπορεί να πάρει τιμές από 0 μέχρι και P . Αν $T_{vote} = P$, τότε όλα τα αποτελέσματα πρέπει να

έχουν συνοχή και με αυτό τον τρόπο περιορίζονται οι πιθανότητες λάθος αποτελεσμάτων, αλλά ταυτόχρονα δεν εμφανίζονται αποτελέσματα που έχουν αρκετές πιθανότητες σωστής αναγνώρισης. Από την άλλη αν το T_{vote} έχει χαμηλή τιμή, ισχύουν τα αντίστροφα. Προτείνεται πως μια καλή τιμή για το T_{vote} είναι:

$$T_{vote} = \lceil \frac{P}{2} \rceil$$

5.1.6 Επισκόπηση Αλγορίθμου

Η συμπεριφορά του αλγορίθμου του Wang [7], μαζί με την προσθήκη της σύντηξης των τοπικών αποφάσεων από το [8], μπορεί να ρυθμιστεί από το σύνολο παραμέτρων που υπάρχουν στον παρακάτω πίνακα:

Όνομα	Περιγραφή	Τιμή
L_{FFT}	Μήκος των παραθύρων του FFT	64ms
t_{hop}	Βήμα μεταξύ των παραθύρων του FFT	32ms
L_{sub}	Μήκος των υποσημάτων του αποσπάσματος	5s
o_{sub}	Ποσοστό επικάλυψης μεταξύ διαδοχικών υποσημάτων του αποσπάσματος	0.5
ΔT_{tile}	Πλάτος των κελιών κατά το χώρισμα του φασματογραφήματος (χρόνος)	0.4s
ΔF_{tile}	Τύψος των κελιών κατά το χώρισμα του φασματογραφήματος (Συχνότητα)	400Hz
ΔT_{max}	Μέγιστη χρονική διαφορά μεταξύ υποψήφιων ζευγών τοπικών μεγίστων στο φασματογράφημα	3s
ΔF_{max}	Μέγιστη συχνοτική διαφορά μεταξύ υποψήφιων ζευγών τοπικών μεγίστων στο φασματογράφημα	350Hz
n_1	Πλήθος δυαδικών ψηφίων κατά την κωδικοποίηση της f_1	13
n_2	Πλήθος δυαδικών ψηφίων κατά την κωδικοποίηση της f_2	13
n_3	Πλήθος δυαδικών ψηφίων κατά την κωδικοποίηση του $t_2 - t_1$	6
δt	Χρονική ανάλυση των ιστογραμμάτων	1s
P	Πλήθος υποσημάτων που χρησιμοποιούνται κατά το βήμα της σύντηξης των τοπικών αποφάσεων	6
T_{vote}	Πλήθος ελάχιστων απαιτούμενων σωστών ταιριασμάτων υποσημάτων, για να θεωρηθεί η αναγνώριση επιτυχημένη	3

Πίνακας 5.1: Πίνακας παραμέτρων αλγορίθμου Shazam

5.2 Βελτίωση του Αλγορίθμου Shazam

5.2.1 Βελτίωση Ευελιξίας στην Μετατόπιση Τονικότητας με την χρήση του CQT

Όπως φαίνεται και παραχάτω στα αποτελέσματα της πειραματικής διαδικασίας, η παραπάνω μέθοδος δεν είναι ευέλικτη όσον αφορά την παραμόρφωση της μετατόπισης της τονικότητας του σήματος, που μπορεί να θεωρηθεί και ως αλλαγή της συχνότητας δειγματοληψίας του σήματος. Η ανάλυση του μοντέλου ακουστικού αποτυπώματος δείχνει πως για τα αποτελέσματα αυτά ευθύνεται η μοντελοποίηση των συχνοτικών σημείων στην τριάδα στοιχείων που αποτελούν το κλειδί του αποτυπώματος.

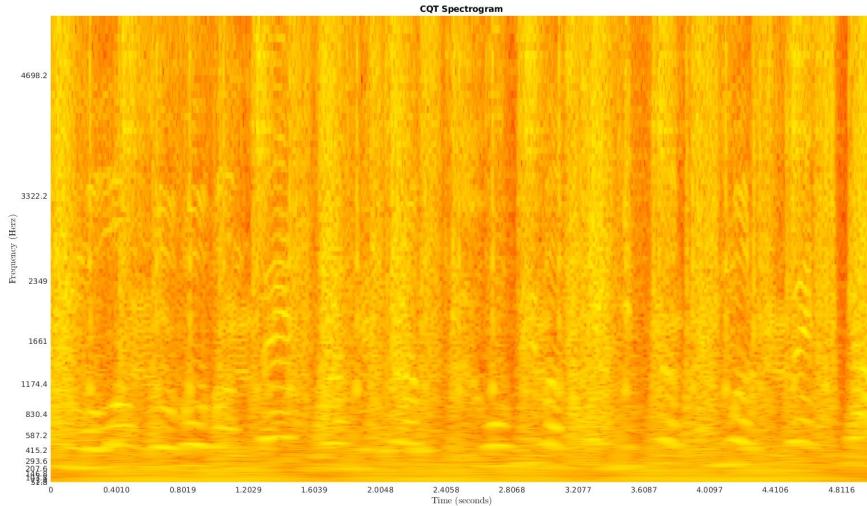
Η μέθοδος του Wang χρησιμοποιεί τον Μετασχηματισμού Φουριέ Σύντομου Χρόνου για να παράξει το φασματογράφημα στο οποίο βασίζεται και το μοντέλο του αποτυπώματος. Το εργαλείο αυτό δίνει την ποσότητα της ενέργειας ενός σήματος για κάθε συχνοτικό εύρος. Όμως, τα συχνοτικά εύρη που εξάγει είναι γραμμικά κατανεμημένα. Η κατανομή αυτή δεν προσαρμόζεται πολύ αποτελεσματικά στις μουσικές εφαρμογές, διότι η συχνοτική κατανομή των μουσικών νοτών ακολουθεί γεωμετρική κατανομή, αντί γραμμικής. Το γεγονός αυτό εξηγεί γιατί η τονική μετατόπιση, που μουσικά μεταφράζεται σε μετατόπιση νοτών κατά μια σταθερά, αντιστοιχεί σε πολλαπλασιασμό των συχνοτήτων.

Για να αντιμετωπισθεί αυτό το πρόβλημα, στο [8] προτείνεται η αντικατάσταση του STFT με τον Μετασχηματισμό Σταθερού-Q (CQT [20]). Ο CQT δίνει εξίσου το ποσό της ενέργειας του σήματος για κάθε συχνοτική μπάντα, όμως η κατανομή των συχνοτήτων ακολουθεί γεωμετρική κατανομή. Λόγω αυτού, η μετατόπιση της τονικότητας του σήματος μεταφράζεται σε μετατόπιση σε συχνοτικές μπάντες. Έτσι, ένα σήμα που έχει ενέργεια στην συχνοτική μπάντα b , όταν μετατοπιστεί τονικά, η ενέργεια αυτή θα μετακινηθεί στην συχνοτική μπάντα $b + k$.

Όπως έχει προαναφερθεί, χρησιμοποιώντας τον CQT μπορεί να γίνει μια ακριβής ταύτιση των συχνοτικών ζωνών του με τις δυτικές μουσικές νότες. Για να γίνει αυτό, πρέπει η αρχική συχνοτική ζώνη f_0 του μετασχηματισμού να αντιστοιχεί στην συχνότητα μιας πραγματικής νότας. Σε αυτή την τροποποίηση του ακουστικού αποτυπώματος, χρησιμοποιούμε τον CQT με 36 φασματικές μπάντες ανά οκτάβα (3 μπάντες ανά νότα) και αρχική συχνότητα f_0 την νότα $G_1\#$ ($51.91Hz$).

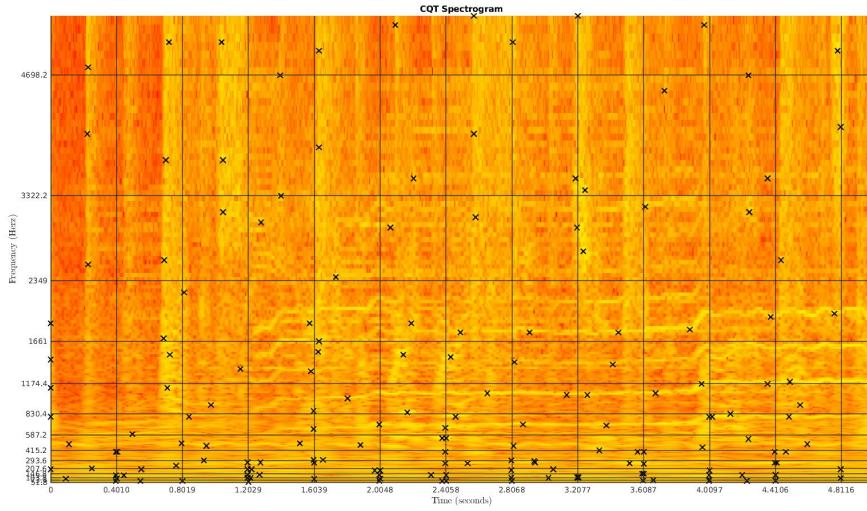
Εδώ, χρειάζεται μια σημαντική σημείωση για το φασματογράφημα που παράγεται από τον CQT και έχει να κάνει με την μορφή του. Αν και οι χρονικές θέσεις στις οποίες υπολογίζονται οι τιμές $X^{CQ}(k, n)$ είναι ίδιες στα πλαίσια μιας οκτάβας, οι θέσεις αυτές υποδιπλασιάζονται ανά οκτάβα. Λόγω αυτής της ιδιομορφίας, η επεξεργασία του φασματογραφήματος είναι πιο δύσκολη από το κλασσικό φασματογράφημα. Για να εξαλειφθεί αυτός ο παράγοντας, στο toolbox που χρησιμοποιείται στην υλοποίηση [20], περιλαμβάνεται επιλογή εμφάνισης του φασματογράμματος στην κλασσική μορφή του φασματογράμματος. Η εμφάνιση αυτή επιτυγχάνεται μέσω παρεμβολής τιμών στα σημεία

του CQT που χρειάζεται έτσι ώστε το πλήθος των τιμών ανά γραμμή να είναι ίδιο για όλο το αποτέλεσμα. Αυτό εξηγεί και γιατί οι χαμηλότερες οκτάβες του φασματογραφήματος φαίνονται σαν να είναι ‘τραβηγμένες’ κατά τον άξονα του χρόνου.



Σχήμα 5.1: Φασματογράφημα του Μετασχηματισμού Σταθερού-Q.

Τα επόμενα βήματα της τροποποίησης είναι αρχετά παρόμοια με την αρχική μέθοδο. Το φασματογράφημα που παράγεται διαιρέίται σε ορθογώνιες περιοχές με μήκος ΔT_{tile} δευτερόλεπτα και ύψος ΔB_{tile} συχνοτικές μπάντες. Σε κάθε περιοχή υπολογίζεται η μέγιστη τιμή και αποθηκεύονται οι συντεταγμένες της. Έπειτα, εξάγονται τα ζευγάρια τοπικών μεγίστων που ικανοποιούν τον χωρικό περιορισμό, όπως αυτός ορίστηκε και στον αρχικό αλγόριθμο. Για δυο σημεία με συντεταγμένες $(t_1, b_1), (t_2, b_2)$ στο φασματογράφημα, πρέπει δηλαδή να ισχύει: $t_2 - t_1 \leq \Delta t_{max}$ και $b_2 - b_1 \leq \Delta b_{max}$. Οι τιμές των συντεταγμένων του ζευγαριού χρησιμοποιούνται για να δημιουργηθεί μια τριάδα τιμών που είναι και το τελικό αλειδί του ακουστικού αποτυπώματος.



