Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Ανάλυση περιεχομένων ροών ήχου με στόχο την κατάτμηση και ταξινόμηση οπτικοακουστικών δεδομένων**
Ονοματεπώνυμο Φοιτητή	**Ιωάννης Μπενέτος**
Πατρώνυμο	**Δημήτριος**
Αριθμός Μητρώου	**ΜΠΠΛ/ 08028**
Επιβλέπων	**Άγγελος Πικράκης, Λέκτορας**

Ημερομηνία Παράδοσης **Οκτώβριος 2011**

Τριμελής Εξεταστική Επιτροπή

Γεώργιος Τσιχριντζής Καθηγητής	Άγγελος Πικράκης Λέκτορας	Χαράλαμπος Κωνσταντόπουλος Λέκτορας

Αφιερώνεται στη σύζυγό μου Παρασκευή
και στα παιδιά μου Δημήτρη, Άγγελο και Ελένη

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. ΠΕΡΙΛΗΨΗ

Σκοπός αυτής της εργασίας είναι η υλοποίηση και διερεύνηση της αξιοπιστίας και της απόδοσης μιας μεθόδου αυτόματης κατάτμησης και αναγνώρισης του περιεχομένου μιας ηχητικής ροής που βασίζεται στην ανάλυση του ηχητικού περιεχομένου. Αυτή η μέθοδος έχει δημοσιευτεί στο περιοδικό "IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCCESING, VOL. 9" στις 4 Μαΐου 2001 με τίτλο "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification". Συγγραφείς της πρωτότυπης εργασίας είναι οι Tong Zhang (Member IEEE) και C.-C. Jay Kuo (Fellow IEEE).

Ενώ οι τρέχουσες προσεγγίσεις του προβλήματος της κατάτμησης και ταξινόμησης των οπτικοακουστικών δεδομένων έχουν εστιάσει κυρίως στα οπτικά γνωρίσματα, τα ηχητικά σήματα μπορεί στην πραγματικότητα να παίξουν έναν πολύ πιο σημαντικό ρόλο στην ανάλυση του περιεχομένου για πολλές εφαρμογές. Προτείνεται μια προσέγγιση της αυτόματης κατάτμησης και ταξινόμησης των οπτικοακουστικών δεδομένων που βασίζεται στην ανάλυση του ηχητικού περιεχομένου. Το ηχητικό σήμα των κινηματογραφικών ταινιών ή των τηλεοπτικών προγραμμάτων κατατμείται και ταξινομείται σε βασικούς τύπους όπως «φωνή», «μουσική», «τραγούδι», «περιβαλλοντικοί ήχοι», «φωνή με μουσική υπόκρουση», «περιβαλλοντικοί ήχοι με μουσική υπόκρουση», «σιγή», κ.λ.π. Εξάγονται απλά ηχητικά χαρακτηριστικά όπως η ενέργεια, ο ρυθμός διέλευσης του μηδενός, η θεμελιώδης συχνότητα και τα ίχνη των φασματικών κορυφών έτσι ώστε να εξασφαλίζεται η επεξεργασία σε πραγματικό χρόνο. Προτείνεται μια διαδικασία που χρησιμοποιεί ευριστικούς κανόνες για την κατάτμηση και ταξινόμηση των ηχητικών σημάτων και που βασίζεται στη μορφολογική και στατιστική ανάλυση αυτών των χρονομεταβλητών ηχητικών χαρακτηριστικών. Τα πειραματικά δεδομένα δείχνουν ότι το προτεινόμενο σχήμα πετυχαίνει ακρίβεια ταξινόμησης 72%.

Λέξεις Κλειδιά: Ηχητική ανάλυση, ηχητική ταξινόμηση, ηχητική κατάτμηση, ανάλυση οπτικοακουστικού περιεχομένου, φιλτράρισμα και ανάκτηση πληροφορίας, διαχείριση πολυμεσικής βάσης δεδομένων.

Abstract

The purpose of this work is to implement and investigate the reliability and performance of a method for automatic segmentation and classification of the contents of an audio stream based on audio content analysis. This method has been published in the journal "IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 9 "on May 4, 2001 entitled "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification". The authors of the original work are Tong Zhang (Member IEEE) and C.-C. Jay Kuo (Fellow IEEE).

While current approaches for audiovisual data segmentation and classification are mostly focused on visual cues, audio signals may actually play a more important role in content parsing for many applications. An approach to automatic segmentation and classification of audiovisual data based on audio content analysis is proposed. The audio signal from movies or TV programs is segmented and classified into basic types such as speech, music, song, environmental sound, speech with music background, environmental sound with music background, silence, etc. Simple audio features including the energy function, the average zero-crossing rate, the fundamental frequency, and the spectral peak tracks are extracted to ensure the feasibility of real-time processing. A heuristic rule-based procedure is proposed to segment and classify audio signals and built upon morphological and statistical analysis of the time-varying functions of these audio features. Experimental results show that the proposed scheme achieves an accuracy rate of 72% in audio classification.

Index Terms: Audio analysis, audio indexing, audio segmentation, audiovisual content parsing, information filtering and retrieval, multimedia database management.

2. ΕΙΣΑΓΩΓΗ – ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Το έργο της αυτόματης κατάτμησης, ταξινόμησης και ανάκτησης των οπτικοακουστικών δεδομένων βρίσκει σημαντικές εφαρμογές στην επαγγελματική παραγωγή, τη διαχείριση του οπτικοακουστικού υλικού, την εκπαίδευση, την ψυχαγωγία, την επιτήρηση κ.λ.π. Για παράδειγμα, μια τεράστια ποσότητα οπτικοακουστικού υλικού είναι αρχειοθετημένη σε τηλεοπτικές και κινηματογραφικές βάσεις δεδομένων. Αν καταστεί δυνατό αυτό το υλικό να κατατμηθεί και ταξινομηθεί σωστά, θα διευκολυνθεί η ανάκτηση εκείνων των τμημάτων του video που χρειάζονται για την επεξεργασία ενός ντοκουμέντου ή ενός διαφημιστικού video κλιπ. Για να δώσουμε και ένα άλλο παράδειγμα, θα ήταν βολικό για τους χρήστες οπτικοακουστικών βιβλιοθηκών ή οικογενειακών ψυχαγωγικών εφαρμογών να μπορούσαν να ανακτήσουν και να παρακολουθήσουν εκείνα τα αποσπάσματα video που τους ενδιαφέρουν. Καθώς ο όγκος του διαθέσιμου υλικού γίνεται τεράστιος, η χειροκίνητη κατάτμηση και ταξινόμηση καθίσταται αδύνατη. Είναι λοιπόν ξεκάθαρη η τάση για αυτόματη κατάτμηση και ταξινόμηση μέσω ηλεκτρονικού υπολογιστή βασισμένη στην ανάλυση του πολυμεσικού περιεχομένου.

Οι μέχρι τώρα προσεγγίσεις του προβλήματος της κατάτμησης και ταξινόμησης των οπτικοακουστικών δεδομένων έχουν επικεντρωθεί κυρίως στα οπτικά χαρακτηριστικά, όπως είναι οι διαφορές στα χρωματικά ιστογράμματα, τα διανύσματα κίνησης και τα καρέ-κλειδιά (keyframes) [1]–[3]. Αντιθέτως, δίνεται σχετικά μικρή προσοχή στο συνοδευτικό ηχητικό σήμα. Υπάρχει όμως μια σημαντική ποσότητα πληροφορίας μέσα στη συνεχή ροή των ηχητικών δεδομένων που ίσως συχνά αναπαριστά το θέμα με έναν απλούστερο τρόπο απ΄ ότι το οπτικό μέρος. Για παράδειγμα, όλες οι πολεμικές σκηνές πρέπει να περιέχουν τον ήχο πυροβολισμών ή εκρήξεων, ενώ το οπτικό περιεχόμενό τους μπορεί να διαφέρει σημαντικά από ένα video κλιπ σε ένα άλλο. Στην αρχή της ταινίας «Washington Square», υπάρχει ένα τμήμα διάρκειας αρκετών λεπτών, που παρουσιάζει τα κτίρια, τους δρόμους και τους ανθρώπους μιας γειτονιάς. Υπάρχουν πολλά εμπλεκόμενα πλάνα, αλλά η συνεχής συνοδευτική μουσική υποδεικνύει ότι στην πραγματικότητα όλα αυτά ανήκουν στην ίδια ηχητική σκηνή. Εκτός τούτου, η πληροφορία της ομιλίας που περιέχεται στα ηχητικά σήματα είναι συνήθως κρίσιμη για την αναγνώριση του περιεχομένου ενός αποσπάσματος video. Ακούγοντας μόνο το διάλογο ενός τμήματος δίχως να βλέπουμε την εικόνα, είναι συνήθως αρκετό για εμάς ώστε να αντιληφθούμε περί τίνος πρόκειται. Όμως, ένας θεατής μπορεί εύκολα να χαθεί βλέποντας μόνο την εικόνα δίχως να ακούει τον ήχο. Έτσι είναι εύλογο να πούμε ότι το ηχητικό σήμα μπορεί στην πραγματικότητα να παίξει έναν πρωταρχικό ρόλο στην ανάλυση των οπτικοακουστικών δεδομένων.

Πρέπει να γίνει συνδυασμός της ηχητικής και της οπτικής πληροφορίας για την ταξινόμηση σε πραγματικό χρόνο. Το πρώτο βήμα είναι να κατατμήσουμε την οπτική ροή σε σημασιολογικές σκηνές που να βασίζονται σε ανάλυση του ηχητικού περιεχομένου. Ονομάζουμε μια τέτοια μονάδα κατάτμησης ως «ηχητική σκηνή» και την ταξινομούμε ως «καθαρή ομιλία», «καθαρή μουσική», «τραγούδι», «ομιλία με μουσικό υπόβαθρο», «περιβαλλοντικό ήχο με μουσικό υπόβαθρο», «σιγή», κ.λ.π. βασιζόμενοι σε αλγορίθμους ηχητικής ταξινόμησης. Στη συνέχεια, γίνεται περαιτέρω κατάτμηση των ηχητικών σκηνών σε πλάνα σύμφωνα με τα οπτικά δεδομένα, και εξάγονται από κάθε πλάνο καρέ-κλειδιά (keyframes) έτσι ώστε να δημιουργηθεί το οπτικό ευρετήριο. Ο συνδυασμός της ηχητικής και της οπτικής ταξινόμησης προσφέρει μεγάλη βοήθεια στους χρήστες όσον αφορά την περιήγηση και την ανάκτηση εκείνων των τμημάτων μιας ταινίας ή ενός τηλεοπτικού προγράμματος που τους ενδιαφέρει. Για παράδειγμα, η ανάκτηση «των τμημάτων που περιέχουν τραγούδια που εκτελεί ο Michael Jackson» μπορεί να επιτευχθεί αναζητώντας τον ηχητικό δείκτη «τραγούδι» και τα καρέ-κλειδιά του «Michael Jackson».

Σε αυτή την εργασία, εστιάζουμε στο πρόβλημα της κατάτμησης και της ταξινόμησης των ηχητικών σημάτων των οπτικοακουστικών δεδομένων με βάση την ανάλυση του ηχητικού περιεχομένου. Η εργασία είναι οργανωμένη ως εξής: Αρχικά γίνεται μια ανασκόπηση των εργασιών που έχουν γίνει σχετικά με την ανάλυση του ηχητικού περιεχομένου. Έπειτα γίνεται μια επισκόπηση του προτεινόμενου συστήματος και αναλύονται οι υπολογισμοί και οι ιδιότητες των ηχητικών χαρακτηριστικών που χρησιμοποιούνται σε αυτή την εργασία. Στη συνέχεια περιγράφονται οι προτεινόμενες διαδικασίες για την κατάτμηση και την ταξινόμηση της ηχητικής ροής. Τέλος, παρουσιάζονται τα πειραματικά δεδομένα, τα συμπεράσματα και οι προτάσεις μας για μελλοντική εργασία.

3. ΠΡΟΗΓΟΥΜΕΝΕΣ ΕΡΓΑΣΙΕΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΤΟΥ ΗΧΗΤΙΚΟΥ ΠΕΡΙΕΧΟΜΕΝΟΥ

Οι εργασίες που έχουν γίνει μέχρι τώρα πάνω στην ανάλυση του ηχητικού περιεχομένου είναι αρκετά περιορισμένες και ακόμα σε προκαταρκτικό στάδιο. Οι τρέχουσες έρευνες μπορεί γενικά να κατηγοριοποιηθούν στις εξής τέσσερις κατευθύνσεις:

3.1 Κατάτμηση και ταξινόμηση του ήχου

Ένα βασικό πρόβλημα στην κατάτμηση και ταξινόμηση του ήχου είναι η διάκριση μεταξύ φωνής και μουσικής, καθώς αυτοί είναι οι δύο πιο σημαντικοί τύποι ήχου. Η προσέγγιση που παρουσιάστηκε από τον Saunders [4] χρησιμοποιούσε μόνο δύο χαρακτηριστικά, το ρυθμό διέλευσης του μηδενός και την ενέργεια, και εφάρμοζε μια απλή διαδικασία κατωφλίου ενώ οι Schreier και Slaney [5] πρότειναν τη χρήση δεκατριών χαρακτηριστικών στα πεδία χρόνου, συχνότητας και cepstrum, καθώς επίσης και μεθόδους ταξινόμησης βασισμένες σε μοντέλο (MAP, GMM, kNN, κ.λ.π) για να επιτύχουν ανθεκτική επίδοση. Και οι δύο προσεγγίσεις ανέφεραν ότι πέτυχαν διάκριση σε πραγματικό χρόνο και με ακρίβεια πάνω από 90%. Καθώς εν γένει, η φωνή και η μουσική έχουν αρκετά διαφορετική φασματική κατανομή και χρονικά μεταβαλλόμενα πρότυπα, δεν είναι πολύ δύσκολο να επιτευχθεί ένα υψηλό επίπεδο διακριτικής ακρίβειας. Σε μια περαιτέρω ταξινόμηση των ηχητικών δεδομένων μπορεί να εξεταστούν και άλλοι τύποι ήχου, εκτός της φωνής και της μουσικής. Οι Wyse και Smoliar [6] εργάστηκαν πάνω στην ταξινόμηση των ηχητικών σημάτων σε «μουσική», «φωνή», και «άλλο». Στην εργασία τους, η μουσική ανιχνεύθηκε με βάση τη μέση χρονική διάρκεια κατά την οποία υπάρχουν κορυφές σε μια στενή περιοχή συχνοτήτων. Έπειτα, η φωνή διαχωρίστηκε με ιχνηλάτηση του ύψους (pitch tracking). Αυτή η μέθοδος αναπτύχθηκε για τη λεκτική ανάλυση των ειδήσεων. Από τους Kimber και Wilcox [7] προτάθηκε επίσης μια προσέγγιση της ακουστικής κατάτμησης, όπου οι ηχογραφήσεις ταξινομήθηκαν σε «φωνή», «σιγή», «γέλιο» και «ήχο δίχως φωνή». Χρησιμοποίησαν ως ηχητικά χαρακτηριστικά τους συντελεστές cepstral και ως ταξινομητή το κρυφό μοντέλο Markov (HMM). Η μέθοδος εφαρμόστηκε κυρίως για την κατάτμηση των ηχογραφημένων συσκέψεων. Η έρευνα των Pfeiffer και άλλων [8] στόχευσε στην ανάλυση του πλάτους, της συχνότητας και του ύψους των ηχητικών σημάτων, όπως επίσης και στην προσομοίωση της ανθρώπινης ηχητικής αντίληψης έτσι ώστε τα αποτελέσματα να μπορούν να χρησιμοποιηθούν για την κατάτμηση των

ηχητικών ροών και την αναγνώριση της μουσικής. Αυτά τα χαρακτηριστικά χρησιμοποιήθηκαν επίσης για την ανίχνευση των ήχων του πυροβολισμού, του κλάματος και της έκρηξης που πιθανόν να σημαίνουν βίαιο περιεχόμενο.

3.2 Ανάκτηση ήχου βάσει περιεχομένου:

Μια ιδιαίτερη τεχνική στην ανάκτηση του ήχου με βάση το περιεχόμενο είναι η επερώτηση με σιγοτραγούδισμα (query by humming), κατά την οποία ανακτάται ένα τραγούδι σιγοτραγουδώντας την αρμονία του. Ένα τυπικό σύστημα για το σκοπό αυτό παρουσιάστηκε από τους Ghias και άλλους [9]. Ο Foote [10] πρότεινε ένα σύστημα ανάκτησης της μουσικής και των ηχητικών εφέ, στο οποίο χρησιμοποιήθηκαν ως ηχητικά χαρακτηριστικά οι συντελεστές συχνότητας Mel-cepstral (Mel-frequency cepstral coefficients, MFCC), και για την ανάκτηση δομήθηκε ένας δενδροειδής ταξινομητής. Επειδή οι MFCC δεν μπορούν να αντιπροσωπεύσουν δεόντως το τέμπο (timbre), αυτή η μέθοδος γενικά αποτυγχάνει στη διάκριση της μουσικής και των περιβαλλοντικών ήχων με διαφορετικούς ρυθμικούς χαρακτήρες (timbre characters). Στην εργασία των Wold [11] και άλλων σχετικά με την ανάκτηση βάσει περιεχομένου, χρησιμοποιήθηκαν οι στατιστικές τιμές (που περιλάμβαναν τις μέσες τιμές, τις διακυμάνσεις και τις αυτοσυσχετίσεις) αρκετών μετρήσεων στα πεδία του χρόνου και της συχνότητας, για να αναπαρασταθούν τα αντιληπτικά χαρακτηριστικά όπως η ακουστότητα (loudness), η φωτεινότητα (brightness), το εύρος ζώνης (bandwidth) και το ύψος (pitch). Δεδομένου ότι ελήφθησαν υπόψη απλά στατιστικά μεγέθη, αυτή η μέθοδος ήταν κατάλληλη μόνο για ήχους με ένα μοναδικό τέμπο (timbre). Από τους Smith και άλλους [12] προτάθηκε μια μέθοδος ανάκτησης του ήχου με ταχεία έρευνα του εκπεμπόμενου ήχου για να ανιχνευθούν και να εντοπιστούν εκείνα τα ηχητικά τμήματα που περιέχουν ένα συγκεκριμένο πρότυπο αναφοράς βάσει ενός αλγορίθμου ενεργούς έρευνας και μοντελοποίησης του ιστογράμματος των χαρακτηριστικών διέλευσης του μηδενός. Σε αυτόν τον αλγόριθμο έπρεπε να είναι εκ των προτέρων γνωστό το προς ανίχνευση ακριβές τμήμα ήχου.

3.3 Ηχητική ανάλυση για οπτική ταξινόμηση:

Οι Liu και άλλοι [13], [14], εφάρμοσαν τα αποτελέσματα της ηχητικής ανάλυσης για τη διάκριση πέντε διαφορετικών οπτικών σκηνών: «δελτίο ειδήσεων», «δελτίο καιρού», «αγώνας καλαθοσφαίρισης», «ποδοσφαιρικός αγώνας» και «διαφήμιση». Τα

υιοθετηθέντα χαρακτηριστικά γνωρίσματα περιελάμβαναν το λόγο σιγής, το λόγο φωνής και το λόγο ενέργειας των υποζωνών συχνοτήτων που εξήχθησαν αντίστοιχα από την κατανομή της έντασης, την περιβάλλουσα του ύψους και το πεδίο συχνοτήτων. Ως ταξινομητές χρησιμοποιήθηκαν το πολυστρωματικό νευρωνικό δίκτυο (MNN) και το κρυφό μοντέλο Markov (HMM). Παρατηρήθηκε ότι, όταν χρησιμοποιείτο το πολυστρωματικό νευρωνικό δίκτυο, η μέθοδος απέδιδε καλά στη διάκριση μεταξύ των δελτίων, των αγώνων και των διαφημίσεων, αλλά παρουσίαζε δυσκολία στην ταξινόμηση των δύο διαφορετικών τύπων δελτίων (ειδήσεων και καιρού) και των δύο διαφορετικών τύπων αγώνων (καλαθοσφαίριση και ποδόσφαιρο). Με τη χρήση του κρυφού μοντέλου Markov, αυξήθηκε ο συνολικός βαθμός ακρίβειας, αλλά υπήρχαν εσφαλμένες ταξινομήσεις μεταξύ και των πέντε τύπων σκηνών. Οι Liu και Huang [15] εφάρμοσαν επίσης το ίδιο σύνολο χαρακτηριστικών για τη διάκριση των δελτίων ειδήσεων από τις διαφημίσεις και τη μουσική στις ειδησιογραφικές εκπομπές. Χρησιμοποιήθηκε ένας ταξινομητής κατωφλίου και ένας ασαφής ταξινομητής. Οι Patel και Sethi [16] πρότειναν την εκτέλεση του ηχητικού χαρακτηρισμού στα συμπιεσμένα κατά MPEG δεδομένα (στην πραγματικότητα, στα δεδομένα του επιπέδου της υποζώνης) για το σκοπό της οπτικής ταξινόμησης. Το ηχητικό σήμα ταξινομήθηκε σε διαστήματα «διαλόγου», «μη διαλόγου» και «σιγής». Τα χαρακτηριστικά γνωρίσματα ελήφθησαν από τα πεδία της ενέργειας, του ύψους, του φασματογραφήματος και του ρυθμού παύσεων, και οργανώθηκαν σε μία διαδικασία κατωφλίων. Υπήρξαν κάπως αρκετά σφάλματα που συνέβαιναν στην ταξινόμηση μεταξύ των διαστημάτων «διαλόγου» και «μη διαλόγου». Από τους Minami και άλλους [17] προτάθηκε μια προσέγγιση για την οπτική ταξινόμηση μέσω της ανίχνευσης της φωνής και της μουσικής, όπου αξιοποιήθηκαν οι τεχνικές επεξεργασίας εικόνας για την ανάλυση του φασματογραφήματος των ηχητικών σημάτων. Οι φασματικές κορυφές αναγνωρίστηκαν εφαρμόζοντας έναν τελεστή εντοπισμού ακμών (edge detection operator), και οι αρμονικές της φωνής ανιχνεύθηκαν με ένα φίλτρο χτένας (comb filter). Παρουσίασαν επίσης δύο εφαρμογές για την επίδειξη αυτής της μεθόδου ταξινόμησης. Η μία εφαρμογή επέτρεπε στους χρήστες να κάνουν άμεση πρόσβαση σε οποιοδήποτε καρέ του video ενώ η άλλη δημιουργούσε περιλήψεις θεατρικών ή κινηματογραφικών έργων αποσπώντας τα σημαντικά τμήματα video με βάση τις θέσεις της μουσικής και της ομιλίας.

3.4 Ενσωμάτωση της ακουστικής και της οπτικής πληροφορίας για την οπτική κατάτμηση και ταξινόμηση

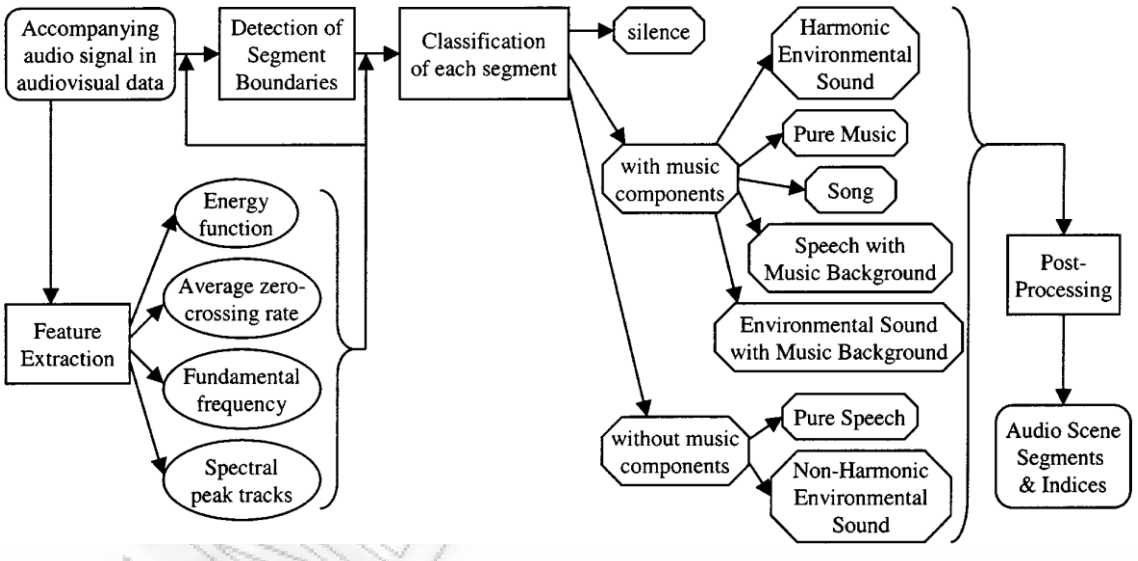
Η νέα τάση για την κατάτμηση και την ταξινόμηση είναι ο συνδυασμός της ηχητικής και της οπτικής πληροφορίας σε ένα ενιαίο πλαίσιο. Αυτή η ιδέα αντικατοπτρίζεται σε τρεις πρόσφατες επιστημονικές δημοσιεύσεις (papers). Ωστόσο, όλα τα ηχητικά χαρακτηριστικά που υιοθετήθηκαν ήταν αρκετά πρωτόγονα, και μέχρις στιγμής δεν εξετάστηκε καμία εκλεπτυσμένη διαδικασία εξαγωγής ηχητικών χαρακτηριστικών γνωρισμάτων για το συγκεκριμένο σκοπό. Στη μέθοδο που προτάθηκε από τον Huang και άλλους [18], το ίδιο σύνολο των ηχητικών χαρακτηριστικών γνωρισμάτων που χρησιμοποιήθηκε στην εργασία του Liu [13] συνδυάστηκε με την πληροφορία χρώματος και κίνησης για να ανιχνευθούν οι σκηνές και τα διαλείμματα λήψης. Στην προσέγγιση που παρουσιάστηκε από τον Naphade [19] και άλλους, συνδυάστηκαν τα δεδομένα της ηχητικής υποζώνης με τα χρωματικά ιστογράμματα για να αποτελέσουν έτσι ένα «Multiject», και χρησιμοποιήθηκαν δύο παραλλαγές του κρυφού μοντέλου Markov για να ταξινομηθούν τα «Multijects». Αναφέρθηκαν τα πειραματικά δεδομένα της ανίχνευσης των γεγονότων της «έκρηξης» και του «καταρράκτη». Στην προσέγγιση των Boreczky και Wilcox [20], χρησιμοποιήθηκαν οι διαφορές των χρωματικών ιστογραμμάτων, οι συντελεστές cepstral των ηχητικών δεδομένων και τα διανύσματα κίνησης μαζί με μία προσέγγιση κρυφού μοντέλου Markov για την κατάτμηση του video σε περιοχές που καθορίζονται από τις λήψεις, τα όρια μεταξύ των λήψεων και την κίνηση της κάμερας κατά τη διάρκεια της λήψης.