Σχήμα 5.2: Απόσπαση τοπικών μεγίστων από το φασματογράφημα του Μετασχηματισμού Σταυθερού-Q.

Η κωδικοποίηση που προτείνεται είναι η εξής:

$$[\hat{b}_1; b_2 - b_1; t_2 - t_1]$$

με $\hat{b}_1 = \lfloor \frac{b_1}{6} \rfloor$, μια υποδιαιρεμένη εκδοχή του b_1 .

Η τριάδα αριθμών είναι παρόμοια με τον αρχικό αλγόριθμο, όμως υπάρχουν κάποιες τροποποιήσεις. Στόχος είναι να επιλεχθεί πληροφορία που είναι ευέλικτη σε παραμορφώσεις και ιδίως στην μετατόπιση τονικότητας. Το πρώτο στοιχείο της αναπαράστασης έχει επιλεγεί έτσι ώστε να δίνει μια χοντροκομμένη εκτίμηση της τοποθεσίας του ζευγαριού στο πεδίο των συχνοτήτων. Για να γίνει αυτό, θα μπορούσε κανείς να επιλέξει την συχνοτική πληροφορία του πρώτου σημείου b_1 , ή του δεύτερου b_2 ή και του μέσου όρου τους $mean(b_1, b_2)$. Όμως η διαχριτική ικανότητα των συχνοτήτων είναι πολύ ακριβής για να είναι ευέλικτη σε περίπτωση τονικής μετατόπισης ακόμα και σε μικρά επίπεδα, κατά την διαδικασία της αναζήτησης.

Σε αυτές τις περιπτώσεις, μια καλή και απλή λύση για να αυξηθεί η ευελιξία των παραμέτρων τέτοιου τύπου, είναι να χρησιμοποιηθεί μια υποδιαιρεση της παραμέτρου. Αυτό φυσικά, μειώνει την ακρίβεια της διαχριτικής ικανότητας της παραμέτρου, αλλά στην περίπτωση αυτή, δεν μας ενοχλεί αυτό, καθώς το κλειδί περιέχει και 2 ακόμα τιμές. Τελικά, η πρώτη παραμέτρος καταλήγει να είναι το $\hat{b}_1 = \lfloor \frac{b_1}{6} \rfloor$. Λαμβάνοντας υπόψιν την ανάλυση του CQT που έχει επιλεγεί (3 μπάντες ανά νότα), η υποδιαιρεση αυτή σημαίνει πως κάθε διαδοχικό νούμερο στο τελικό πρώτο στοιχείο του κλειδιού καλύπτει δύο διαδοχικές νότες της χρωματικής κλίμακας, που είναι και αρκετό για συνηθισμένες χρήσεις τονικής μετατόπισης.

Το δεύτερο στοιχείο του κλειδιού είναι το συχνοτικό εύρος των δυο σημείων του ζευγαριού. Τα δυο πρώτα στοιχεία $[b_1; b_2 - b_1]$ αντιστοιχούν στα δυο στοιχεία του πρωτότυπου αλγορίθμου $[f_1; f_2]$. Πρωτικά, τα δυο αυτά σετ πληροφορίας περιέχουν συνδυασμούς δυο ορθογώνιων μεταβλητών πληροφορίας: την απόλυτη συχνότητα του ζευγαριού και την σχετική διαφορά συχνότητας μεταξύ των δυο σημείων.

Το τρίτο και τελευταίο στοιχείο πληροφορίας του κλειδιού είναι το χρονικό εύρος του ζευγαριού, το οποίο και μένει ίδιο με την αναπαράσταση στον αρχικό αλγόριθμο. Καθώς περέχει μόνο σχετική χρονική πληροφορία, είναι ευέλικτο στην αποκοπή αποσπάσματος από το κομμάτι αναφοράς. Επίσης, όπως είδαμε, αν η αναπαράσταση του $t_2 - t_1$ έχει αρκετά χαμηλή διακριτικότητα, η ποσότητα αυτή είναι ευέλικτη στην τονική μετατόπιση.

Οι υπόλοιπες επιμέρους διαδικασίες του αλγορίθμου παραμένουν ίδιες: απομόνωση των τοπικών μεγίστων του φασματογραφήματος, χρήση κριτηρίων απόστασης και στον χρόνο και στην συχνότητα, αποκλεισμός της απόλυτης χρονικής πληροφορίας. Αυτό σημαίνει πως η καινούρια αναπαράσταση κληρονομεί και την ευελιξία που έχει προς τις άλλες παραμορφώσεις (προσθήκη θορύβου, ισοστάθμιση, συμπίεση δυναμικού εύρους, επιλογή τυχαίου αποσπάσματος).

Όσον αφορά την μετατόπιση τονικότητας, μπορούμε να δούμε τα εξής. Ένα σήμα που έχει ένα ζευγάρι σημείων με συντεταγμένες (t_1, b_1) και (t_2, b_2) θα παραμορφωθεί με τέτοιο τρόπο, έτσι ώστε τα σημεία του να μετακινηθούν στις θέσεις $(t_1, b_1 + \kappa)$ και $(t_2, b_2 + \kappa)$. Το κλειδί αυτού του ζευγαριού εν τέλει μετατρέπεται σε αυτό:

$$\widehat{[b_1 + \kappa; (b_2 + \kappa) - (b_1 + \kappa); t_2 - t_1]} =$$

Με δεδομένη χαμηλή διακριτική ανάλυση στην αναπαράσταση του πρώτου συστατικού στοιχείου, τις περισσότερες φορές ισχύει:

$$\widehat{b_1 + \kappa} = \widehat{b_1}$$

Αξίζει να αναφερθεί πως η μετατόπιση τονικότητας θα μεταφέρει κάποιες τιμές από τα όρια των νέων συχνοτικών κελιών $\widehat{b_i}$ στο επόμενο. Όμως, όπως και στον αρχικό αλγόριθμο δεν απαιτείται ακριβές ταίριασμα μεταξύ όλων των ζευγαριών. Όπως είδαμε το βήμα της συγκέντρωσης τιμών στα ιστογράμματα διατηρείται όμοιο, το οποίο απαιτεί μόνο την διατήρηση της πλειοψηφίας των ζευγαριών. Επειδή τα συχνοτικά κελιά $\widehat{b_i}$ είναι αρκετά μεγάλα, θεωρείται πως η ποσότητα των οριακών τιμών είναι στατιστικά πολύ μικρή για να μεταβάλει τα τελικά αποτελέσματα της αναγνώρισης.

Επίσης, φαίνεται και πως το δεύτερο στοιχείο παραμένει αμετάβλητο, αφού:

$$(b_2 + \kappa) - (b_1 + \kappa) = b_2 - b_1$$

Όταν η παραμόρφωση συνοδεύεται και από χρονική παραμόρφωση στα πλαίσια της αλλαγής συχνότητας δειγματοληψίας, η χρονική αυτή παραμόρφωση απορροφάται από την

διαχριτικότητα που χρησιμοποιείται στην αναπαράσταση του $t_2 - t_1$, όπως φαίνεται και στα αποτελέσματα της πειραματικής διαδικασίας του αρχικού αλγορίθμου παρακάτω.

Επομένως, το ανανεωμένο κλειδί που προτείνεται φαίνεται να είναι ευέλικτο στα πλαίσια της μετατόπισης τονικότητας, της χρονικής παραμόρφωσης, της αλλαγής συχνότητας δειγματοληψίας και ταυτόχρονα διατηρεί τα πλεονεκτήματα που κληρονομεί από τον πρωτότυπο αλγόριθμο του Wang.

Όνομα	Περιγραφή	Τιμή
fs	Συχνότητα δειγματοληψίας	11,025Hz
r	Πλήθος συχνοτικών ζωνών ανά νότα	3
B	Πλήθος συχνοτικών ζωνών ανά οκτάβα	36
f_{min}	Συχνότητα ελάχιστης συχνοτικής ζώνης	51.91Hz
f_{max}	Συχνότητα μέγιστης συχνοτικής ζώνης	5,512.5Hz
Δt_{tile}	Πλάτος κελιών στον χρόνο	0.4sec
ΔB_{tile}	Τύφος κελιών στη συχνότητα ανά συχνοτικές ζώνες	18bins
ΔT_{max}	Μέγιστη χρονική διαφορά μεταξύ υποψήφιων ζευγών τοπικών μεγίστων στο φασματογράφημα	1.2sec
ΔF_{max}	Μέγιστη συχνοτική διαφορά μεταξύ υποψήφιων ζευγών τοπικών μεγίστων στο φασματογράφημα	24bins

Πίνακας 5.2: Πίνακας παραμέτρων αλγορίθμου με CQT

5.2.2 Μείωση του Όγκου Κλειδιών στη Βάση Δεδομένων - Σημαντικότητα Κλειδιού

Όπως έχει αναφερθεί, το μοντέλο της βάσης δεδομένων που χρησιμοποιείται είναι ζευγαριών κλειδιού-τιμών και είναι δομημένη με τέτοιο τρόπο έτσι ώστε να επιτρέπεται ένα κλειδί να περιέχει πολλές διαφορετικές τιμές.

Έχει ενδιαφέρον να συγχρίνει κανές τη δομή αυτής της βάσης δεδομένων με το γλωσσάρι που περιέχεται στο τέλος κάποιου βιβλίου. Οι λέξεις που αναφέρονται, περιέχουν δείκτες, δηλαδή τον αριθμό της σελίδας, που περιέχεται η λέξη-κλειδί. Με τον ίδιο τρόπο, στην περίπτωση μας, τα κλειδιά περιέχουν δείκτες που αναφέρουν τον μουσικό τίτλο στον οποίο αναφέρονται. Συνεχίζοντας τη σύγκριση αυτή, παρατηρούμε πως στα γλωσσάρια, δεν περιέχεται κάθε λέξη του βιβλίου. Οι περισσότερες λέξεις επαναλαμβάνονται σε βαθμό τέτοιο που δεν έχει νόημα να βρίσκονται στο γλωσσάρι σε αντίθεση με κάποιους τεχνικούς συγκεκριμένους όρους. Είναι πλέον εύλογο να εξεταστεί το ενδεχόμενο του να υπάρχουν κλειδιά που περιέχονται σε σημαντικό ποσοστό του σύνολου των μουσικών τίτλων της βάσης δεδομένων, και πως αν όντως υπάρχουν αυτά τα κλειδιά, περιέχουν ελάχιστη πληροφορία, άρα και μπορούν να διαγραφούν από την βάση.

Ένας τρόπος διερεύνησης του ζητήματος είναι να μελετηθεί η κατανομή του αριθμού

αναφορών μουσικών τίτλων ανά κλειδί. Το αποτέλεσμα του πλήθους των αναφορών για ένα κλειδί αποτελεί μια εκτίμηση της τυχαίας μεταβλητής αριθμός αναφορών. Επαναλαμβάνοντας την διαδικασία αυτή για όλα τα κλειδιά της βάσης και αποθηκεύοντας τα αποτελέσματα σε ένα κανονικοποιήμενο ιστόγραμμα, προχύπτει μια εκτίμηση της συνάρτησης πυκνότητας-πιθανότητας του αριθμού αναφορών.