Ένα άλλο πεδίο έρευνας που είναι αρκετά σημαντικό για την ανάλυση του ηχητικού περιεχομένου είναι η *ανάλυση ηχητικών σκηνών (Audio Scene Analysis ή ASA)*, που ονομάστηκε έτσι μετά την κλασσική εργασία του Bregman [21]. Σκοπός αυτού του πεδίου είναι η κατανόηση του τρόπου με τον οποίο το ακουστικό σύστημα και ο ανθρώπινος εγκέφαλος επεξεργάζονται τα περίπλοκα ηχητικά περιβάλλοντα, στα οποία είναι παρούσες πολλαπλές πηγές που μεταβάλλονται χρονικά ανεξάρτητα. Οι Brown και Cooke [22], ονόμασαν το ερευνητικό πεδίο της κατασκευής υπολογιστικών μοντέλων για το διαχωρισμό των ηχητικών πηγών ως «υπολογιστική ανάλυση ηχητικών σκηνών» (computational audio scene analysis ή CASA). Ένα παράδειγμα είναι η εργασία του Weintraub [23] στην οποία χρησιμοποιήθηκε ένα πλαίσιο δυναμικού προγραμματισμού γύρω από το αυτοσυσχετιστικό μοντέλο Licklider για να απομονωθούν οι φωνές δύο ομιλητών που αλληλοπαρεμβαλλόντουσαν σε μία ηχογράφηση. Ένα άλλο παράδειγμα είναι το σύστημα που κατασκευάστηκε από τον Ellis [24], με στόχο την ανάλυση του ήχου και το διαχωρισμό αντιληπτικών χαρακτηριστικών μέσα από θορυβώδη ηχητικά μίγματα όπως «η ατμόσφαιρα του δρόμου της πόλης». Ο δομημένος ήχος (structured audio) του MPEG-4 ενοποιεί πολλές ιδέες και προσπάθειες στον τομέα αυτό και παρέχει σημασιολογικές και συμβολικές περιγραφές (semantic and symbolic descriptions) του ήχου (ο αποκωδικοποιητής είναι τυποποιημένος ενώ για τα επόμενα χρόνια απομένει

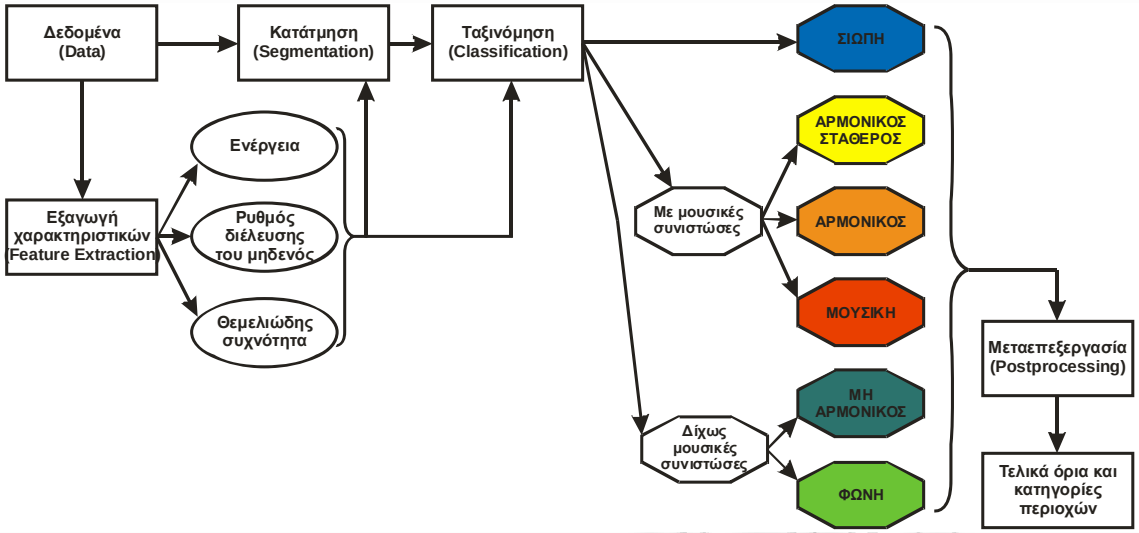
να αναπτυχθούν ώριμες τεχνολογίες για τον κωδικοποιητή). Μια σύνοψη αυτής της εργασίας δόθηκε από τον Vercoe και άλλους στην αναφορά [25]. Αυτή η τεχνική είναι χρήσιμη κατά τη μετάδοση με υπερβολικά χαμηλό ρυθμό bit, στην ευέλικτη σύνθεση (flexible synthesis), και στον αντιληπτικά βασισμένο χειρισμό και ανάκτηση των ήχων.

4. ΕΠΙΣΚΟΠΗΣΗ ΤΟΥ ΠΡΟΤΕΙΝΟΜΕΝΟΥ ΣΥΣΤΗΜΑΤΟΣ

Σε αυτή την εργασία, προτείνουμε ένα σύστημα αυτόματης κατάτμησης και ταξινόμησης των οπτικοακουστικών δεδομένων που βασίζεται στην ανάλυση του ηχητικού περιεχομένου. Εξάγουμε τέσσερα ηχητικά χαρακτηριστικά: την ενέργεια βραχέως χρόνου, το ρυθμό διέλευσης του μηδενός βραχέως χρόνου, τη θεμελιώδη συχνότητα βραχέως χρόνου και τα ίχνη των φασματικών κορυφών. Εκτελούμε μορφολογική και στατιστική ανάλυση των χρονικών καμπυλών αυτών των χαρακτηριστικών για να αποκαλύψουμε τις διαφορές μεταξύ των διαφόρων ηχητικών τύπων. Έπειτα, χρησιμοποιούμε μια ευριστική διαδικασία κανόνων για την κατάτμηση και ταξινόμηση των ηχητικών σημάτων βάσει αυτών των χαρακτηριστικών. Το διάγραμμα ροής της πρωτότυπης εργασίας φαίνεται στο Σχήμα 1, και του υλοποιηθέντος συστήματος στο Σχήμα 2. Οι διαφορές των δύο διαγραμμάτων οφείλονται στους λόγους που εξηγούνται στην ενότητα 6 της παρούσας εργασίας.



Σχήμα 1: Το προτεινόμενο σύστημα κατάτμησης και ταξινόμησης της πρωτότυπης εργασίας



Σχήμα 2: Το υλοποιηθέν σύστημα κατάτμησης και ταξινόμησης

Πρώτα ανιχνεύονται τα όρια των τμημάτων εντοπίζοντας τις απότομες μεταβολές των ηχητικών χαρακτηριστικών βραχέως χρόνου. Έπειτα, κάθε τμήμα ταξινομείται σε έναν από τους βασικούς ηχητικούς τύπους. Ξεχωρίζονται τα σιωπηλά τμήματα, και τα μη σιωπηλά χωρίζονται σε δύο κατηγορίες, με ή δίχως μουσικές συνιστώσες, ανιχνεύοντας την ύπαρξη ή μη συνεχόμενων κορυφών στο φάσμα του ηχητικού σήματος. Τα ηχητικά τμήματα της πρώτης κατηγορίας ταξινομούνται περαιτέρω σε «αρμονικό περιβαλλοντικό ήχο», «καθαρή μουσική», «τραγούδι», «φωνή με μουσική υπόκρουση» και «περιβαλλοντικό ήχο με μουσική υπόκρουση», βάσει της ανάλυσης των ηχητικών χαρακτηριστικών. Τα ηχητικά τμήματα της δεύτερης κατηγορίας ταξινομούνται ως «καθαρή φωνή» ή ως ένας από τέσσερις τύπους «μη αρμονικού περιβαλλοντικού ήχου». Τέλος, εφαρμόζεται ένα στάδιο μεταεπεξεργασίας για τη μείωση των πιθανών σφαλμάτων κατάτμησης.

***Σημείωση**: Στο σύστημα που υλοποιήσαμε δεν υπάρχουν οι κατηγορίες «τραγούδι», «φωνή με μουσική υπόκρουση» και «περιβαλλοντικός ήχος με μουσική υπόκρουση» λόγω ελλιπούς τεκμηρίωσης του πρωτότυπου άρθρου, όπως αναφέρουμε αναλυτικά στην ενότητα 6. Για τους ίδιους λόγους υπάρχει μόνο ένας τύπος «μη αρμονικού περιβαλλοντικού ήχου».*

Για την ηχητική ταξινόμηση αξιοποιούνται τα ηχητικά χαρακτηριστικά με συνδυαστικό τρόπο. Για παράδειγμα, το χαρακτηριστικό της ενέργειας βραχέως χρόνου, του ρυθμού διέλευσης του μηδενός βραχέως χρόνου και της θεμελιώδους συχνότητας βραχέως χρόνου συνδυάζονται αποτελεσματικά για τη διάκριση της φωνής, της μουσικής και της σιγής. Δεν χρησιμοποιούμε μόνο τις τιμές των

χαρακτηριστικών, αλλά επίσης και τα πρότυπα των χρονικών αλλαγών τους και τις σχέσεις μεταξύ των τριών χαρακτηριστικών.

Παρότι το προτεινόμενο σύστημα καλύπτει μια ευρεία γκάμα ηχητικών τύπων, η πολυπλοκότητά του είναι χαμηλή λόγω του ότι τα επιλεχθέντα ηχητικά χαρακτηριστικά γνωρίσματα είναι εύκολο να υπολογιστούν και η διαδικασία ταξινόμησης είναι ταχεία. Τα περισσότερα ηχητικά χαρακτηριστικά που χρησιμοποιήθηκαν στο σύστημα είναι βραχέως χρόνου και μονοδιάστατα, κάνοντας έτσι εφικτή την επεξεργασία των οπτικοακουστικών δεδομένων σε πραγματικό χρόνο. Μεταξύ των τριών χαρακτηριστικών βραχέως χρόνου, η θεμελιώδης συχνότητα είναι η πλέον υπολογιστικά δαπανηρή, λόγω του ότι απαιτεί έναν γρήγορο μετασχηματισμό Fourier (FFT) 512 σημείων ανά 100 δείγματα εισόδου.

Τέλος, η προτεινόμενη προσέγγιση της κατάτμησης και ταξινόμησης του ήχου βασίζεται στην εξέταση των διαφορετικών τύπων ηχητικών σημάτων και των φυσικών χαρακτηριστικών τους, που είναι γενική και ανεξάρτητη μοντέλου. Κατά συνέπεια, μπορεί εύκολα να εφαρμοστεί σαν πρώτο βήμα επεξεργασίας των οπτικοακουστικών δεδομένων, σε σχεδόν οποιοδήποτε σύστημα διαχείρισης οπτικοακουστικού υλικού. Για παράδειγμα, μπορεί να χρησιμοποιηθεί σαν εργαλείο κατάτμησης και ταξινόμησης των ραδιοφωνικών και τηλεοπτικών προγραμμάτων σε πραγματικό χρόνο. Έτσι, μπορεί να δημιουργηθεί αυτόματα ένας πίνακας ευρετηρίου για κάθε εκπεμπόμενο πρόγραμμα, και ο χρήστης να μπορεί να επιλέγει να περιηγηθεί σε συγκεκριμένα τμήματα (π.χ. σε αυτά με καθαρή μουσική). Ειδικά στα τηλεοπτικά προγράμματα, η συμπερίληψη ενός καρέ-κλειδιού (keyframe) σε κάθε τμήμα θα διευκολύνει το έργο της ανάκτησης.

5. ΑΝΑΛΥΣΗ ΤΩΝ ΗΧΗΤΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

5.1 Ενέργεια βραχέως χρόνου (Short-Time Energy)

Η ενέργεια βραχέως χρόνου ενός ηχητικού σήματος ορίζεται ως εξής:

$$E_n = \frac{1}{N}\sum_m [x(m)w(n-m)]^2$$

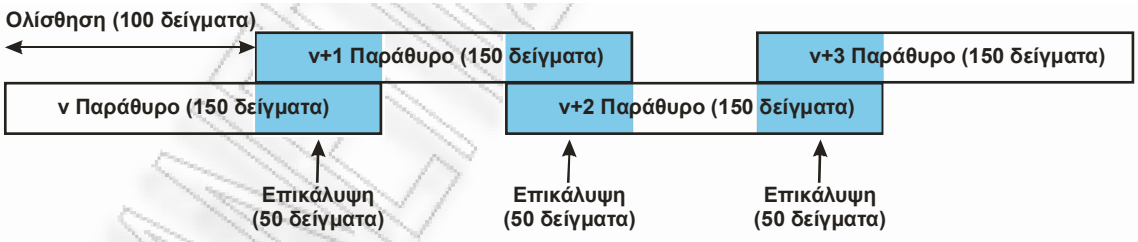
όπου:

x(m) : το ηχητικό σήμα διακριτού χρόνου

n : ο δείκτης χρόνου της ενέργειας βραχέως χρόνου

w(n) : ορθογώνιο παράθυρο μήκους N

Η ενέργεια βραχέως χρόνου αποτελεί μια βολική αναπαράσταση της χρονικής μεταβολής του πλάτους του σήματος. Υποθέτοντας ότι το ηχητικό σήμα μεταβάλλεται σχετικά αργά μέσα σε ένα μικρό χρονικό διάστημα, υπολογίζουμε την En για κάθε 100 δείγματα με συχνότητα δειγματοληψίας 11025 δείγματα/sec. Θέτουμε τη διάρκεια του παραθύρου w(n) να είναι 150 δείγματα έτσι ώστε να υπάρχει μια επικάλυψη μεταξύ των γειτονικών πλαισίων, όπως φαίνεται στο Σχήμα 3.

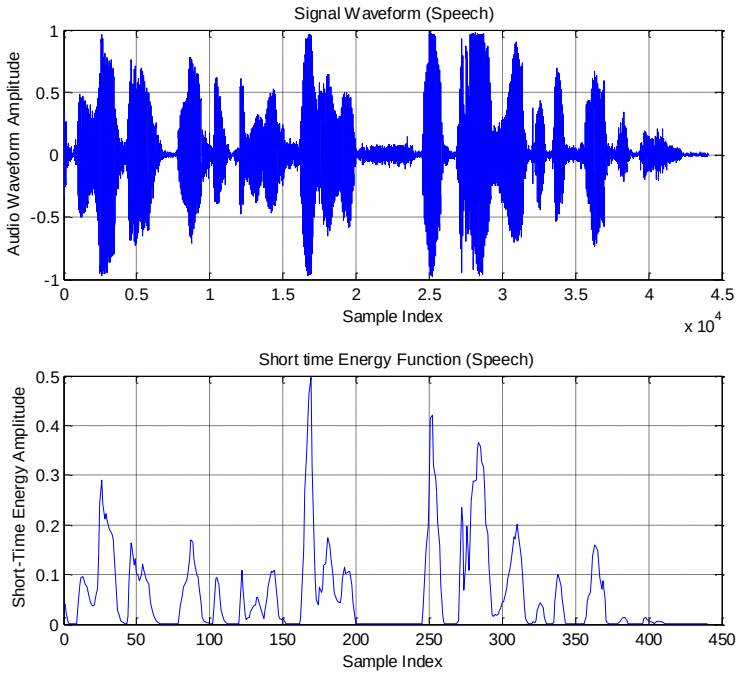


Σχήμα 3: Η τεχνική του ολισθαίνοντος παραθύρου

Στο Σχήμα 4 φαίνεται η κυματομορφή και η αντίστοιχη χρονική καμπύλη της ενέργειας βραχέως χρόνου ενός τμήματος φωνής. Θα πρέπει να σημειωθεί ότι ο δείκτης δείγματος για την καμπύλη της ενέργειας είναι σε λόγο 1:100 σε σχέση με τον αντίστοιχο δείκτη δείγματος της κυματομορφής του σήματος λόγω του ότι εξάγεται μία τιμή της ενέργειας για κάθε 100 δείγματα του σήματος.

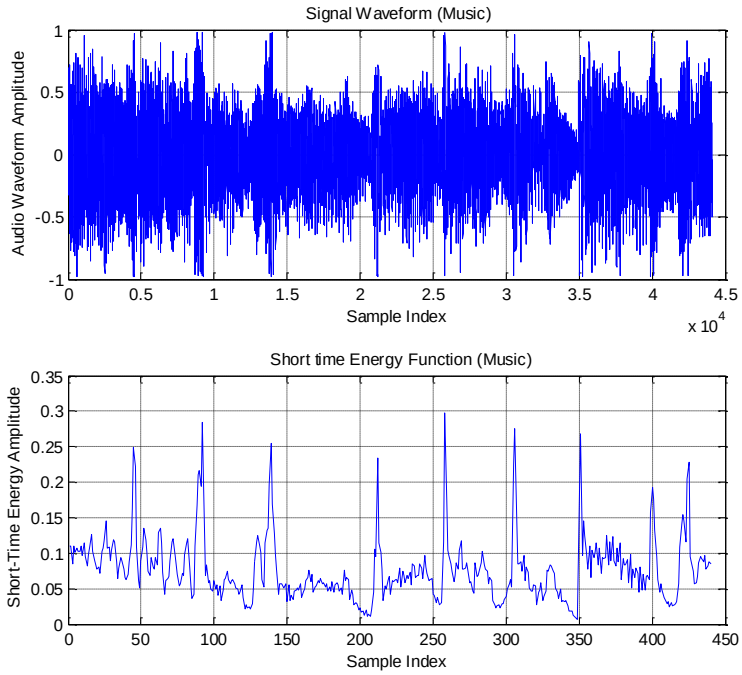
Οι βασικοί λόγοι που χρησιμοποιείται η ενέργεια βραχέως χρόνου στην εργασία αυτή είναι οι εξής:

1. Για τα σήματα φωνής, αποτελεί τη βάση της διάκρισης μεταξύ εμφώνων και αφώνων μερών του λόγου επειδή η τιμή της είναι σημαντικά μικρότερη για τα άφωνα σε σχέση με τα έμφωνα μέρη, όπως μπορεί να παρατηρηθεί από τις κορυφές και τις κοιλάδες της καμπύλης ενέργειας που φαίνεται στο Σχήμα 4.

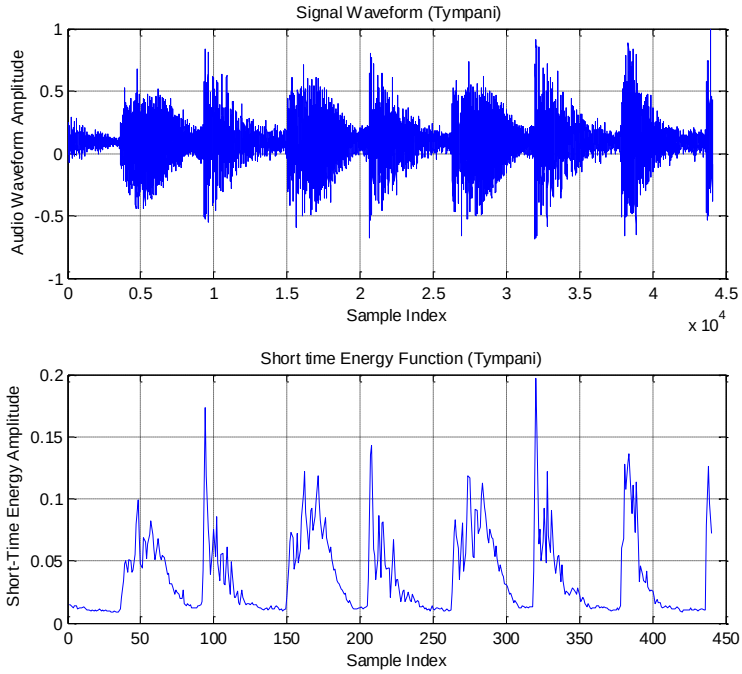


Σχήμα 4: Η κυματομορφή και η ενέργεια βραχέως χρόνου ενός τμήματος φωνής διάρκειας 4sec

2. Μπορεί να χρησιμοποιηθεί για τη διάκριση μεταξύ σιγής και αντιληπτού ήχου όταν ο λόγος SNR είναι υψηλός, όπως μπορεί και πάλι να παρατηρηθεί στο Σχήμα 4, όπου μεταξύ του $2*10^4$ και του $2.5*10^4$ δείγματος της κυματομορφής υπάρχει μια σιγή που έχει σαν αποτέλεσμα την εμφάνιση στην καμπύλη ενέργειας ενός αντίστοιχου τμήματος με σχεδόν μηδενική τιμή.
3. Το πρότυπο της χρονικής αλλαγής της μπορεί να αποκαλύψει το ρυθμό και την περιοδικότητα του ήχου όπως μπορεί να φανεί από την περιοδικότητα των κορυφών της καμπύλης ενέργειας του τμήματος μουσικής που φαίνεται στο Σχήμα 5 και της καμπύλης ενέργειας του ρυθμικού ήχου ενός τύμπανου που φαίνεται στο Σχήμα 6.



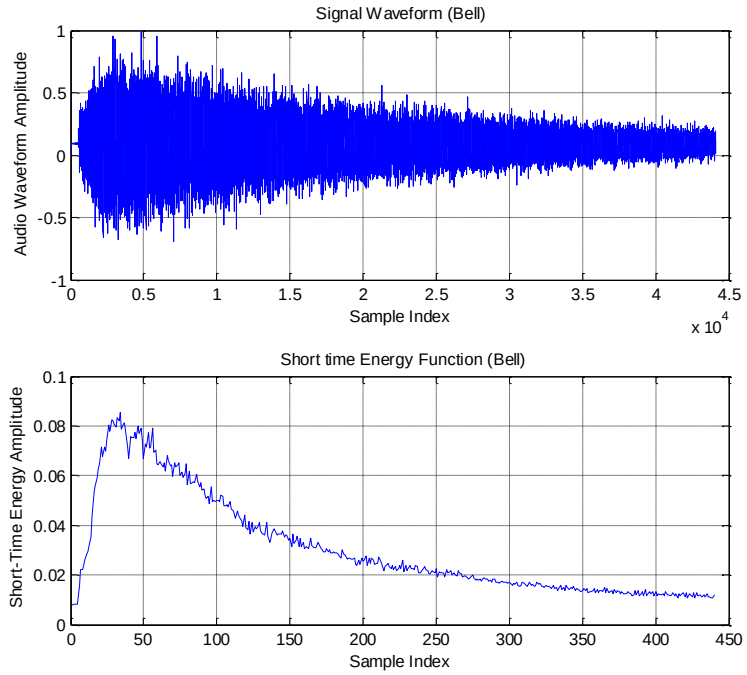
Σχήμα 5: Η κυματομορφή και η ενέργεια βραχέως χρόνου ενός τμήματος μουσικής διάρκειας 4sec



Σχήμα 6: Η κυματομορφή και η ενέργεια βραχέως χρόνου του ήχου ενός τύμπανου διάρκειας 4sec

Επίσης, στο Σχήμα 7 βλέπουμε την κυματομορφή και την καμπύλη ενέργειας του ήχου μιας καμπάνας όπου η ομαλή πτώση του πλάτους της ενέργειας αναπαριστά με

ακρίβεια την ακουστικά παρατηρούμενη σταδιακή πτώση της έντασης του ήχου της καμπάνας.



Σχήμα 7: Η κυματομορφή και η ενέργεια βραχέως χρόνου του ήχου μιας καμπάνας διάρκειας 4sec

5.2 Ρυθμός διέλευσης του μηδενός βραχέως χρόνου (Short-Time Zero-Crossing Rate)

Για τα σήματα διακριτού χρόνου, ορίζεται ότι συμβαίνει μια διέλευση από το μηδέν όταν δύο διαδοχικά δείγματα έχουν αντίθετο πρόσημο. Ο ρυθμός υπό τον οποίο συμβαίνουν οι διελεύσεις από το μηδέν είναι ένα απλό μέτρο του συχνοτικού περιεχομένου ενός σήματος. Ο ρυθμός διέλευσης του μηδενός βραχέως χρόνου (ZCR) ορίζεται ως εξής:

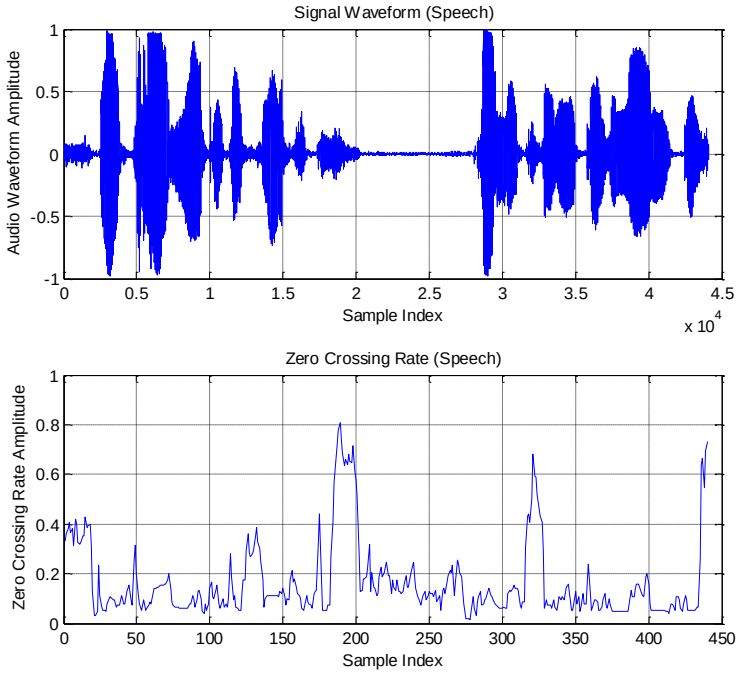
$$Z_n = \frac{1}{2}\sum_m \left|\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]\right| w(n-m)$$

όπου:

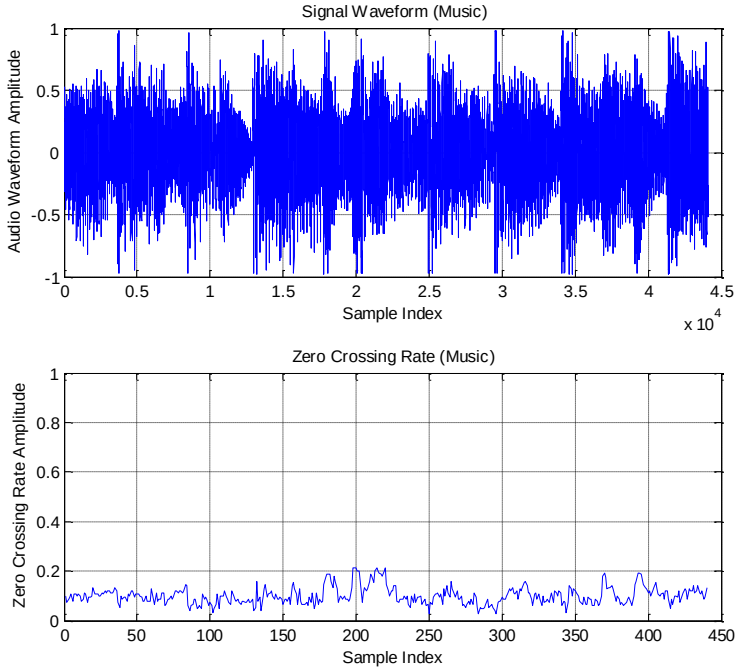
$$\operatorname{sgn}[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases}$$

και w(n) είναι ένα ορθογώνιο παράθυρο.

Όμοια με τον υπολογισμό της ενέργειας βραχέως χρόνου, επιλέγουμε να υπολογίζουμε την τιμή του ρυθμός διέλευσης του μηδενός για κάθε 100 δείγματα εισόδου, και θέτουμε τη διάρκεια του παραθύρου στα 150 δείγματα.