Όπως φαίνεται και στο ανάλογο υποκεφάλαιο της πειραματικής διαδικασίας, προχύπτει πως ένα σεβαστό ποσοστό των κλειδιών της βάσης αναφέρονται αρκετά υψηλό αριθμό αναφορών, και όπως προαναφέρθηκε, φαίνεται συνετό να θεωρηθεί πως αυτά τα κλειδιά δεν βοηθούν στην διαδικασία αναγνώρισης. Προτείνεται λοιπόν ένα επιπλέον βήμα μείωσης του όγκου των κλειδιών της βάσης, το οποίο θα βοηθήσει στην βελτίωση της πολυπλοκότητας όλης της διαδικασίας.

Στο [8] προτείνεται η εξής διαδικασία: Για κάθε κλειδί k της βάσης δεδομένων, ονομάζουμε N_k τον αριθμό των αναφορών στις οποίες το κλειδί εμφανίζεται τουλάχιστον μια φορά. Ορίζουμε την σημαντικότητα ενός κλειδιού k ως:

$$s(k) = \frac{N - N_k}{N}$$

όπου N το συνολικό πλήθος των μουσικών τίτλων της βάσης δεδομένων. Στην ουσία, ένα κλειδί που εμφανίζεται σε πολλούς τίτλους έχει χαμηλή σημαντικότητα, ενώ αντίστοιχα, ένα σπάνιο κλειδί έχει υψηλή σημαντικότητα.

Στην προτεινόμενη μέθοδο, όταν ένα κλειδί αποσπάται από ένα άγνωστο απόσπασμα προς αναγνώριση, όλα τα ιστογράμματα των μουσικών τίτλων που περιέχουν το κλειδί ανανεώνονται. Αφού τα κλειδιά με χαμηλή σημαντικότητα αναφέρονται σε πολλούς τίτλους, θα πρέπει να γίνει ένας σημαντικός αριθμός ανανεώσεων, που συνεπάγεται και αντίστοιχη υπολογιστική ισχύ. Επιπλέον, αφού το εν λόγω κλειδί αναφέρεται τόσο συχνά σε μουσικούς τίτλους, υπάρχει και μεγάλη πιθανότητα να υπάρχει και στο άγνωστο απόσπασμα. Άρα, αυτά τα κλειδιά προκαλούν αρκετά περισσότερους υπολογισμούς από τα κλειδιά με χαμηλό συντελεστή σημαντικότητας.

Καταλήγοντας, η διαδικασία απλοποίησης της βάσης γίνεται διατηρώντας κάθε κλειδί για το οποίο ισχύει $s(k) < T_s$ και διαγράφοντας από αυτή όλα τα άλλα, όπου T_s κάτω όριο του συντελεστή σημαντικότητας. Η τιμή του T_s πρέπει να είναι τέτοια που να εξασφαλίζεται πως ο αριθμός των κλειδιών που απομένει στην βάση είναι αρκετός για να μπορεί να επιτευχθεί η διαδικασία σωστής αναγνώρισης, το οποίο εξαρτάται από την στατιστική κατανομή των κλειδιών, όπως αυτή προαναφέρθηκε. Όπως φαίνεται και στην πειραματική διαδικασία, με μια τυπική τιμή $T_s = 0.5$, επιτυγχάνεται σημαντικό κέρδος στην υπολογιστική πολυπλοκότητα ενώ ταυτόχρονα τα αποτελέσματα είναι στα ίδια ποσοστά επιτυχίας.

6 Πειραματική Διαδικασία

6.1 Περιγραφή πλαισίου πειραμάτων - Καθορισμός χριτηρίων απόδοσης

Η υλοποίηση των αλγορίθμων και η εκτέλεση της πειραματικής διαδικασίας έγινε στην Matlab R2017b-64bit, στο λειτουργικό σύστημα Ubuntu 16.04 και σύστημα με επεξεργαστή Intel i5-7200U @ 2.50GHz και 8GB RAM.

Για να μελετηθεί η απόδοση του αλγορίθμου, κρίνεται αναγκαία η εκτέλεση μιας σειράς πειραμάτων στην οποία να εξετάζονται αρκετές πτυχές του.

Για τον έλεγχο της απόδοσης κατά την αναγνώριση μουσικών κομματιών, δημιουργούμε μια βάση δεδομένων από 1000 μουσικά κομμάτια, από διαφορετικά είδη μουσικής. Τα αρχεία αυτά είναι της μορφής .wav, και η προεπεξεργασία των αρχείων περιλαμβάνει τις απαραίτητες διαδικασίες ώστε να καταλήξουν να είναι της μορφής: μονοχάναλα αρχεία(mono) των 16bit/sample με συχνότητα δειγματοληψίας fs στα $11.025Hz$. Από κάθε κομμάτι όμως χρησιμοποιήσουμε τα 60 πρώτα δευτερόλεπτα, τα οποία όμως περάσουν από όλη την διαδικασία της εξαγωγής ακουστικού αποτυπώματος και όμως αποθηκευτούν στην βάση δεδομένων.

Η διαδικασία της αναγνώρισης αποτελείται από την δημιουργία μιας τεχνητής ροής μουσικών κομματιών. Κάθε κομμάτι παίζεται για σταθερή χρονική διάρκεια, και χωρίζεται σε πλαίσια μήκους $L_{frame} = L_{sub} + (P - 1) * L_{sub0sub}$, δηλαδή όσο μήκος χρειάζεται για να τρέξει μια αναγνώριση ανάλογα με τις ρυθμίσεις του αλγορίθμου, χωρίς διαδοχική επικάλυψη. Τα πλαίσια εξετάζονται ζεχωριστά, έτσι στο κάθε κομμάτι να υπάρχει ένα σύνολο από διαφορετικά αποτελέσματα αναγνώρισης.

6.2 Περιγραφή του Audio Degradition Toolbox

Για την δημιουργία των παραμορφώσεων αλλά κυρίως της εξομοίωσης πραγματικών συνθηκών αναπαραγγής του ήχου χρησιμοποιείται και το Audio Degradition Toolbox[11], το οποίο παρέχει συναρτήσεις που υλοποιούν μονάδες παραμορφώσεων και σε δεύτερο στάδιο, συνδυασμό αυτών που εξομοιώνουν πραγματικές συνθήκες. Συγκεκριμένα, στην πειραματική διαδικασία χρησιμοποιήθηκαν οι εξής συνθήκες:

Ζωντανή Ηχογράφηση

Τλοποιείται με τον συνδυασμό 2 μονάδων παραμόρφωσης. Αρχικά, γίνεται συνέλιξη με την χρονική απόχριση ενός μεγάλου δωματίου και στην συνέχεια προστίθεται ελαφρύς ροζ θόρυβος.

Ραδιοφωνική Αναμετάδοση

Τλοποιείται με τον συνδυασμό 2 μονάδων παραμόρφωσης. Αρχικά, εφαρμόζεται συμπίεση του δυναμικού εύρους του σήματος ώστε να εξομοιώθει το χαρακτηριστικό της υψηλής έντασης των περισσότερων ραδιοφωνικών σταθμών. Στη συνέχεια εφαρμόζεται χρονική επιτάχυνση του σήματος κατά 2%, όπως συνήθως γίνεται έτσι ώστε να δημιουργηθεί χρόνος για περισσότερες διαφημίσεις.

Αναπαραγωγή μέσω Κινητού

Τλοποιείται με τον συνδυασμό 2 μονάδων παραμόρφωσης. Αρχικά, εφαρμόζεται η χρονοστική απόκριση του ηχείου του μοντέλου Google Nexus One, το οποίο έχει ομοιότητες με υψηλότερο φίλτρο με συχνότητα αποκοπής περίπου στα $500Hz$ και έπειτα το σήμα περνάει από ελαφρύ ροζ θόρυβο.

Ηχογράφηση μέσω Κινητού

Τλοποιείται με τον συνδυασμό 4 μονάδων παραμόρφωσης. Αρχικά, εφαρμόζεται η χρονοστική απόκριση του μικροφώνου του μοντέλου Google Nexus One. Έπειτα, το σήμα περνάει από συμπίεση δυναμικού εύρους, ώστε να εξομοιώθει το σύστημα αυτόματης ενίσχυσης έντασης του κινητού. Ακολουθεί το ‘χλιπάρισμα’ του σήματος σε ποσοστό 3%. Τέλος, προστίθεται ροζ θόρυβος μέτριας έντασης.

Αναπαραγωγή από Βινύλιο

Τλοποιείται με τον συνδυασμό 4 μονάδων παραμόρφωσης. Αρχικά, εφαρμόζεται η χρονοστική απόκριση ενός τυπικού μοντέλου πικάπ. Έπειτα, προστίθεται ο χαρακτηριστικός ‘ζεστός’ ήχος του βινυλίου στο σήμα μέσω ζεχωριστού αρχείου που παρέχεται στο toolbox. Στη συνέχεια, το σήμα περνάει από την μονάδα ”WoW Resample”, η οποία λειτουργεί σαν την κλασσική αλλαγή δειγματοληψίας, με την διαφορά πως ο συντελεστής της παραμόρφωσης είναι μεταβαλλόμενος. Η συχνότητα μεταβολής wow ρυθμίζεται έτσι ώστε να ταιριάζει με την ταχύτητα περιστροφής του βινυλίου στα $33rpm$. Τέλος, προστίθεται ελαφρύς ροζ θόρυβος.

6.3 Πειραματική Διαδικασία Αλγορίθμου Shazam

6.3.1 Απόδοση αλγορίθμου με καθαρό σήμα

Αρχικά, ενδιαφερόμαστε να μελετήσουμε την απόδοση του με τις παραμέτρους όπως αυτές ορίζονται στον Πίνακα 5.1, όπου ως είσοδος προς αναγνώριση είναι 500 κομμάτια που υπάρχουν στην βάση δεδομένων χωρίς την προσθήκη επιπλέον παραμορφώσεων.

Αναγνώριση στο αποθηκευμένο μέρος των κομματιών αναφοράς

Στο παρακάτω πείραμα, για κάθε πλαίσιο σε κάθε κομμάτι ορίζεται: ως επιτυχημένη αναγνώριση, η περίπτωση που η έξοδος του συστήματος έχει το σωστό αναγνωριστικό δείκτη του κομματιού αναφοράς και το σωστό δευτερόλεπτο εκκίνησης του αποσπάσματος.

Η λάθος αναγνώριση είναι η περίπτωση που η έξοδος του συστήματος

Θα έχει λάθος αναγνωριστικό κομματιού ανεξαρτήτως του χρόνου εκκίνησης, και τέλος η μη αναγνώριση, όταν το σύστημα δεν έχει ως έξοδο κάποιο αναγνωριστικό.

Παρακάτω παρουσιάζονται τα ποσοστά επιτυχίας, λάθος και μη αναγνώρισης για διαφορετικές τιμές του πλήρους των υποσημάτων P :

Πλήρος υποσημάτων P	Σωστή Αναγνώριση	Λάθος αναγνώριση	Δεν βρέθηκε
6	94.27%	0.53%	5.2%
3	91.14%	0.16%	8.7%
1	96.85%	3.15%	Ø

Πίνακας 6.1: Αποτελέσματα αναζήτησης με καθαρό σήμα για το αποθηκευμένο μέρος των κομματιών αναφοράς.