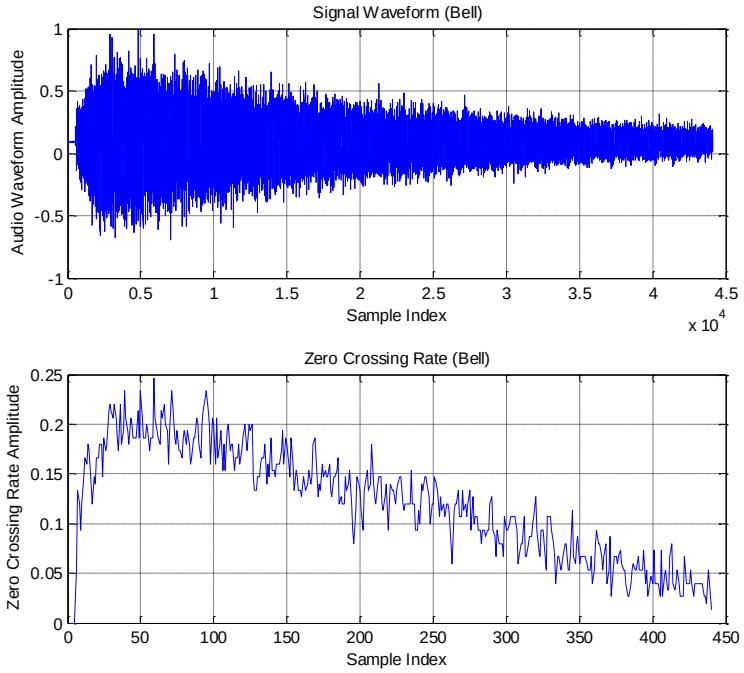
Σχήμα 8: Η κυματομορφή και ο ZCR βραχέως χρόνου ενός τμήματος φωνής διάρκειας 4sec



Σχήμα 9: Η κυματομορφή και ο ZCR βραχέως χρόνου ενός τμήματος μουσικής διάρκειας 4sec

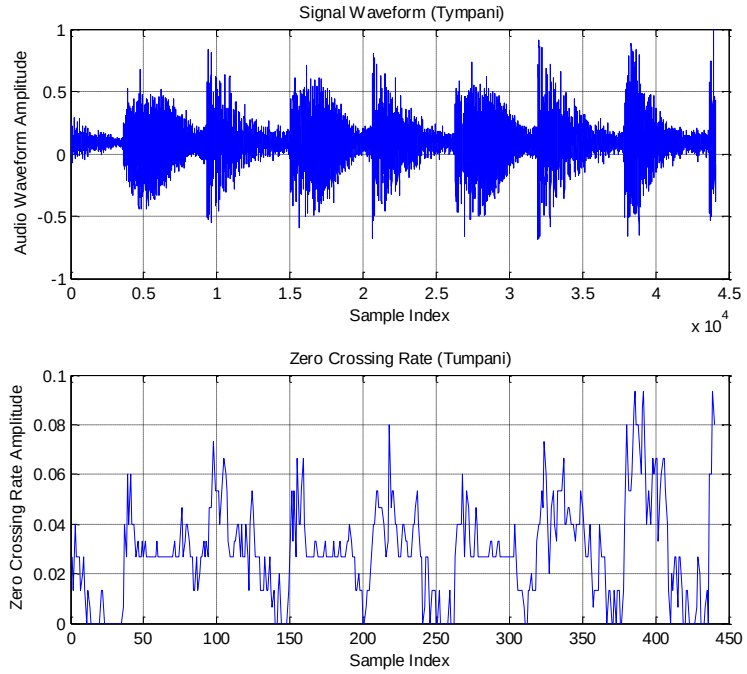
Στο Σχήμα 8 φαίνεται η καμπύλη του ρυθμού διέλευσης του μηδενός βραχέως χρόνου ενός τμήματος φωνής, και στο Σχήμα 9 ενός τμήματος μουσικής.

Ο ρυθμός διέλευσης του μηδενός βραχέως χρόνου μπορεί να χρησιμοποιηθεί σαν ένα άλλο μέτρο για τη διάκριση μεταξύ των εμφώνων και των αφώνων μερών του λόγου, διότι τα άφωνα μέρη έχουν συνήθως πολύ υψηλότερο ZCR σε σχέση με τα έμφωνα [26]. Όπως φαίνεται και στο Σχήμα 8, η καμπύλη του ZCR της φωνής έχει κορυφές και κοιλάδες που οφείλονται στα άφωνα και στα έμφωνα μέρη αντίστοιχα. Αυτό έχει ως αποτέλεσμα μια μεγάλη διακύμανση και μια ευρεία περιοχή του πλάτους της καμπύλης του ZCR. Ας σημειωθεί επίσης ότι η καμπύλη του ZCR έχει μια σχετικά χαμηλή και σταθερή γραμμή βάσης με υψηλές κορυφές πάνω από αυτήν. Συγκριτικά, η καμπύλη του ZCR για τη μουσική έχει μια πολύ χαμηλότερη διακύμανση και μέσο πλάτος, γεγονός που υποδηλώνει ότι ο ρυθμός διέλευσης του μηδενός της μουσικής είναι κανονικά πολύ πιο σταθερός κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου. Οι καμπύλες του ZCR για τη μουσική έχουν γενικά ακανόνιστες κυματομορφές με μεταβαλλόμενη γραμμή βάσης και σχετικά μικρό εύρος πλατών. Καθώς οι περιβαλλοντικοί ήχοι αποτελούνται από ήχους διαφόρων προελεύσεων, οι καμπύλες του ZCR τους μπορεί να έχουν πολύ διαφορετικές ιδιότητες. Για παράδειγμα, ο ZCR μιας καμπάνας αποκαλύπτει μια συνεχή πτώση του κεντροειδούς της συχνότητας συναρτήσει του χρόνου όπως φαίνεται στο Σχήμα 10.



Σχήμα 10: Η κυματομορφή και ο ZCR βραχέως χρόνου του ήχου μιας καμπάνας διάρκειας 6sec

Επίσης, το μικρό πλάτος και η μορφή της καμπύλης του ZCR του ρυθμικού κτυπήματος ενός συμφωνικού τύμπανου αποκαλύπτει τον απότομο, «κρουστικό» και μικρής διάρκειας ήχο του οργάνου όπως φαίνεται στο Σχήμα 11.

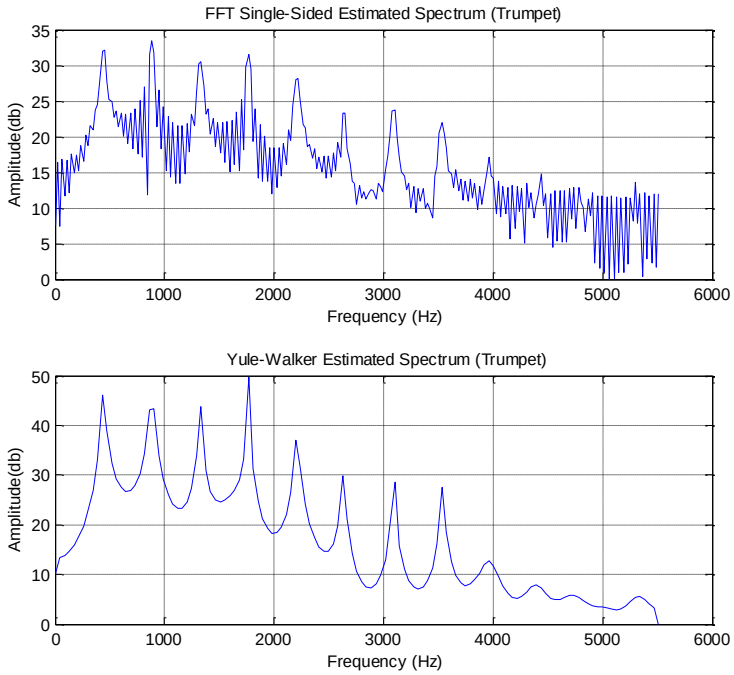


Σχήμα 11: Η κυματομορφή και ο ZCR βραχέως χρόνου του ήχου ενός τυμπάνου διάρκειας 4sec

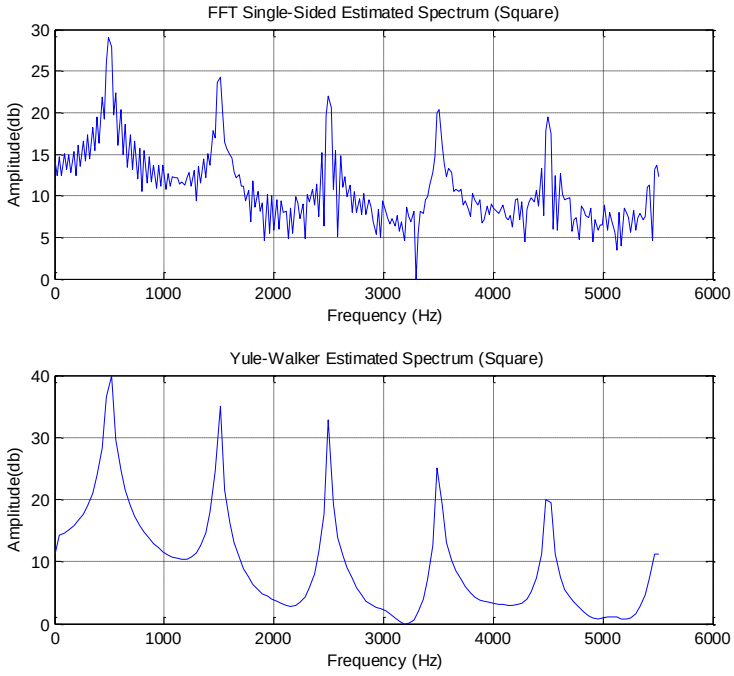
Μπορούμε λοιπόν εν ολίγοις, να ταξινομήσουμε τους περιβαλλοντικούς ήχους σύμφωνα με τις ιδιότητες της καμπύλης ZCR όπως την κανονικότητα, την περιοδικότητα, τη σταθερότητα και την περιοχή του πλάτους.

5.3 Θεμελιώδης συχνότητα βραχέως χρόνου (Short-Time Fundamental Frequency)

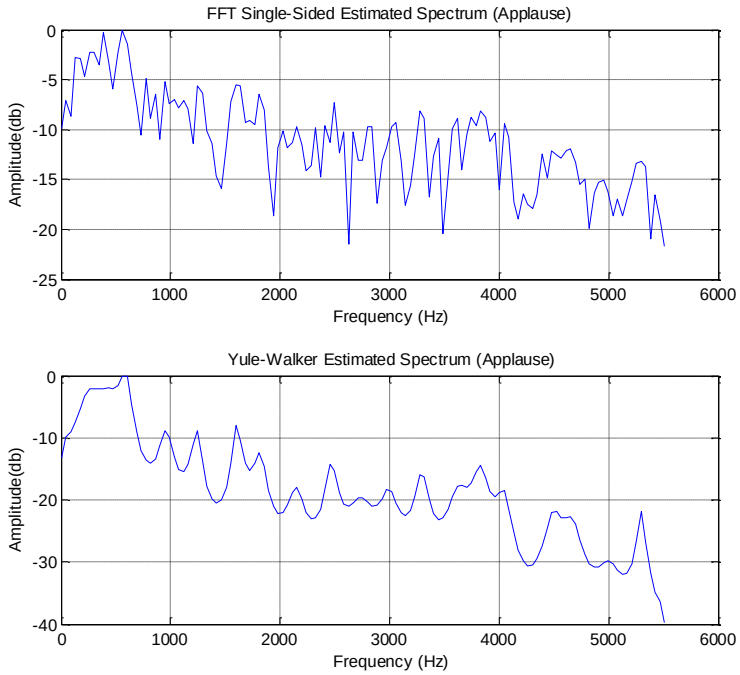
Ένας αρμονικός ήχος αποτελείται από μία θεμελιώδη συχνότητα και μια σειρά ακεραίων πολλαπλασίων της που ονομάζονται αρμονικές συχνότητες. Κατ΄ αυτή την έννοια, μπορούμε να κατατάξουμε τους ήχους σε δύο κατηγορίες, στους αρμονικούς και στους μη αρμονικούς.



Σχήμα 12: Φάσμα του ήχου μιας τρομπέτας όπως εκτιμάται α) απευθείας από τον FFT και β) από την αυτοπαλινδρομική μέθοδο Yule-Walker.



Σχήμα 13: Φάσμα ενός τετραγωνικού σήματος 500Hz όπως εκτιμάται α) απευθείας από τον FFT και β) από την αυτοπαλινδρομική μέθοδο Yule-Walker.



Σχήμα 14: Φάσμα του ήχου χειροκροτήματος όπως εκτιμάται α) απευθείας από τον FFT και β) από την αυτοπαλινδρομική μέθοδο Yule-Walker.

Στα σχήματα 12, 13 και 14, φαίνονται αντίστοιχα τα φάσματα του ήχου μιας τρομπέτας, ενός τετραγωνικού σήματος και του χειροκροτήματος όπως εκτιμώνται α) απευθείας από τον FFT (Fast Fourier Transformation) και β) από την αυτοπαλινδρομική μέθοδο Yule-Walker (Yule-Walker Autoregressive method). Είναι φανερό, από την ύπαρξη αρμονικών συχνοτήτων, ότι τα δύο πρώτα σήματα είναι αρμονικά ενώ το τρίτο είναι μη αρμονικό καθώς δεν εμφανίζονται σε αυτό αρμονικές συχνότητες. Αν ένας ήχος είναι αρμονικός ή μη εξαρτάται από την πηγή του ήχου. Οι ήχοι των περισσοτέρων μουσικών οργάνων είναι αρμονικοί. Η φωνή είναι ένα μίγμα αρμονικών και μη αρμονικών ήχων, λόγω του ότι τα έμφωνα τμήματά της είναι αρμονικά ενώ τα άφωνα είναι μη αρμονικά. Οι περισσότεροι περιβαλλοντικοί ήχοι, όπως το χειροκρότημα, ο ήχος των βημάτων και οι εκρήξεις, είναι μη αρμονικοί ήχοι. Όμως, υπάρχουν επίσης παραδείγματα ηχητικών εφέ που είναι αρμονικοί και σταθεροί ήχοι, όπως η ήχος του κουδουνιού και του τηλεφωνικού πληκτρολογίου, αλλά και παραδείγματα που είναι μίγματα αρμονικών και μη αρμονικών ήχων, όπως είναι το γέλιο και το γαύγισμα.

Προκειμένου να μετρήσουμε το αρμονικό χαρακτηριστικό του ήχου, ορίζουμε τη θεμελιώδη συχνότητα βραχέως χρόνου (SFuF) ως εξής:

Όταν ο ήχος είναι αρμονικός η τιμή της SFuF ισούται με τη θεμελιώδη συχνότητα που εκτιμάται εκείνη τη χρονική στιγμή, ενώ όταν ο ήχος είναι μη αρμονικός η τιμή της SFuF τίθεται το μηδέν.