Τα αποτελέσματα είναι αναμενόμενα: Συγκεκριμένα, βλέπουμε πως με την χρήση της τεχνικής της σύντηξης των τοπικών αποφάσεων, τα ποσοστά των λάθος αναγνωρίσεων βρίσκονται σε πολύ μικρό ποσοστό, ενώ ταυτόχρονα η χρήση μεγαλύτερου αριθμού υποσημάτων ανεβάζει τα ποσοστά της σωστής αναγνώρισης. Επιπλέον, μπορεί να παρατηρηθεί πως και η σωστή αναγνώριση για ένα μοναδικό σήμα μήκους $L = 5sec$ έχει λίγο καλύτερα ποσοστά, ανεβαίνοντας και τα ποσοστά της λάθος αναγνώρισης ακόμα και όταν αναζητούνται αυτούσια τα κομμάτια από την βάση δεδομένων, πόσο μάλλον όταν θα είναι υπαρκτή η προσθήκη παραμορφώσεων.

Αναγνώριση στο μη αποθηκευμένο μέρος των κομματιών αναφοράς

Σε αυτό το σημείο, ιδιαίτερο ενδιαφέρον έχει η αποτελεσματικότητα του αλγορίθμου στα χρονικά σημεία των κομματιών που βρίσκονται εκτός του αποθηκευμένου πρώτου λεπτού. Όπως φαίνεται παρακάτω, η απόδοση του αλγορίθμου μειώνεται σε μεγάλο βαθμό καθώς η αναγνώριση εξαρτάται από τις ομοιότητες που παρουσιάζονται στην πορεία του κομματιού.

Στο παρακάτω πείραμα, για κάθε πλαίσιο σε κάθε κομμάτι ορίζεται: ως επιτυχημένη αναγνώριση, η περίπτωση που η έξοδος του συστήματος έχει το σωστό αναγνωριστικό δείκτη του κομματιού αναφοράς χωρίς να υπολογίζεται ο χρόνος εκκίνησης του αποσπάσματος. Η λάθος αναγνώριση είναι η περίπτωση που η έξοδος του συστήματος θα έχει λάθος αναγνωριστικό κομματιού, και τέλος η μη αναγνώριση, όταν το σύστημα δεν έχει ως έξοδο κάποιο αναγνωριστικό.

Ο λόγος που ο χρόνος εκκίνησης του αποσπάσματος δεν λαμβάνεται υπόψιν είναι προφανής. Από την στιγμή που στην βάση δεδομένων είναι αποθηκευμένο μόνο το πρώτο λεπτό από το κάθε κομμάτι, ο αλγόριθμος δεν μπορεί να εξάγει το σωστό χρόνο εκκίνησης αφού αυτός είναι μετά το πρώτο λεπτό.

Πλήθος υποσημάτων P	Σωστή Αναγνώριση	Λάθος αναγνώριση	Δεν βρέθηκε
6	24.39%	0.26%	75.35%
3	15.35%	0.79%	83.86%
1	55.95%	44.05%	Ø

Πίνακας 6.2: Αποτελέσματα αναζήτησης με καθαρό σήμα για κομμάτι του μη αποθηκευμένου μέρους των κομματιών αναφοράς.

Τα αποτελέσματα εμφανίζουν τα ίδια χαρακτηριστικά με το προηγούμενο πείραμα. Αναφορικά, παρατηρείται καλύτερη απόδοση για $P = 6$ από $P = 3$, ενώ για ένα μοναδικό υποσήμια αν και τα ποσοστά της σωστής αναγνώρισης είναι αρκετά καλύτερα, τα ποσοστά της λάθος αναγνώρισης βρίσκονται στο ίδιο επίπεδο.

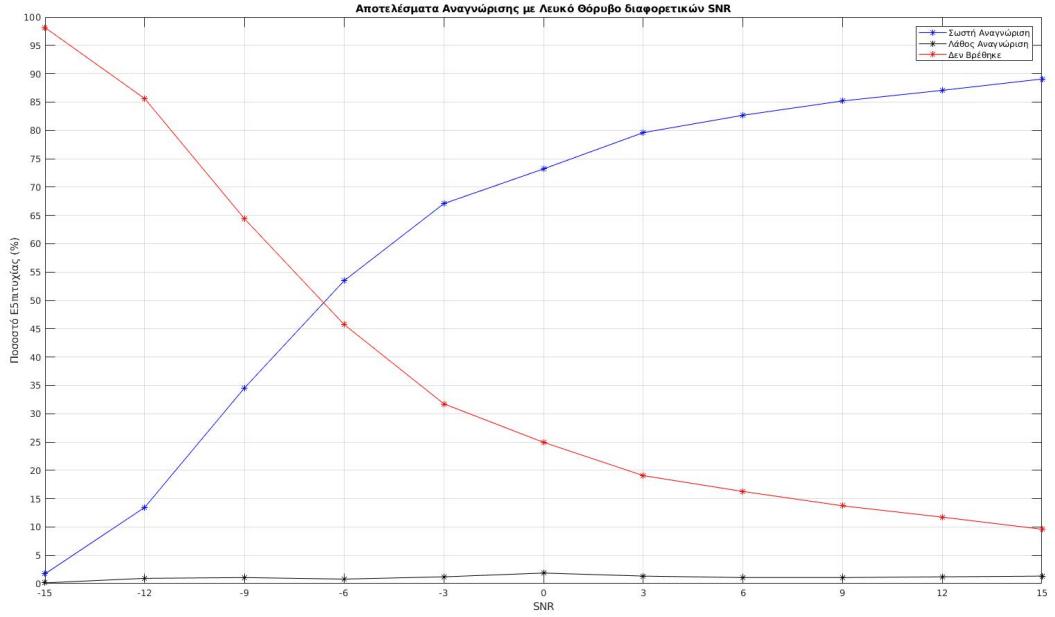
6.3.2 Απόδοση αλγορίθμου με θόρυβο

Έπειτα, εξετάζεται η απόδοση με τις παραμέτρους όπως αυτές ορίζονται στον Πίνακα 5.1, όπου ως είσοδος προς αναγνώριση θα βρίσκεται το πρώτο λεπτό από 250 κομμάτια που υπάρχουν στην βάση δεδομένων με την προσθήκη θορύβου με διαφορετικές τιμές SNR.

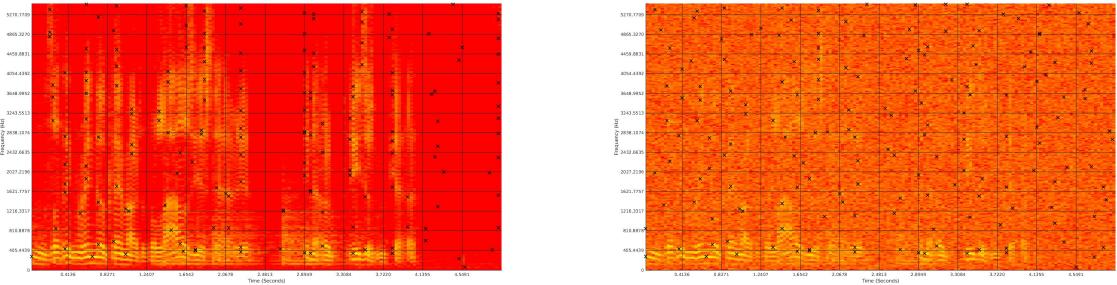
Προσθήκη Λευκού Γκαουσιανού Θορύβου

Στην παρακάτω σειρά πειραμάτων εξετάζεται η απόδοση του αλγορίθμου, με είσοδο προς αναγνώριση 250 αποσπάσματα από τα κομμάτια της βάσης δεδομένων, στα οποία έχει προστεθεί λευκός γκαουσιανός θόρυβος με τιμές $SNR = [15 : -3 : -15]$.

Παρουσιάζονται οι 3 πιθανές έξοδοι του συστήματος:



Σχήμα 6.1: Αποτελέσματα αναζήτησης για σήματα με διαφορετικές τιμές SNR.



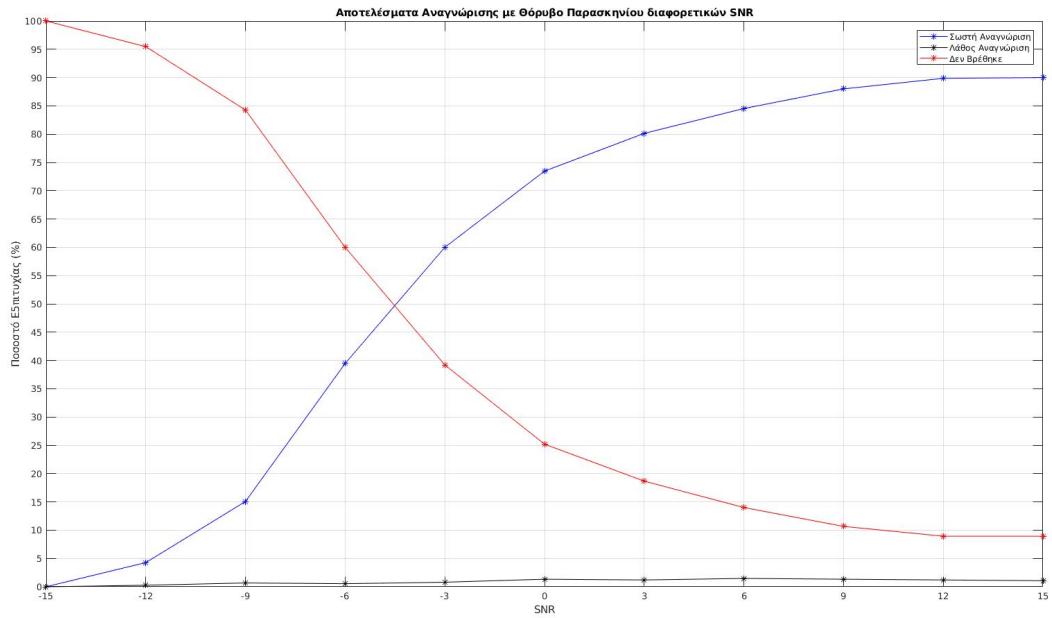
(α') Φασματογράφημα αποσπάσματος χώρις προσθήκη(β') Φασματογράφημα αποσπάσματος με προσθήκη λευκού θορύβου με $SNR = 1$.

Σχήμα 6.2: Παράδειγμα σύγχρισης φασματογραφήματος με προσθήκη λευκού θορύβου.

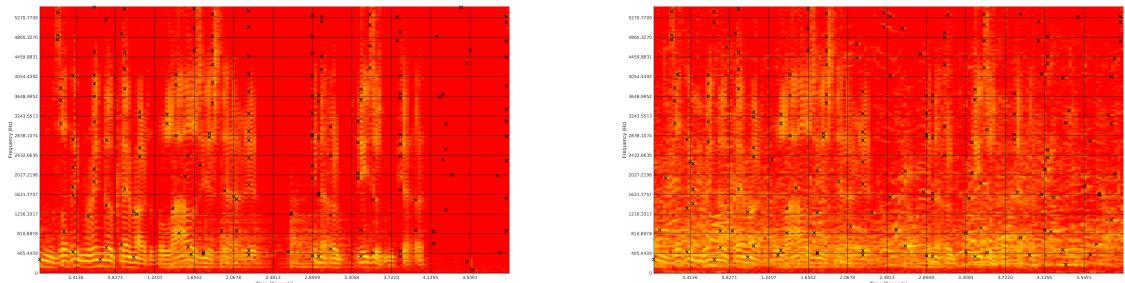
Παρατηρούμε ότι αλγόριθμος έχει ποσοστό επιτυχημένης αναγνώρισης 50% μέχρι και την τιμή $SNR = -6$. Όπως φαίνεται και στο παρακάτω παράδειγμα, η προσθήκη λευκού θορύβου επηρεάζει το φασματογράφημα του αποσπάσματος ομοιόμορφα σε όλο το συχνοτικό του φάσμα, όμως η πλειοψηφία των τοπικών μεγίστων παραμένει στα σημεία του αρχικού ήχου αναφοράς, όταν οι τιμές του SNR δεν είναι πολύ χαμηλές.

Προσθήκη περιβάλλοντος ομιλίας

Στην παρακάτω σειρά πειραμάτων εξετάζεται η απόδοση του αλγορίθμου, με είσοδο προς αναγνώριση 250 αποσπάσματα από τα κομμάτια της βάσης δεδομένων, στα οποία έχει προστεθεί θόρυβος από ένα πλήθος που μιλάει ταυτόχρονα, με τιμές $SNR = [15 : -3 : -15]$. Παρουσιάζονται οι 3 πιθανές έξοδοι του συστήματος:



Σχήμα 6.3: Αποτελέσματα αναζήτησης για σήματα με διαφορετικές τιμές SNR.



(α') Φασματογράφημα αποσπάσματος.

(β') Φασματογράφημα αποσπάσματος με προσθήκη παρασκηνίου με $SNR = 1$

Σχήμα 6.4: Παράδειγμα σύγχρισης φασματογραφήματος με προσθήκη θόρυβου παρασκηνίου με $SNR = 1$.

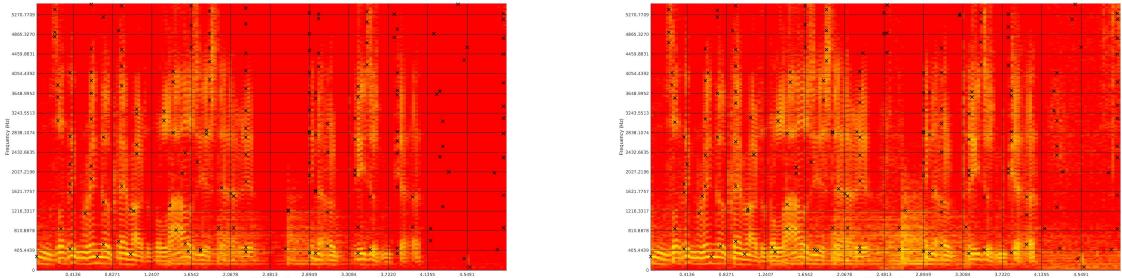
Παρατηρούμε πως τα αποτελέσματα είναι παρόμοια με τα προηγούμενα, όμως ο αλγόριθμος έχει ποσοστό επιτυχημένης αναγνώρισης 50% μέχρι και την τιμή $SNR = -3$. Όπως φαίνεται και στο παρακάτω παράδειγμα, η προσθήκη θορύβου από ομιλία κοινού επηρεάζει το φασματογράφημα του αποσπάσματος σε όλο το συχνοτικό του φάσμα, όμως η πλειοψηφία των τοπικών μεγίστων παραμένει στα σημεία του αρχικού ήχου αναφοράς, όταν οι τιμές του SNR δεν είναι πολύ χαμηλές.

6.3.3 Απόδοση αλγορίθμου με Συμπίεση Δυναμικού Εύρους

Στην παρακάτω σειρά πειραμάτων εξετάζεται η απόδοση του αλγορίθμου, με είσοδο προς αναγνώριση αποσπάσματα από τα κομμάτια της βάσης δεδομένων, τα οποία έχουν υποστεί συμπίεση δυναμικού εύρους. Η συμπίεση έχει ορισθεί να ξεκινά για τιμές πλάτους πάνω από $-40dB$, και συμπίεση πλάτους με συντελεστή κλίσης 0.8. Παρουσιάζονται οι 3 πιθανές έξοδοι του συστήματος:

P	Σωστή Αναγνώριση	Λάθος αναγνώριση	Δεν βρέθηκε
6	93.87%	0.8%	5.33%
1	92.4%	7.6%	\emptyset

Πίνακας 6.3: Αποτελέσματα αναζήτησης με σήμα που έχει υποστεί συμπίεση δυναμικού εύρους, για κομμάτι του μη αποθηκευμένου μέρους των κομματιών αναφοράς.



(α') Φασματογράφημα αποσπάσματος.

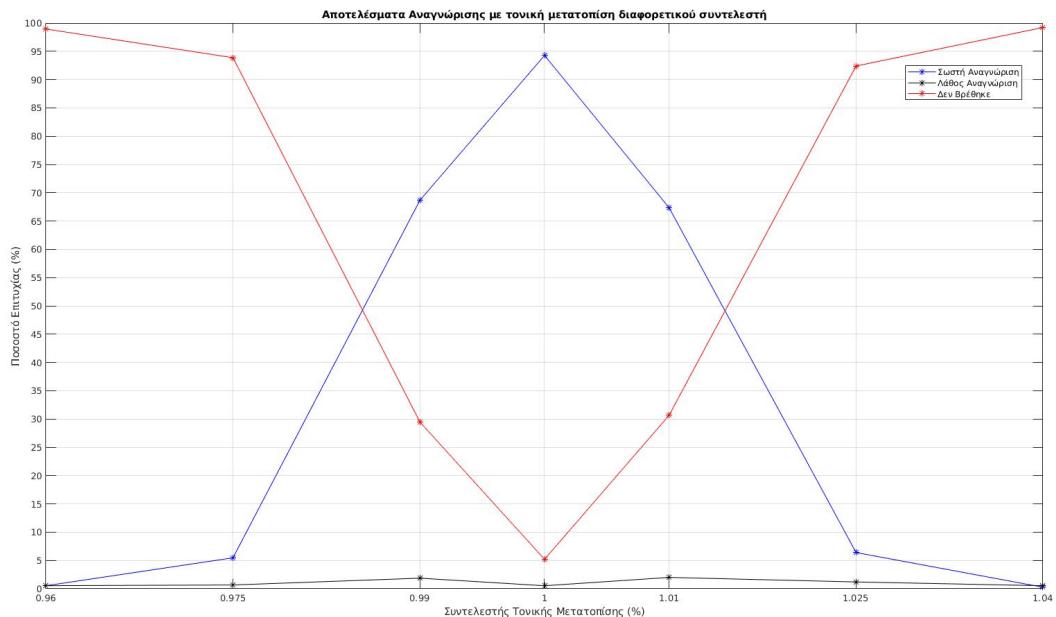
(β') Φασματογράφημα αποσπάσματος με προσθήκη συμπίεσης δυναμικού εύρους.

Σχήμα 6.5: Παράδειγμα συγκρίσης φασματογραφήματος με προσθήκη συμπίεσης δυναμικού εύρους.

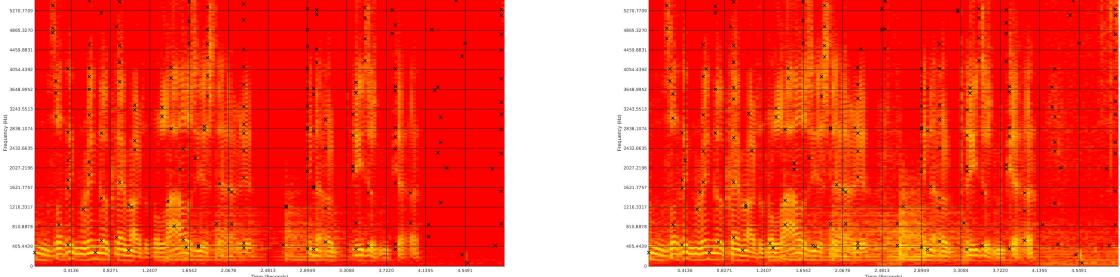
Παρατηρούμε πως η συμπίεση του δυναμικού εύρους των αποσπασμάτων δεν επηρεάζει την απόδοση του αλγορίθμου. Το αποτέλεσμα αυτό είναι αναμενόμενο, διότι με την εφαρμογή της παραμόρφωσης αυτής δεν επηρεάζεται η πλειοψηφία της θέσης των τοπικών μεγίστων στο φασματογράφημα.

6.3.4 Απόδοση αλγορίθμου με τονική μετατόπιση/χρονική παραμόρφωση

Στην παρακάτω σειρά πειραμάτων εξετάζεται η απόδοση του αλγορίθμου, με είσοδο προς αναγνώριση 250 αποσπάσματα από τα κομμάτια της βάσης δεδομένων, τα οποία έχουν υποστεί τονική μετατόπιση που ορίζεται ως απλή αλλαγή του ρυθμού δειγματοληψίας του σήματος. Σε αυτή την περίπτωση η τονική μετατόπιση πηγαίνει ταυτόχρονα με την χρονική παραμόρφωση και οι δύο συντελεστές (χρόνου - συχνότητας) είναι αντιστρόφως ανάλογοι (βλ. 2.7). Παρουσιάζονται οι 3 πιθανές έξοδοι του συστήματος για διαφορετικές τιμές συντελεστών:



Σχήμα 6.6: Αποτελέσματα αναζήτησης για σήματα με διαφορετικές τιμές συντελεστή τονικής μετατόπισης.



(α') Φασματογράφημα αποσπάσματος χωρίς τονική(β') Φασματογράφημα αποσπάσματος με τονική μετατόπιση συντελεστή 4%.

Σχήμα 6.7: Παράδειγμα σύγκρισης φασματογραφήματος με τονική μετατόπιση συντελεστή +4%.

Για την εξήγηση της πολύ χαμηλής απόδοσης του αλγορίθμου, κοιτάμε τις αντιστοιχίες των τοπικών μεγίστων πριν και μετά την παραμόρφωση με συντελεστή κ . Βλέπουμε πως ένα σημείο με συντεταγμένες (t_1, f_1) θα μετατραπεί στο $(\frac{1}{\kappa}t_1, \kappa f_1)$. Έτσι ένα υποψήφιο κλειδί-ζευγάρι δυο τοπικών μεγίστων, θα έχει την μορφή:

$$[f_1, f_2, t_2 - t_1] \rightarrow [\kappa f_1, \kappa f_2, \frac{1}{\kappa}(t_2 - t_1)]$$

Για να μπορεί να επιτευχθεί η ταυτοποίηση του κλειδιού θα πρέπει και τα τρία επιμέρους στοιχεία στην κωδικοποιημένη τους μορφή να παραμένουν σταθερά και μετά την εφαρμογή του παράγοντα κ .

Μπορούμε να δούμε πως η χρονική παραμόρφωση μπορεί να απορροφηθεί από την κωδικοποιημένη διακριτικότητα του $t_2 - t_1$. Η σχέση $\kappa^{\frac{\Delta t_{max}}{2^{n_3}-1}} < 2t_{hop} = 0.064sec$ εξακολουθεί να ισχύει για συνηθισμένες τιμές του συντελεστή κ , επομένως στις περισσότερες περιπτώσεις, η χρονική διαφορά θα παραμένει σταθερή.