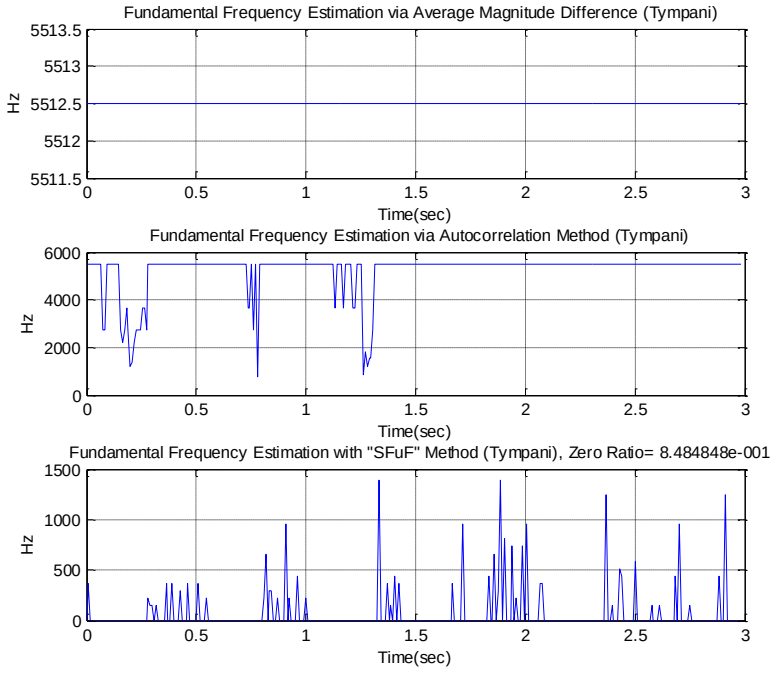
Αν και έχουν προταθεί πολλές μέθοδοι για την εκτίμηση της θεμελιώδους συχνότητας ή την ανίχνευση του ύψους του ήχου (pitch detection) στην ανάλυση της φωνής και της μουσικής [26]-[29] (αξίζει να τονιστεί ότι η θεμελιώδης συχνότητα είναι μια φυσική και μετρήσιμη ποσότητα ενώ το ύψος αποτελεί έναν αντιληπτικό όρο [30]), καμία από αυτές τις μεθόδους δεν είναι απολύτως ικανοποιητική για το χαρακτηρισμό μιας ευρείας γκάμας ηχητικών σημάτων. Καθώς ο πρωταρχικός μας σκοπός για την εκτίμηση της θεμελιώδους συχνότητας είναι η ανίχνευση της αρμονικής ιδιότητας για όλα τα είδη των ηχητικών σημάτων, φροντίσαμε να αναπτύξουμε μια μέθοδο που να είναι αποδοτική και ανθεκτική αλλά όχι κατ΄ ανάγκη απολύτως ακριβής. Σε αυτή την εργασία, ο υπολογισμός της θεμελιώδους συχνότητας βραχέως χρόνου βασίζεται στην ανίχνευση των κορυφών του ηχητικού φάσματος. Το φάσμα εξάγεται με την αυτοπαλινδρομική (autoregressive ή AR) παραμετρική μέθοδο Yule-Walker που υπάρχει στην εργαλειοθήκη επεξεργασίας σήματος (Signal Processing Toolbox) του MATLAB. Το φάσμα που γεννάται από αυτή τη μέθοδο είναι μια εξομαλυσμένη εκδοχή του συχνοτικού περιεχομένου του σήματος. Επί πλέον, καθώς το αυτοπαλινδρομικό (AR) μοντέλο είναι μία έκφραση μόνο πόλων, οι κορυφές κατέχουν προεξέχουσα θέση στο φάσμα. Συγκρίνοντας τα φάσματα που παράγονται με το αυτοπαλινδρομικό μοντέλο με αυτά που παράγονται απ΄ ευθείας από τον FFT για τους ήχους των σχημάτων 12, 13 και 14, μπορούμε να δούμε ότι η ανίχνευση των κορυφών των αρμονικών συχνοτήτων είναι πολύ ευκολότερο να γίνει με το φάσμα του αυτοπαλινδρομικού μοντέλου απ΄ ότι με το απ΄ ευθείας υπολογιζόμενο από τον FFT φάσμα. Επιλέγουμε η τάξη του αυτοπαλινδρομικού μοντέλου να είναι 40. Με αυτή την τάξη, οι αρμονικές κορυφές είναι αξιοσημείωτες, ενώ εμφανίζονται επίσης και μη αρμονικές κορυφές. Όμως οι μη αρμονικές κορυφές, συγκρινόμενες με τις αρμονικές, όχι μόνο στερούνται ακριβούς αρμονικής σχέσης, αλλά εμφανίζονται επίσης να είναι λιγότερο οξείες και να έχουν μικρότερο ύψος. Τοιουτοτρόπως, για να θεωρηθεί ένας ήχος ως αρμονικός πρέπει για ορισμένες από τις φασματικές κορυφές του να ισχύουν τα ακόλουθα:

- οι θέσεις τους να έχουν ελάχιστο κοινό διαιρέτη
- μερικές τουλάχιστον από αυτές να έχουν μεγάλο ύψος και
- να είναι αρκετά οξείες

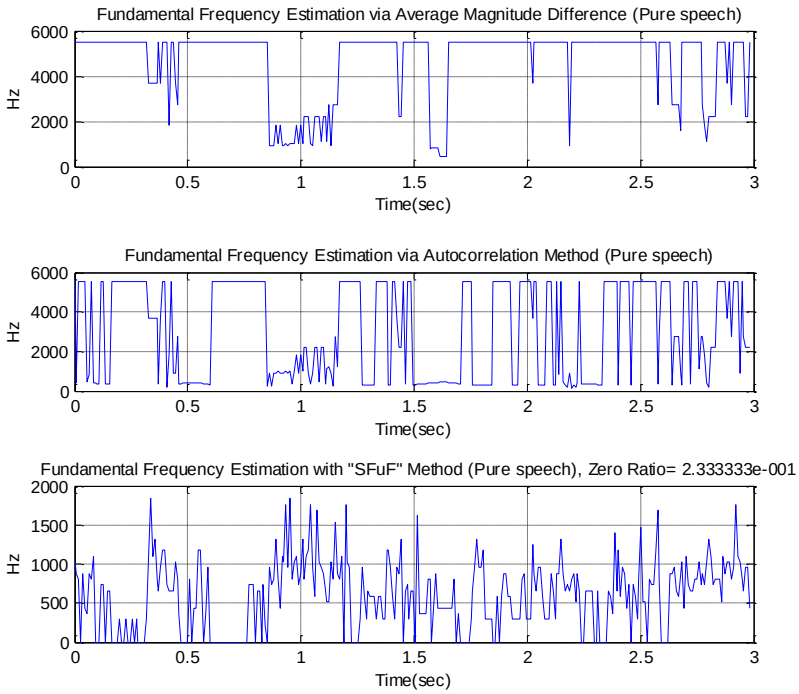
Για το σκοπό αυτό, εντοπίζονται όλα τα μέγιστα του φάσματος ως εν δυνάμει αρμονικές κορυφές, και υπολογίζεται το ύψος, το πλάτος και η οξύτητα κάθε

κορυφής χρησιμοποιώντας μορφολογική ανάλυση. Έπειτα, εξετάζεται αν κάποιες από τις θέσεις αυτών των κορυφών έχουν κοινό διαιρέτη και αν τουλάχιστον ορισμένες από αυτές έχουν οξύτητα, ύψος και πλάτος που να ικανοποιούν κάποια συγκεκριμένα κριτήρια. Αν πληρούνται όλες οι παραπάνω προϋποθέσεις, εκτιμάται ως τιμή της θεμελιώδους συχνότητας βραχέως χρόνου (SFuF) η συχνότητα που αντιστοιχεί στο μέγιστο κοινό διαιρέτη των θέσεων των αρμονικών κορυφών, αλλιώς η SFuF τίθεται στο μηδέν. Η SFuF υπολογίζεται για κάθε 100 δείγματα εισόδου. Αφού εξαχθεί η χρονική καμπύλη της SFuF για ένα τμήμα ορισμένου μήκους, ακολουθεί ένα στάδιο μεταεπεξεργασίας κατά το οποίο αφαιρούνται τα μεμονωμένα σημεία για να βελτιωθεί η ακρίβεια της εκτίμησης της SFuF.

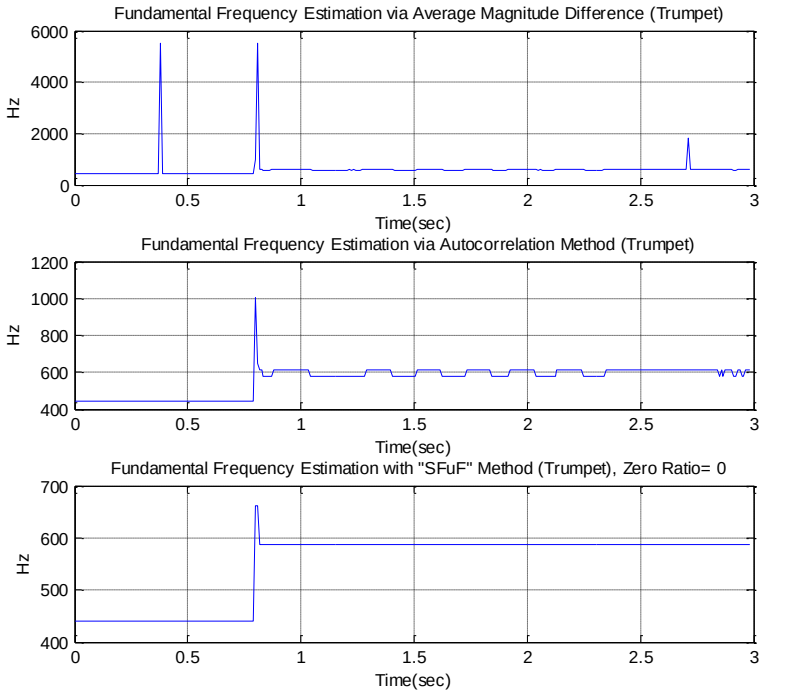
Στα σχήματα 15 έως 21 φαίνονται οι καμπύλες της SFuF για έξι παραδείγματα ήχων και παρουσιάζονται επίσης οι αντίστοιχες καμπύλες εκτίμησης της θεμελιώδους συχνότητας με τις μεθόδους της μέσης διαφοράς έντασης (average magnitude difference) και της αυτοσυσχέτισης (autocorrelation) για λόγους σύγκρισης των τριών εκτιμητικών μεθόδων.



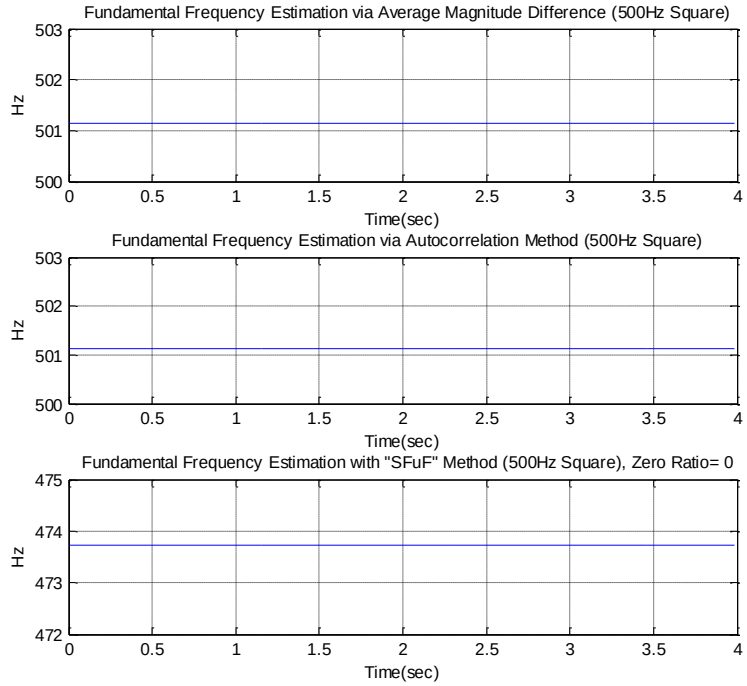
Σχήμα 15: Εκτίμηση της θεμελιώδους συχνότητας για ένα τμήμα ήχου τύμπανου



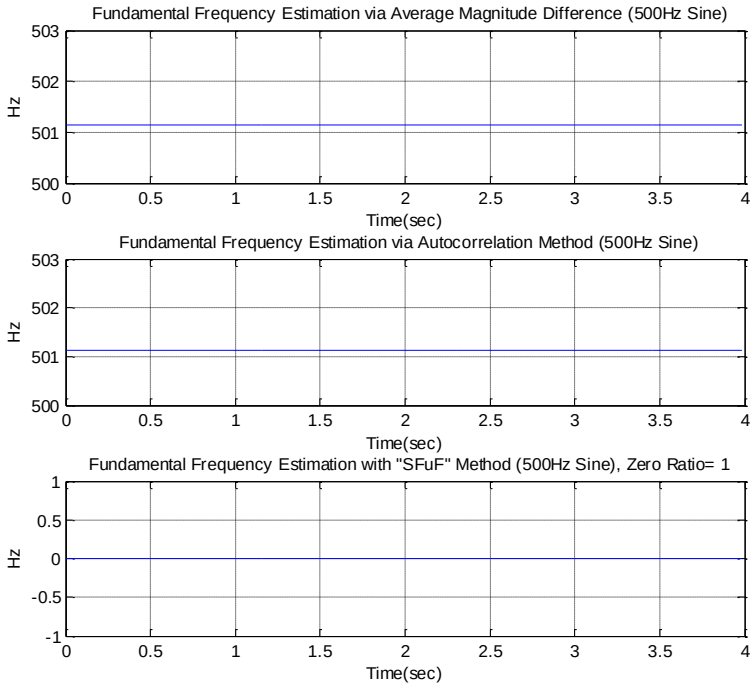
Σχήμα 16: Εκτίμηση της θεμελιώδους συχνότητας για ένα τμήμα φωνής