$\kappa(\%)$	Κωδικοποιημένη Διαχριτικότητα (sec)
+0	0.0476
+1	0.0480
+2	0.0485
+3	0.0490
+4	0.0495
+5	0.0500
+6	0.0504
+7	0.0509
+8	0.0514
+9	0.0519
+10	0.0523

Πίνακας 6.4: Αντιστοιχία συντελεστή τονικής μετατόπισης κ με την κωδικοποιημένη διαχριτικότητα των χρονικών διαφορών $t_2 - t_1$.

Αντίστοιχα, η διαχριτικότητα των συχνοτήτων δίνεται από τον αντίστροφο του μήκους του παραθύρου που χρησιμοποιείται στο φασματογράφημα $1/L_{FFT} = 15.6Hz$, για τις τιμές των παραμέτρων των πειραμάτων. Επομένως, ο πολλαπλασιασμός των συχνοτήτων με τον συντελεστή κ , επιτρέπει την σωστή αντιστοίχηση μέχρι μια μέγιστη συχνότητα, στην οποία η μετατόπιση υπερβαίνει τα $15.6Hz$, όπως φαίνεται και στον παρακάτω πίνακα.

$\kappa(\%)$	Μέγιστη σωστή συχνότητα (Hz)
+1	1563
+2	782
+3	521
+4	391
+5	313
+6	261
+7	224
+8	196
+9	174
+10	154

Πίνακας 6.5: Μέγιστη τιμή σωστής αντιστοίχισης συχνοτήτων για διαφορετικές τιμές το συντελεστή τονικής μετατόπισης κ .

Εν τέλει, η πλειοψηφία των κλειδιών του παραμορφωμένου αποσπάσματος έχει 2 από τα 3 επιμέρους στοιχεία διαφορετικά. Από την στιγμή που η αναζήτηση απαιτεί επακριβής ταύτιση, τα κλειδιά αυτά δεν μπορούν να ταυτιστούν με το αντίστοιχο κομμάτι αναφοράς στην βάση δεδομένων. Κατ' επέκταση, το βέλτιστο αποτέλεσμα σε κάθε πλαίσιο αναζήτησης θα είναι λανθασμένο, και στην συνέχεια, η σύντηξη των τοπικών αποφάσεων θα αποτύχει

να εξάγει ένα τελικό αποτέλεσμα, γεγονός που εξηγεί τα υψηλά ποσοστά μη αναγνώρισης για μεγαλύτερες τιμές συντελεστή παραμόρφωσης.

6.3.5 Απόδοση Αλγορίθμου σε εξομοίωση διαφορετικών πραγματικών καταστάσεων

Σε αυτή την περίπτωση εξετάζεται η απόδοση του αλγορίθμου σε εξομοίωση διαφορετικών πραγματικών καταστάσεων με την χρήση του Audio Degradation Toolbox. Ως είσοδο χρησιμοποιούνται 100 διαφορετικοί μουσικοί τίτλοι της βάσης δεδομένων. Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα:

Συνθήκη Εξομοίωσης	Σωστή Αναγνώριση	Λάθος αναγνώριση	Δεν βρέθηκε
Ηχογράφηση από Κινητό	98.5%	0%	1.5%
Αναπαραγωγή από Κινητό	100%	0%	0%
Ζωντανή Ηχογράφηση	99.5%	0%	0.5%
Ηχογράφηση από Βινύλιο	100%	0%	0%
Ραδιοφωνική Αναμετάδοση	18,5%	0%	81,5%

Πίνακας 6.6: Αποτελέσματα αναζήτησης σε εξομοίωση διαφορετικών πραγματικών καταστάσεων.

6.4 Πειραματική Διαδικασία Βελτίωσης Αλγορίθμου με χρήση CQT

6.4.1 Απόδοση Βελτίωσης Αλγορίθμου με CQT για καθαρό σήμα

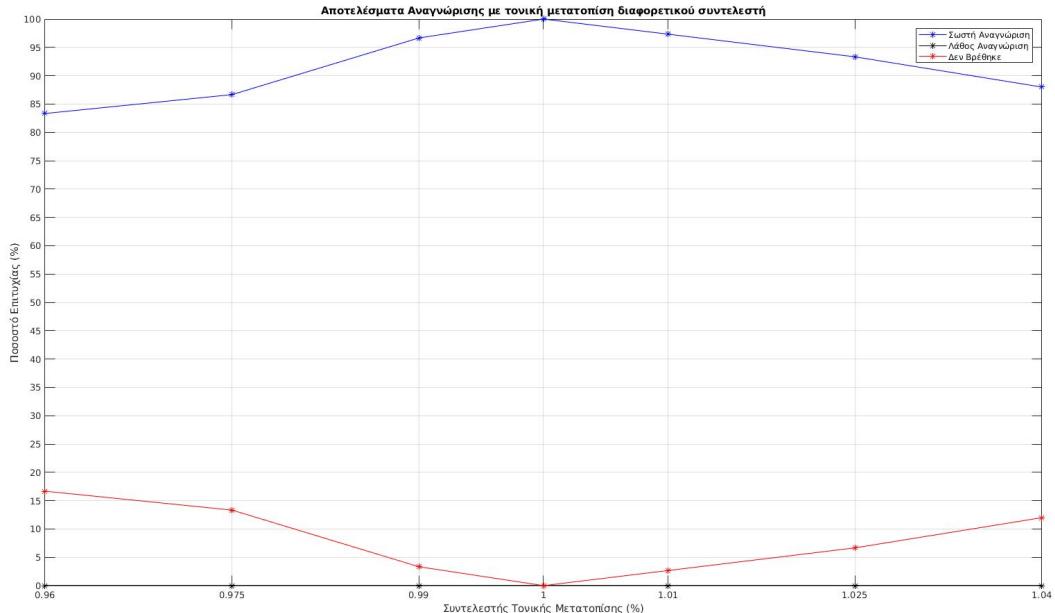
Σε αυτή την περίπτωση εξετάζεται η απόδοση της βελτίωσης του αλγορίθμου με την προτεινόμενη αντικατάσταση της χρήσης STFT με τον CQT. Ως είσοδος προς αναγνώριση είναι 100 κομμάτια που υπάρχουν στην βάση δεδομένων χωρίς την προσθήκη επιπλέον παραμορφώσεων. Παρακάτω παρουσιάζονται τα ποσοστά επιτυχίας, λάθος και μη αναγνώρισης για διαφορετικές τιμές του πλήθους των υποσημάτων P :

Πλήθος υποσημάτων P	Σωστή Αναγνώριση	Λάθος αναγνώριση	Δεν βρέθηκε
6	100%	0%	0%
3	100%	0%	0%
1	98%	2%	Ø

Πίνακας 6.7: Αποτελέσματα αναζήτησης με CQT για είσοδο καθαρού σήματος.

6.4.2 Απόδοση Βελτίωσης Αλγορίθμου με CQT με τονική μετατόπιση/χρονική παραμόρφωση

Στην παρακάτω σειρά πειραμάτων εξετάζεται η απόδοση του αλγορίθμου, με είσοδο προς αναγνώριση 100 αποσπάσματα από τα κομμάτια της βάσης δεδομένων, τα οποία έχουν υποστεί τονική μετατόπιση που ορίζεται ως απλή αλλαγή του ρυθμού δειγματοληψίας του σήματος με τον ίδιο τρόπο που ορίστηκε προηγουμένως. Παρουσιάζονται οι 3 πιθανές έξοδοι του συστήματος για διαφορετικές τιμές συντελεστών:



Σχήμα 6.8: Αποτελέσματα αναζήτησης για σήματα με διαφορετικές τιμές συντελεστή τονικής μετατόπισης.

Τα αποτελέσματα των πειραμάτων είναι πολύ ικανοποιητικά, ιδίως σε σχέση με τον αρχικό αλγόριθμο. Βλέπουμε πως η απόδοση μειώνεται για μεγαλύτερες τιμές του συντελεστή παραμόρφωσης, όμως τα ποσοστά επιτυχημένης αναγνώρισης εξακολουθούν να είναι υψηλά για τις συνηθισμένες τιμές που χρησιμοποιούνται κατά την εφαρμογή της τονικής μετατόπισης.

6.4.3 Απόδοση Βελτίωσης Αλγορίθμου με CQT σε εξομοίωση διαφορετικών πραγματικών καταστάσεων

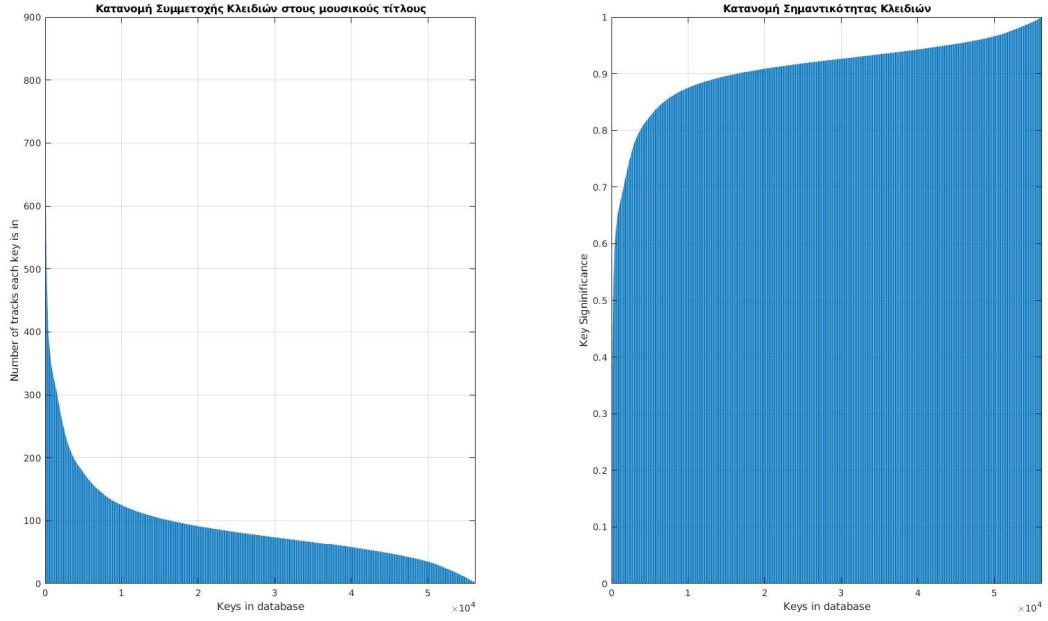
Σε αυτή την περίπτωση εξετάζεται η απόδοση του αλγορίθμου σε εξομοίωση διαφορετικών πραγματικών καταστάσεων με την χρήση του Audio Degradation Toolbox. Ως είσοδο χρησιμοποιούνται 100 διαφορετικοί μουσικοί τίτλοι της βάσης δεδομένων. Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα:

Συνθήκη Εξομοίωσης	Σωστή Αναγνώριση	Λάθος αναγνώριση	Δεν βρέθηκε
Ηχογράφηση από Κινητό	100%	0%	0%
Αναπαραγωγή από Κινητό	100%	0%	0%
Ζωντανή Ηχογράφηση	93%	0%	7%
Ηχογράφηση από Βινύλιο	100%	0%	0%
Ραδιοφωνική Αναμετάδοση	100%	0%	0%

Πίνακας 6.8: Αποτελέσματα αναζήτησης σε εξομοίωση διαφορετικών πραγματικών καταστάσεων.

6.5 Αξιολόγηση της μεθόδου σημαντικότητας των κλειδιών της βάσης δεδομένων

Σε αυτό το σημείο, στοχεύουμε να αξιολογήσουμε την μέθοδο μείωσης όγκου κλειδιών από τη βάση δεδομένων. Για να επιτευχθεί αυτό, η βάση δεδομένων που χρησιμοποιήθηκε κατά την διάρκεια των πειραμάτων με την χρήση του CQT περνάει από την διαδικασία της δημιουργίας της παραμέτρου σημαντικότητας των κλειδιών, με τα παρακάτω αποτελέσματα.



Σχήμα 6.9: Ταξινομημένη παρουσίαση της κατανομής της σημαντικότητας των χλειδιών στην βάση δεδομένων (αύξουσα ταξινόμηση).

Η βάση δεδομένων αυτή περιέχει 1000 διαφορετικούς μουσικούς τίτλους που μετατρέπονται συνολικά σε 56,241 μοναδικά χλειδιά, τα οποία μπορεί να είναι κοινά μεταξύ διαφορετικών τίτλων. Αρχικά, υπολογίζεται το ποσοστό των χλειδιών της βάσης που έχουν συντελεστή σημαντικότητας T_{prune} χαμηλότερο από τρεις διαφορετικές τιμές.

Κατώφλι T_{prune}	Κατάλληλα Μοναδικά Κλειδιά	Τιμές Κατάλληλων Μοναδικών Κλειδιών
0.5	0.38%	5.87%
0.6	0.76%	9.05%
0.7	2.82%	19.61%

Πίνακας 6.9: Το ποσοστό των μοναδικών χλειδιών της βάσης που ικανοποιούν την συνθήκη σημαντικότητας χλειδιού για διαφορετικές τιμές κατωφλίωσης T_{prune} , μαζί με το αντίστοιχο ποσοστό των συνολικών τιμών που περιέχονται στην βάση δεδομένων. Είναι εμφανές πως αν και τα χλειδιά που διαγράφονται είναι πολύ λίγα, περιέχουν σημαντικό ποσοστό της συνολικής πληροφορίας της βάσης δεδομένων.

Έπειτα, οι τρεις νέες, μειωμένες σε όγκο, βάσεις δεδομένων χρησιμοποιούνται για να γίνει σύγχριση με την αρχική βάση μέσω δύο σειρών αναγνώρισεων. Ως είσοδο προς αναγνώριση χρησιμοποιούνται 100 τίτλοι της βάσης δεδομένων σε δύο διαφορετικές

εκδοχές. Πρώτα, χωρίς καμία παραμόρφωση, και σε δεύτερη φάση, παραμορφωμένοι έντονα με την χρήση του Audio Degradation Toolbox. Συγκεκριμένα, οι παραμορφώσεις που χρησιμοποιούνται είναι αυτές της εξομοίωσης ζωντανού περιβάλλοντος και ταυτόχρονα ηχογράφησης μέσω κινητού. Στους παραχάτω πίνακες αναφέρονται τα αποτελέσματα του μέσου όρου χρόνου αναζήτησης των άγνωστων αποσπασμάτων διάρκειας 5 δευτερολέπτων, και στη συνέχεια τα ποσοστά επιτυχίας των δύο πειραμάτων:

Κατώφλι T_{prune}	Μέσος όρος χρόνου αναζήτησης αποσπάσματος 5sec
0	4.2 sec
0.5	2.5 sec
0.6	1.7 sec
0.7	1.3 sec

Πίνακας 6.10: Αποτελέσματα Μέσου Όρου χρόνου αναζήτησης με CQT σε βάσεις διαφορετικού κατωφλιού T_{prune} .

Κατώφλι T_{prune}	Σωστή Αναγνώριση	Λάθος Αναγνώριση	Δεν βρέθηκε
0	100%	0%	0%
0,5	100%	0%	0%
0,6	100%	0%	0%
0,7	100%	0%	0%

Πίνακας 6.11: Αποτελέσματα αναζήτησης με CQT για είσοδο καθαρού σήματος με βάσεις διαφορετικού κατωφλιού T_{prune} .

Κατώφλι T_{prune}	Σωστή Αναγνώριση	Λάθος Αναγνώριση	Δεν βρέθηκε
0	27%	0%	73%
0.5	51%	0%	49%
0.6	54%	0%	46%
0.7	49%	0%	51%

Πίνακας 6.12: Αποτελέσματα αναζήτησης με CQT για είσοδο βαριά παραμορφωμένου σήματος με βάσεις διαφορετικού κατωφλιού T_{prune} .

Τα αποτελέσματα είναι αρκετά ενδιαφέροντα. Αρχικά, βλέπουμε πως όντως ο χρόνος αναζήτησης μειώνεται σε ποσοστά μεγαλύτερα του 50%. Στη συνέχεια βλέπουμε πως με είσοδο καθαρού σήματος δεν υπάρχει καμία διαφορά απόδοσης για το σύνολο των μουσικών τίτλων της βάσης δεδομένων. Τέλος, ιδιαίτερο ενδιαφέρον έχουν τα αποτελέσματα της αναζήτησης με βαριά παραμορφωμένο σήμα εισόδου.

Εδώ, παρατηρείται πως τα ποσοστά επιτυχημένης αναζήτησης είναι αρκετά καλύτερα όταν από την βάση δεδομένων έχουν διαγραφεί τα κλειδιά με χαμηλό δείκτη σημαντικότητας.

Αν και η αρχική σκέψη για την διαγραφή των λιγότερο σημαντικών κλειδιών αφορούσε μόνο και μόνο την βελτίωση στον χρόνο αναζήτησης στην βάση δεδομένων, φαίνεται να επηρεάζει και την απόδοση του αλγορίθμου υθεικά.

Αυτό οφείλεται μάλλον στο γεγονός πως όταν το απόσπασμα είναι βαριά παραμορφωμένο, δημιουργούνται κλειδιά που είναι από τη φύση τους παραπλήσια των ‘καθαρών’ κλειδιών, αλλά και αυτά παραμορφωμένα με τη σειρά τους. Όταν η βάση δεδομένων περιέχει τα κλειδιά χαμηλής σημαντικότητας, τα οποία εμπεριέχονται στους περισσότερους μουσικούς τίτλους της βάσης δεδομένων, αλλά πολύ πιθανόν και στο άγνωστο απόσπασμα, υπάρχουν περισσότερες πιθανότητες τα παραμορφωμένα κλειδιά του αποσπάσματος να ταυτιστούν με ένα κλειδί χαμηλής σημαντικότητας από ότι αυτά που παραμένουν στην βάση δεδομένων μετά από την διαδικασία ‘φίλτραρισματος’ με χριτήριο την σημαντικότητα.

7 Επίλογος

7.1 Ανακεφαλαίωση

Σε αυτή την διπλωματική εργασία έγινε μια εισαγωγή στον τομέα της αναγνώρισης μουσικής με την χρήση μοντέλων ακουστικών αποτυπώματων. Αρχικά, έγινε μια περιγραφή σε διαφορετικές προσεγγίσεις για την λύση του προβλήματος. Είδαμε την δομή ενός τυπικού συστήματος αναγνώρισης μουσικής και επιστροφής των μεταδεδομένων ενός άγνωστου αποσπάσματος, η οποία και αποτελείται από ένα κεντρικό εξυπηρετητή που περιέχει μια μεγάλη βάση δεδομένων με την κωδικοποιημένη συμπαγή μορφή μουσικών τίτλων, και τους χρήστες του συστήματος, οι οποίοι κάνουν εφωτήματα με άγνωστα μουσικά αποσπάσματα, τα οποία ο κεντρικός εξυπηρετητής πρέπει να είναι σε θέση να απαντήσει με ακρίβεια και ταχύτητα.

Έπειτα, περιγράφηκαν διαφορετικές τεχνικές παραγωγής μοντελοποίησης του ακουστικού αποτυπώματος. Αυτές περιλαμβάνουν την εκμάίευση του αποτυπώματος κατευθείαν από την χρονική αναπαράσταση - κυματομορφή του σήματος. Επίσης, από μια συμπυκνωμένη μορφή του φασματογραφήματος του Μετασχηματισμού Φουριέ Σύντομου Χρόνου που πάρνει στοιχεία είτε από τοπικά χαρακτηριστικά σύντομης χρονικής διάρκειας, είτε γενικότερα χαρακτηριστικά υψηλότερης διάστασης μεγάλης χρονικής διάρκειας. Τέλος, το αποτύπωμα μπορεί να προκύπτει από μετατροπή του σήματος σε μια ακολουθία πεπερασμένων χαρακτήρων που μπορεί να αντιπροσωπεύει μουσικά χαρακτηριστικά.

Εξίσου σημαντική αλλά και συνδεδεμένη διαδικασία με την επιλογή του μοντέλου του ακουστικού αποτυπώματος είναι και η επιλογή της τεχνικής αναζήτησης στην βάση δεδομένων. Υπάρχουν δυο μεγάλες κατηγορίες που βασίζονται στην παραδοχή ή όχι της υπόθεσης πως τα χαρακτηριστικά του αποτυπώματος είναι τόσο ευέλικτα ώστε να διατηρούνται στο απόλυτο υπό την παρουσία παραμορφώσεων. Στην πρώτη περίπτωση, αρκεί η αναζήτηση σε μια βάση δεδομένων με ζευγάρια δεικτών-τιμών με μοναδική απαίτηση την ακριβής ταύτιση των δεικτών των καταχωρήσεων. Στην δεύτερη περίπτωση, οι αλγόριθμοι αναζήτησης θεωρούν πως τα χαρακτηριστικά μπορούν να είναι παραμορφωμένα, και για αυτό χρησιμοποιούν στρατηγικές αναζήτησης κατά προσέγγισης. Όμως, η αναζήτηση σε αυτού του τύπου συστήματα είναι αρκετά πιο περίπλοκη.

Στην συνέχεια έγινε αναλυτική περιγραφή και υλοποίηση συστήματος αναγνώρισης μουσικών τίτλων βασισμένου στο [7]. Η τεχνική αυτή χρησιμοποιεί το φασματογράφημα του Μετασχηματισμού Φουριέ Σύντομου Χρόνου, την πληροφορία του οποίου απλοποιεί, κρατώντας από αυτό μόνο τα σημεία που είναι τοπικά μέγιστα ανά σταύρερά χρονικά και συχνοτικά διαστήματα. Τα σημεία αυτά αποτελούν τα δομικά στοιχεία για την δημιουργία των κλειδιών του ακουστικού αποτυπώματος. Τέλος, η τεχνική αναζήτησης περιορίζεται στην κατηγόρια ακριβούς ταύτισης.