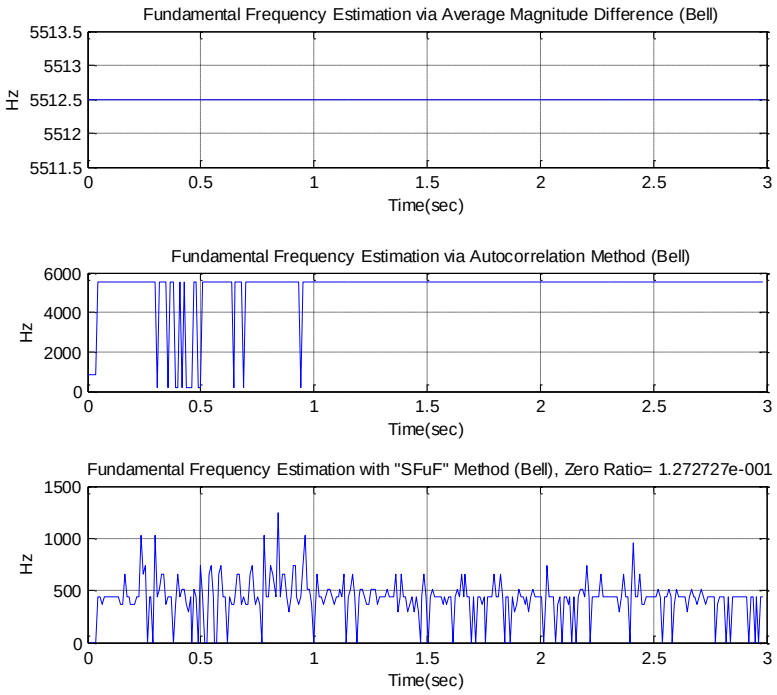
Σχήμα 17: Εκτίμηση της θεμελιώδους συχνότητας για ένα τμήμα ήχου τρομπέτας



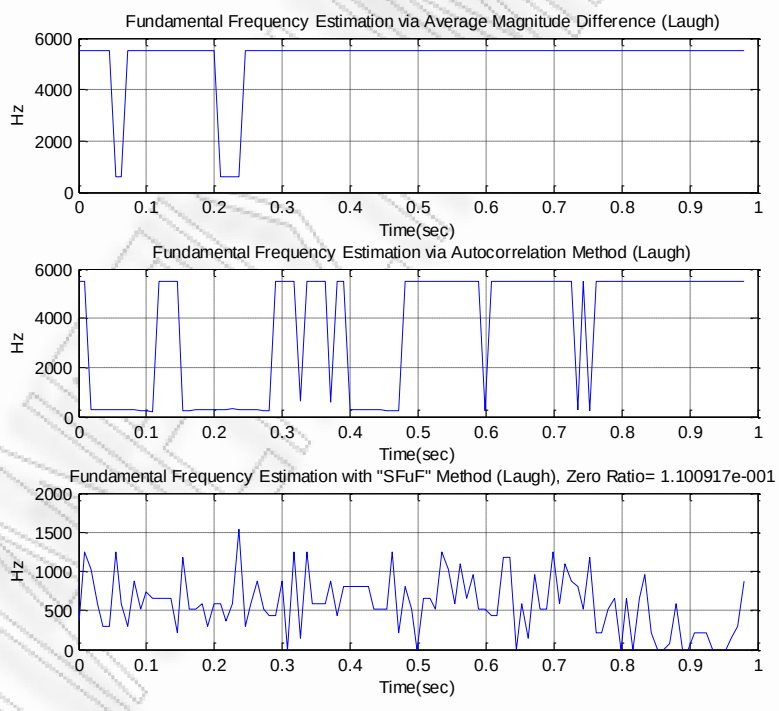
Σχήμα 18: Εκτίμηση της θεμελιώδους συχνότητας για έναν τετραγωνικό παλμό 500Hz



Σχήμα 19: Εκτίμηση της θεμελιώδους συχνότητας για ένα ημιτονικό σήμα 500Hz



Σχήμα 20: Εκτίμηση της θεμελιώδους συχνότητας για ένα τμήμα ήχου καμπάνας



Σχήμα 21: Εκτίμηση της θεμελιώδους συχνότητας για ένα τμήμα ήχου γέλιου

Σε κάθε μία από τις παραπάνω καμπύλες SFuF σημειώνεται ο «λόγος του μηδενός» (zero ratio) αυτού του τμήματος ήχου, ο οποίος ορίζεται ως ο λόγος του αριθμού των

πλαισίων με μηδενική τιμή SFuF προς το συνολικό αριθμό πλαισίων της καμπύλης SFuF:

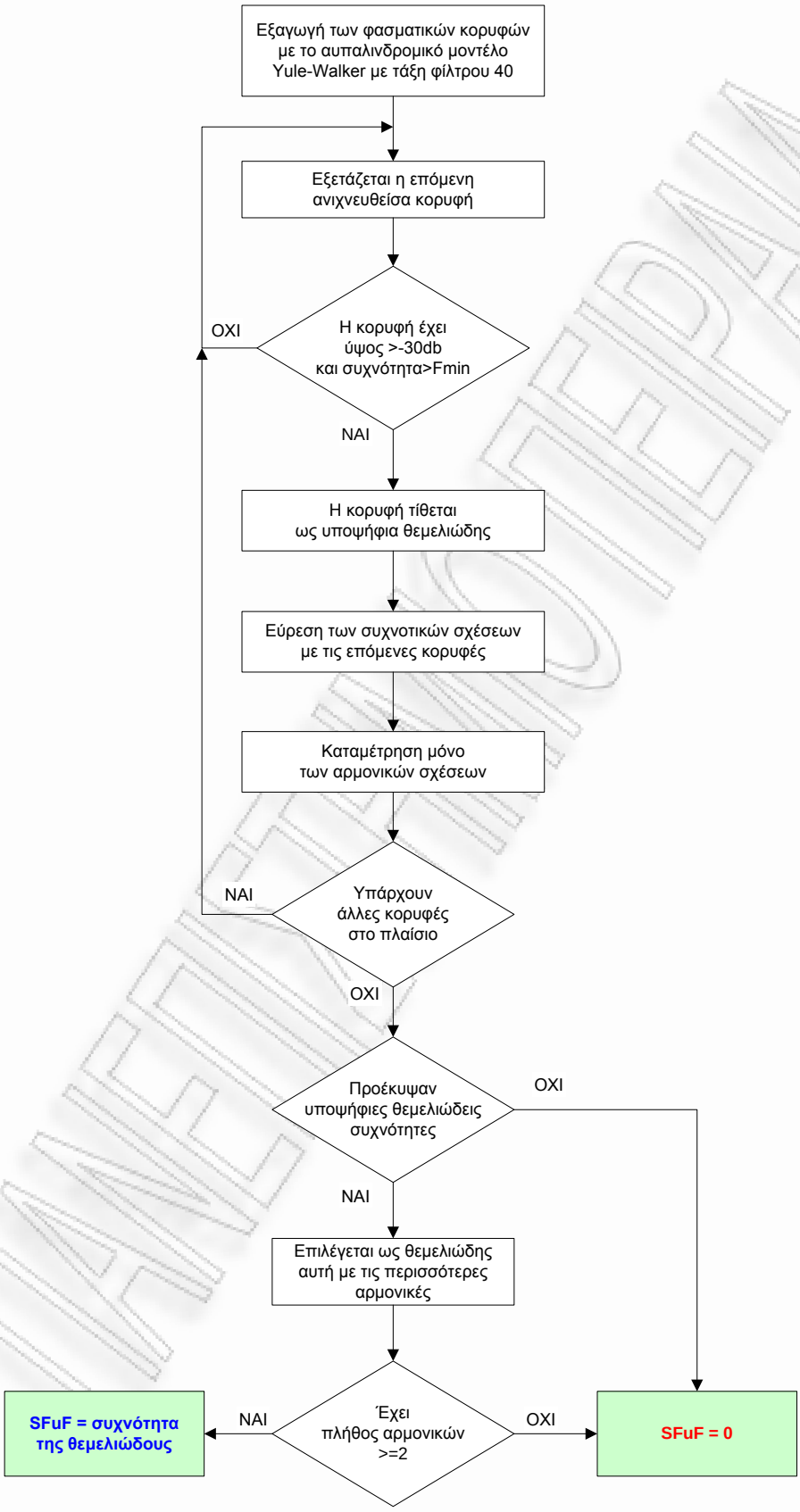
$$Zero\_ratio = \frac{\text{Αριθμός\_πλαισίων\_με\_SFuF} = 0}{\Sigma υνολικός\_αριθμός\_πλαισίων}$$

Από τα παραπάνω σχήματα μπορεί να φανεί ότι η μουσική είναι γενικά συνεχώς αρμονική. Επίσης, η θεμελιώδης συχνότητά της αλλάζει συνήθως πιο αργά σε σχέση με άλλα είδη ήχων, και η τιμή της SFuF τείνει να επικεντρωθεί σε μια συγκεκριμένη συχνότητα για μια μικρή χρονική περίοδο. Στην καμπύλη SFuF του σήματος της φωνής, εμφανίζονται εναλλάξ αρμονικά και μη αρμονικά μέρη καθώς τα έμφωνα μέρη της ομιλίας είναι αρμονικά ενώ τα άφωνα είναι μη αρμονικά. Η θεμελιώδης συχνότητα των εμφώνων μερών βρίσκεται κανονικά στην περιοχή των 100-300Hz. Οι περισσότεροι περιβαλλοντικοί ήχοι, είναι μη αρμονικοί με λόγους του μηδενός πάνω από 0.9. Ένα παράδειγμα μικτού αρμονικού και μη αρμονικού ήχου είναι ο ήχος του γέλιου, στον οποίο τα έμφωνα μέρη είναι αρμονικά, ενώ οι ενδιάμεσες παύσεις καθώς επίσης και τα μεταβατικά μέρη είναι μη αρμονικά. Έχει λόγο μηδενός 0.11 που είναι παρόμοιος με αυτόν της φωνής.

Παρατηρούμε επίσης ότι στην περίπτωση του ημιτονικού σήματος, η μέθοδος εκτιμά ορθά σαν θεμελιώδη συχνότητα το μηδέν καθότι το ημιτονικό σήμα στερείται αρμονικών συχνοτήτων. Τέλος, στην περίπτωση του τετραγωνικού παλμού, που έχει περιττές αρμονικές συχνότητες, η μέθοδος εκτιμά με αρκετή ακρίβεια την τιμή της θεμελιώδους συχνότητας (~473.7Hz αντί για την πραγματική που είναι 500Hz).

Σημείωση: Στην τεκμηρίωση της πρωτότυπης εργασίας υπάρχουν ελλείψεις, π.χ. δεν αναφέρεται ποιο αυτοπαλινδρομικό μοντέλο χρησιμοποιήθηκε,, όπως επίσης και ασάφειες, π.χ. αναφέρεται αορίστως ότι: *«εξετάζεται αν κάποιες από τις θέσεις αυτών των κορυφών έχουν κοινό διαιρέτη και αν τουλάχιστον ορισμένες από αυτές έχουν οξύτητα, ύψος και πλάτος που να ικανοποιούν κάποια συγκεκριμένα κριτήρια».* Παρά τις δυσκολίες αυτές, προσπαθήσαμε και πετύχαμε να υλοποιήσουμε τη διαδικασία εκτίμησης της θεμελιώδους συχνότητας όσο το δυνατόν πλησιέστερα προς την αδρά περιγραφόμενη διαδικασία στην πρωτότυπης εργασίας, και έτσι πετύχαμε αριθμητικά αποτελέσματα παρόμοια με τα δημοσιευθέντα. Η διαδικασία που υλοποιήσαμε για την εκτίμηση της θεμελιώδους συχνότητας (SfuF) φαίνεται στο διάγραμμα ροής του σχήματος 22 και είναι η εξής:

1. Εξάγεται το φάσμα με τη μέθοδο Yule-Walker, με τάξη φίλτρου 40.
2. Οι συχνότητες όλων των κορυφών με ύψος πάνω από ένα κατώφλι ανιχνεύονται ως υποψήφιες θεμελιώδεις συχνότητες. **Το κατώφλι ύψους βρέθηκε πειραματικά να είναι -30db** σε σχέση με το μέγιστο ύψος των κορυφών.
3. Για κάθε μία υποψήφια θεμελιώδη συχνότητα καταμετράται ο αριθμός των επομένων από αυτήν κορυφών που η συχνότητά τους έχει αρμονική σχέση με την υποψήφια θεμελιώδη συχνότητα.
4. Αν δεν προκύψουν υποψήφιες θεμελιώδεις συχνότητες, τότε ως τιμή της θεμελιώδους συχνότητας τίθεται το μηδέν.
5. Από όλες τις υποψήφιες θεμελιώδεις συχνότητες επιλέγεται αυτή που βρέθηκε να έχει τις περισσότερες αρμονικές.
6. Τέλος, εξετάζεται αν το πλήθος των αρμονικών της υποψήφιας θεμελιώδους είναι πάνω από μία τιμή κατωφλίου, αλλιώς τίθεται ως θεμελιώδης συχνότητα το μηδέν. Βρέθηκε πειραματικά ότι **το κατώφλι του πλήθους των αρμονικών είναι 2**.