Η απόδοση του συστήματος δοκιμάστηκε εκτενώς με άγνωστα αποσπάσματα από

μουσικούς τίτλους της βάσης δεδομένων που φτιάχτηκε. Τα αποσπάσματα εισόδου δέχθηκαν επεξεργασία από διάφορες μονάδες παραμόρφωσης που αντιπροσωπεύουν τα πιο συνηθισμένα είδη παραμόρφωσης ήχου που συναντούνται. Επίσης, τα αποσπάσματα εισόδου πέρασαν από μονάδες που μετατρέπουν το σήμα σε συνθήκες εξομοίωσης πραγματικών συνθηκών, όπως αναπαραγωγή μέσω κινητής συσκευής, ή ραδιοφωνικής ηχογράφησης. Τα αποτελέσματα ήταν επί το πλείστον ικανοποιητικά, με εξαίρεση το εφέ της τονικής μετατόπισης (Pitch-Shifting), που συνδέεται και με την χρονική παραμόρφωση (Time-Stretching). Αιτία αυτού, η ίδια η γραμμική δομή του Μετασχηματισμού Φουρέ Σύντομου Χρόνου, ή αλλιώς, της ομοιόμορφης κατανομής των συχνοτικών ζωνών του φασματογραφήματος.

Για την λύση αυτού του προβλήματος, στην συνέχεια γίνεται υλοποίηση παραλλαγής του ακουστικού αποτυπώματος βασισμένο στο [8]. Ο Μετασχηματισμός Φουρέ Σύντομου Χρόνου αντικαθίσταται με τον Μετασχηματισμό Σταθερού-Q, ο οποίος χρησιμοποιεί γεωμετρική κατανομή στα πεδία της συχνότητας. Τα αποτελέσματα είναι πολύ ικανοποιητικά, καθώς με το ανανεωμένο μοντέλο του αποτυπώματος, η απόδοση του συστήματος με είσοδο παραμορφωμένη με τονική μετατόπιση βρίσκεται σε υψηλά επίπεδα, σε αντίθεση με το αρχικό μοντέλο.

Τέλος, γίνεται μια δοκιμή για την προσπάθεια της μείωσης του όγκου δεδομένων της βάσης, που βασίζεται στην δημιουργία της παραμέτρου "σημαντικότητας" των κλειδιών, δηλαδή το ποσοστό των μουσικών τίτλων από όλη την βάση δεδομένων στους οποίους εμφανίζεται κάθε μοναδικό κλειδί. Έπειτα, εφαρμόζεται διαγραφή αυτών που δεν ικανοποιούν μια συνθήκη κατωφλίωσης βασισμένη σε αυτή την παράμετρο. Τα αποτελέσματα είναι πολύ καλά, καθώς εκτός από την μείωση του χρόνου αναζήτησης στο μισό, παρατηρείται πως η διαγραφή των λιγότερο σημαντικών κλειδιών της βάσης μπορεί να αυξήσει και την απόδοση του αλγορίθμου σε συνθήκες βαριάς παραμόρφωσης.

7.2 Μελλοντικές επεκτάσεις

Με την όλο και πιο διαδεδομένη χρήση συστημάτων αυτόματης αναγνώρισης μουσικής αλλά και ποσοστών επιτυχίας αυτών σε όλο και περισσότερες συσκευές και εφαρμογές, θα μπορούσε να πει κανείς πως οι λύσεις σε αυτόν τον τομέα ικανοποιούν τα περισσότερα πιθανά σενάρια εφαρμογής. Χαρακτηριστικό σημείο αυτής της θέσης είναι πως σε τελευταία μοντέλα κινητών, η αναγνώριση μουσικής μέσω του μικροφώνου της συσκευής τρέχει στο παρασκήνιο συνεχώς, χωρίς σημαντική επιβάρυνση των λοιπών λειτουργιών.

Στην πορεία της διπλωματικής εργασίας και την μελέτη των επιστημονικών συγγραμμάτων, είδαμε μια πληθώρα τεχνικών αναγνώρισης μουσικών τίτλων με την χρήση ακουστικών αποτυπωμάτων. Έχει σημασία όμως να τονιστεί το γεγονός, πως τα συστήματα που υλοποιήθηκαν ανήκουν στην κατηγορία της ακριβούς αναζήτησης όσον αφορά την επιλογή στρατηγικής αναζήτησης στην βάση δεδομένων. Οι τεχνικές αυτές έχουν πολύ καλά αποτελέσματα απέναντι σε ένα ευρύ φάσμα παραμορφώσεων του σήματος, και αν

και πάντα μπορεί να υπάρξει ένας συνδυασμός παραμορφώσεων που ίσως είναι αρκετός για να μειωθούν τα ποσοστά των επιτυχημένων αναγνωρίσεων, με προσεκτική ανάλυση της παραμόρφωσης, μάλλον όταν βρεθεί η αντίστοιχη επεξεργασία του αποτυπώματος για να επέλθουν θετικά αποτελέσματα.

Ο περιορισμός αυτών των μεθόδων όμως, οφείλεται σε αυτό ακριβώς το χαρακτηριστικό. Τα συστήματα που χρησιμοποιούν ακριβής ταύτιση ως τεχνική αναζήτησης είναι σε θέση να αναγνωρίσουν τον μουσικό τίτλο, όχι με μουσικά χαρακτηριστικά υψηλού επιπέδου, αλλά συγχρίνοντας τοπικά χαρακτηριστικά του σήματος. Στην περίπτωση διαφορετικών εκδοχών/ηχογραφήσεων του ίδιου μουσικού τίτλου, η απόδοση πέφτει κατακόρυφα. Αυτό είναι λογικό, αφού τα χαρακτηριστικά που χρησιμοποιούνται για το ακουστικό αποτύπωμα είναι αρκετά χαμηλού επιπέδου και δεν μπορεί να θεωρηθεί πως διατηρούνται σε ένα διαφορετικό σήμα, ακόμα και αν στην ανθρώπινη ακοή, τα δύο μουσικά κομμάτια ακούγονται παρόμοια ή ίδια.

Η εξερεύνηση του τομέα της ταύτισης κατά προσέγγισης, και το συνδυασμό του με την αναγνώριση διασκευών μουσικών τίτλων, είναι το επόμενο λογικό βήμα για κάθε υποψήφιου ερευνητή του πεδίου της Αναγνώρισης Μουσικής (Music Information Retrieval). Η αναγνώριση διασκευών μουσικών τίτλων έχει στόχο την ανάπτυξη μεθόδων που να μπορούν να αξιολογούν και να συγχρίνουν αποτελεσματικά την μουσική ομοιότητα δύο σημάτων.

Η μεθοδολογία ως προς την επίτευξη αυτού του στόχου κυμαίνεται από την απόσπαση μουσικών χαρακτηριστικών από το σήμα, όπως η μελωδία [25] και ο ρυθμός της μουσικής [24]. Άλλες τεχνικές επικεντρώνουν στην ταύτιση χαρακτηριστικών που παράγονται από την αναπαράσταση του σήματος σε μορφή χρωματογραφήματος [26] και εκμεταλλεύονται τεχνικές και αλγορίθμους από τον τομέα της βιοπληροφορικής (πχ. BLAST) για το κομμάτι της αναζήτησης και της ταύτισης κατά προσέγγιση.

Βιβλιογραφία

- [1] P. Cano, E. Battle, E. Gomez, L. de C.T. Gomes, and M. Bonnet “Audio Fingerprinting: Concepts and Applications”, pp 233–245 (2005).
- [2] P. Cano, E. Battle, H. Mayer, and H. Neuschmied “Robust Sound Modeling for Song Detection in Broadcast Audio,” . *in AES, 112th Audio Engineering Society Convention, (Munich, Germany), p. 5531, May 2002.*
- [3] P. Cano, E. Battle, T. Kalker, and J. Haitsma “A Review of Algorithms for Audio Fingerprinting,”. *in IEEE Workshop on Multimedia Signal Processing, St.Thomas, Virgin Islands, USA, Dec. 2002, pp. 169 – 173.*
- [4] Julius O. Smith III “Spectral Audio Processing”. *W3K Publishing, 2011*
- [5] <http://cse16-iiith.virtual-labs.ac.in/exp06/indexie.html>
- [6] Γεώργιος Β. Μουστακίδης ‘Βασικές τεχνικές ψηφιακής επεξεργασίας σημάτων’, *Τζιόλα, 2004*
- [7] A. Wang “An Industrial-strength Audio Search Algorithm”, *in International Symposium on Music Information Retrieval, Baltimore, Maryland, USA, Oct. 2003, pp. 7 – 13.*
- [8] Sébastien Fenet, Gaël Richard, and Yves Grenier, “A Scalable Audio Fingerprint Method with Robustness to Pitch-Shifting”, *in Proceedings of the International Society for Music Information Retrieval (ISMIR), Miami, USA, October 2011, pp. 121-126.*
- [9] Michele Covell and Shumeet Baluja. Known-audio detection using waveprint: Spectrogram fingerprinting by wavelet hashing. *In Proc. International Conference on Acoustics, Speech and Signal Processing (2007)*
- [10] J. Haitsma, T. Kalker, and J. Oostveen, “Robust audio hashing for content identification,” *in International Workshop on Content-Based Multimedia Indexing, Brescia, Italy, September 2001.*
- [11] Matthias Mauch and Sebastian Ewert, “The Audio Degradation Toolbox and its Application to Robustness Evaluation,” *in Proceedings of the 14th International Society for Music Information Retrieval Conference (IMIR 2013), 2013, pp. 83–88*
- [12] P. Cano, E. Battle, H. Mayer, and H. Neuschmied, “Robust Sound Modeling for Song Detection in Broadcast Audio,” *in Audio Engineering Society Convention, Munich, Germany, May 2002*
- [13] H. Neuschmied, H. Mayer, and E. Battle, ”Identification of Audio Titles on the Internet,” *in International Conference on Web Delivering of Music (2001)*

- [14] Fenet, Sébastien, Yves Grenier and Gaël Richard. “An Extended Audio Fingerprint Method with Capabilities for Similar Music Detection.” *ISMIR* (2013).
- [15] M. Ramona and G. Peeters, “Audio identification based on spectral modeling of Bark-bands energy and synchronization through onset detection,” in *International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, May 2011*, pp. 477–480.
- [16] M. Ramona and G. Peeters, “AudioPrint: an efficient audio fingerprint system based on a novel costless synchronization scheme,” in *International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, May 2013*
- [17] H. Özer, B. Sankur, and N. Memon, “Robust audio hashing for audio identification,” in *European Signal Processing Conference, vol. 3, Vienna, Austria, Sep. 2004*
- [18] B. C. J. Moore, Hearing—Handbook of Perception and Cognition. 2nd ed., 1995
- [19] J. C. Brown, “Calculation of a constant Q spectral transform,” *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, 1991.
- [20] Christian Schörkhuber and Anssi Klapuri, ”Constant-Q Transform Toolbox for Music Processing,” 2010
- [21] Judith C. Brown and Miller S. Puckette, ”An efficient algorithm for the calculation of a constant Q transform,” *J. Acoust Soc. Am.*, 92(5):2698–2701, 1992
- [22] <https://symas.com/lmdb/>
- [23] <https://github.com/kyamagu/matlab-lmdb>
- [24] A. Holzapfel and Y. Stylianou, “Rhythmic similarity of music based on dynamic periodicity warping,” in *International Conference on Acoustics, Speech and Signal Processing, 2008*
- [25] M. Marolt, “A mid-level melody-based representation for calculating audio similarity,” in *International Symposium on Music Information Retrieval, Victoria, Canada, Oct. 2006*
- [26] B. Martin, D. G. Brown, P. Hanna, and P. Ferraro, “BLAST for Audio Sequences Alignment: A Fast Scalable Cover Identification Tool,” in *International Symposium on Music Information Retrieval, Porto, Portugal, Oct. 2012*