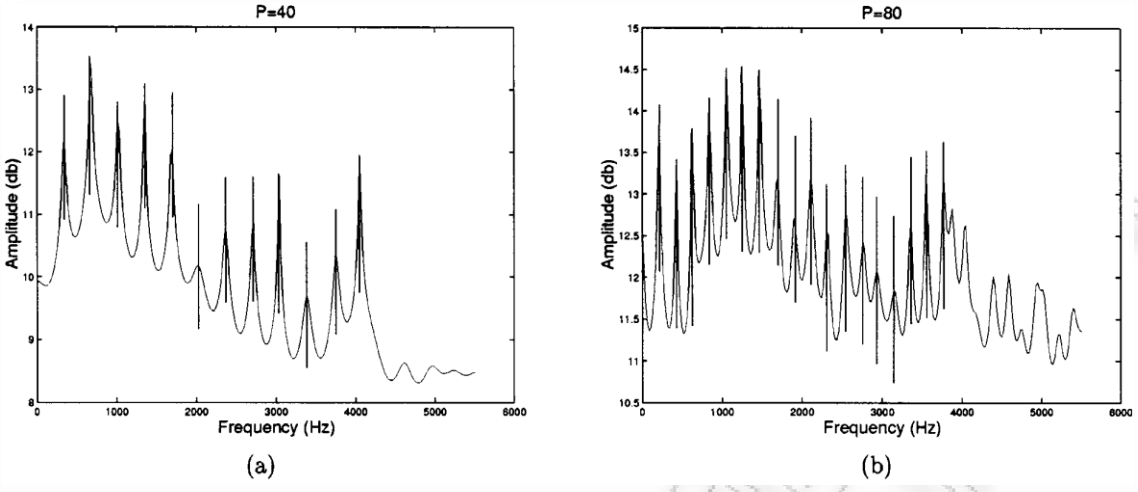
Σχήμα 22: Υπολογισμός της θεμελιώδους συχνότητας ενός πλαισίου

5.4 Ίχνη των φασματικών κορυφών (Spectral peak Track)

Τα ίχνη των κορυφών σε ένα φασματογράφημα ηχητικού σήματος, αποκαλύπτουν συχνά τα χαρακτηριστικά του τύπου του ήχου. Για παράδειγμα, τα ίχνη των φασματικών κορυφών των μουσικών οργάνων παραμένουν στα ίδια επίπεδα συχνότητας και έχουν συγκεκριμένη χρονική διάρκεια. Τα ίχνη των αρμονικών κορυφών της ανθρώπινης φωνής ευθυγραμμίζονται σε μορφή χτένας. Τα ίχνη των φασματικών κορυφών στα τμήματα με τραγούδι μπορεί να εκτείνονται σε ευρεία περιοχή συχνοτήτων, και η θεμελιώδης συχνότητά τους να εκτείνεται από 87Hz έως 784Hz. Στα τραγούδια υπάρχουν σχετικά μακρά ίχνη που είναι σταθερά λόγω του ότι η φωνή μπορεί να παραμείνει σε μία συγκεκριμένη νότα για μια χρονική περίοδο, και έχουν συχνά κυματοειδές σχήμα λόγω της δόνησης των φωνητικών χορδών. Τα ίχνη των φασματικών κορυφών στα τμήματα με ομιλία ευρίσκονται κανονικά σε χαμηλότερες ζώνες συχνοτήτων, και είναι πιο κοντά μεταξύ τους λόγω του ότι η περιοχή συχνοτήτων της θεμελιώδους συχνότητας είναι 100-300Hz. Τείνουν επίσης να έχουν μικρότερο μήκος διότι υπάρχουν διακοπές μεταξύ των εκφερόμενων συλλαβών, κα μπορεί να παρουσιάζουν μια αργή διακύμανση λόγω του ότι το ύψος μπορεί να αλλάζει κατά τη διάρκεια της προφοράς ορισμένων συλλαβών.

Σε αυτή την εργασία, εξάγουμε τα ίχνη των φασματικών κορυφών για να χαρακτηρίσουμε έναν ήχο ως τραγούδι ή ως ομιλία. Αυτό γίνεται ανιχνεύοντας τις κορυφές στο φάσμα που δημιουργείται από τις παραμέτρους του AR μοντέλου και εξετάζοντας τις αρμονικές σχέσεις μεταξύ των κορυφών. Η έκταση της θεμελιώδους συχνότητας των υπό εξέταση αρμονικών κορυφών τίθεται στους 80Hz-800Hz λόγω αυτής της ιδιότητας του τραγουδιού και της ομιλίας. Η ανάλυση συχνότητας ενός FFT 512 σημείων είναι αρκετή για να ανιχνευθούν αρμονικές κορυφές σε αυτή την περιοχή συχνοτήτων αν παράλληλα εκλεγεί κατάλληλα και η τάξη του AR μοντέλου. Για παράδειγμα, με τάξη P=40, είναι δυνατόν να ανιχνευθούν εύκολα αρμονικές κορυφές που έχουν θεμελιώδη συχνότητα μεγαλύτερη από 250Hz, πράγμα που ταιριάζει για τα περισσότερα μουσικά τμήματα. Όμως, αυτή η ανάλυση δεν είναι αρκετή για τα περισσότερα τμήματα με ανδρική ή γυναικεία φωνή. Βρήκαμε πειραματικά, ότι η τάξη P=80 ήταν κατάλληλη για τη γυναικεία φωνή (με ύψος περίπου 150-250Hz), αλλά η ανδρική φωνή όπου το ύψος της είναι μεταξύ 100-150Hz ίσως απαιτήσει τάξη P=100. Ωστόσο, με αυτές τις υψηλότερες τιμές της τάξης P, θα εμφανίζονται τεχνηματικές (artifact) κορυφές στο εκτιμώμενο φάσμα των ήχων που έχουν υψηλότερη θεμελιώδη συχνότητα, η ύπαρξη των οποίων μπορεί να επηρεάσει σοβαρά την ποιότητα της ανίχνευσης των κορυφών αυτών των ήχων.

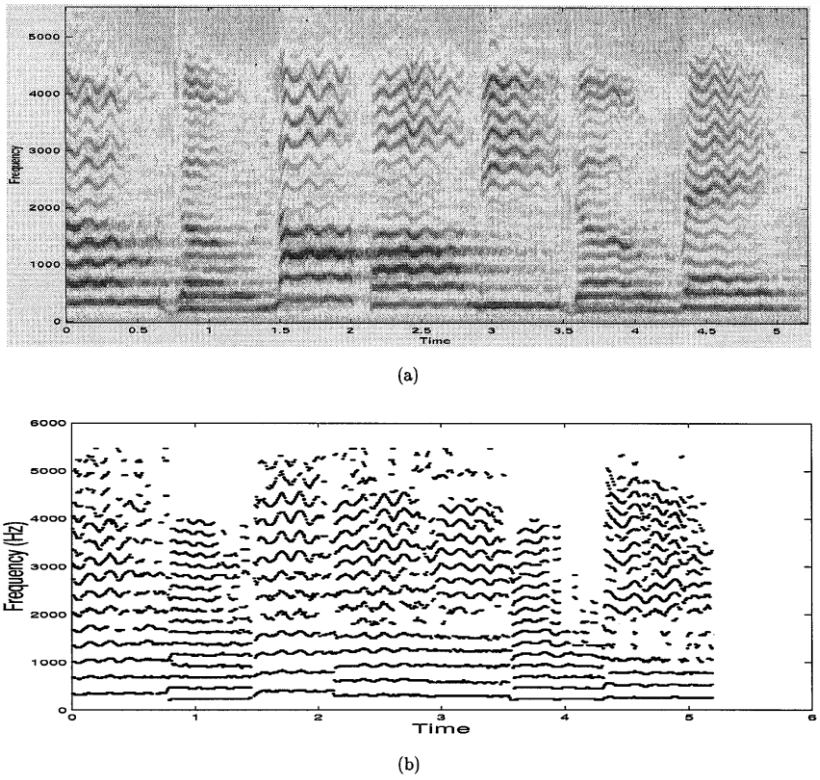
Θέτουμε την τάξη του AR μοντέλου σε τρεις τιμές: 40, 80 και 100. Η ιδέα είναι ότι το μοντέλο πρέπει να μπορέσει να ανιχνεύσει αρμονικές κορυφές με μία από αυτές τις τάξεις για τους ήχους που μας ενδιαφέρουν. Η διαδικασία του καθορισμού της κατάλληλης τάξης αναπτύσσεται παρακάτω. Αν έχουν ανιχνευθεί αρμονικές κορυφές στο φάσμα του προηγουμένου πλαισίου που δημιουργήθηκε από το AR μοντέλο τάξης P1 (το P1 μπορεί να είναι 40, 80, ή 100), τότε αρχίζουμε την ανίχνευση των αρμονικών κορυφών στο φάσμα του τρέχοντος πλαισίου που δημιουργήθηκε με τάξη P1. Αν βρεθούν αρμονικές κορυφές στο φάσμα, προχωρούμε στο επόμενο πλαίσιο. Αλλιώτικα, προσπαθούμε με τα φάσματα που δημιουργήθηκαν με τις άλλες δύο τάξεις. Αν δεν έχουν ανιχνευθεί αρμονικές κορυφές στο προηγούμενο πλαίσιο, δοκιμάζουμε για το τρέχον πλαίσιο τις τρεις τάξεις μία-μία μέχρις ότου βρεθούν αρμονικές κορυφές ή να εξαχθεί το συμπέρασμα ότι δεν υπάρχουν αρμονικές κορυφές. Οι αρμονικές κορυφές πρέπει να έχουν μεταξύ τους αρμονικές σχέσεις και να ικανοποιούν ορισμένες συνθήκες οξύτητας, ύψους και πλάτους. Δεδομένου ότι υπάρχουν πολλές νόθες κορυφές στα φάσματα που δημιουργούνται με τάξη P=80 ή 100, προσθέτουμε σε αυτές τις περιπτώσεις, βασιζόμενοι στα χαρακτηριστικά των σημάτων της ομιλίας, τους περιορισμούς ότι οι αρμονικές κορυφές πρέπει να ευθυγραμμίζονται διαδοχικά στη χαμηλομεσαία περιοχή συχνοτήτων και ότι η θεμελιώδης συχνότητα πρέπει να είναι κάτω από 250Hz,. Επίσης, εφαρμόζουμε ένα επίπεδο εμπιστοσύνης στο αποτέλεσμα της ανίχνευσης όταν η τάξη είναι P=40, το οποίο τίθεται στο 1 όταν οι ευρεθείσες αρμονικές κορυφές ικανοποιούν ορισμένα κριτήρια, διαφορετικά αυτό τίθεται στο 0. Αν το επίπεδο εμπιστοσύνης είναι 1, προχωρούμε στο επόμενο πλαίσιο του σήματος. Διαφορετικά, επιχειρούμε να ανιχνεύσουμε αρμονικές κορυφές με υψηλότερη τάξη (π.χ P=80 και 100). Αν δεν βρεθούν φασματικές κορυφές ούτε σε αυτά τα φάσματα, ερχόμαστε πίσω και κρατάμε το αποτέλεσμα για P=40. Διαφορετικά, υιοθετούμε τις αρμονικές κορυφές που ανιχνεύθηκαν σε ένα φάσμα μεγαλύτερης τάξης.



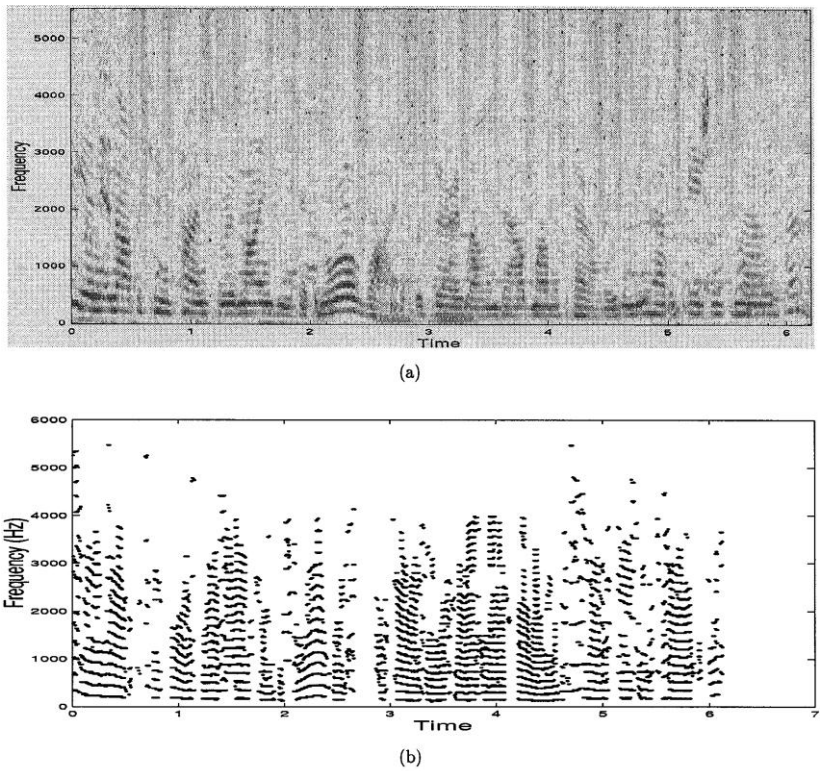
Σχήμα 23: Ανίχνευση αρμονικών κορυφών στο φάσμα που δημιουργείται με τις παραμέτρους του AR μοντέλου για τμήματα τραγουδιού και φωνής: (a) γυναικείο τραγούδι με P=40 και (b) γυναικεία φωνή με P=80. P είναι η τάξη του AR μοντέλου

Στο Σχήμα 23 φαίνονται οι αρμονικές κορυφές που ανιχνεύθηκαν με την παραπάνω διαδικασία για δύο πλαίσια τραγουδιού και φωνής αντίστοιχα, όπου κάθε ανιχνευθείσα κορυφή είναι σημειωμένη με μια κάθετη γραμμή.

Οι αρμονικές κορυφές ανιχνεύονται για κάθε 100 δείγματα εισόδου, και κάθε πλαίσιο περιέχει 512 δείγματα. Οι θέσεις των ανιχνευθέντων κορυφών τοποθετούνται σε χρονική σειρά για να αποτελέσουν τα ίχνη των φασματικών κορυφών. Προκειμένου να διορθώσουμε τα σφάλματα ανίχνευσης, εφαρμόζουμε στα δημιουργηθέντα ίχνη δύο στάδια μεταεπεξεργασίας. Το πρώτο στάδιο καλείται «σύνδεση» (linking), κατά το οποίο προστίθενται τα σημεία που λείπουν στα ίχνη έτσι ώστε αυτά να γίνουν πλήρη. Αυτό γίνεται αναζητώντας στα ίχνη οπές εύρους ενός έως τριών δειγμάτων. Αυτά τα απόντα σημεία μπορεί να προέρχονται από ασθενείς ή αλληλοεπικαλυπτόμενες αρμονικές κορυφές που είναι δύσκολο να ανιχνευθούν. Το δεύτερο στάδιο καλείται «καθάρισμα» (cleaning), κατά το οποίο αφαιρούνται μεμονωμένα σημεία που είναι έξω από τη γραμμή οποιουδήποτε ίχνους. Τα φασματογραφήματα και τα ίχνη των φασματικών κορυφών που εκτιμώνται με την προτεινόμενη μέθοδο για δύο τμήματα τραγουδιού και φωνής φαίνονται στο Σχήμα 24 και Σχήμα 25, αντίστοιχα.



Σχήμα 24: Φασματογράφημα και ιχνηλάτηση φασματικών κορυφών ενός γυναικείου φωνητικού σόλο.



Σχήμα 25: Φασματογράφημα και ιχνηλάτηση φασματικών κορυφών μιας γυναικείας φωνής με μουσική και θόρυβο στο υπόβαθρο.

Το Σχήμα 24 αφορά ένα γυναικείο φωνητικό σόλο επτά νότων με σειρά «5-1-6-4-3-1-2». Μπορούμε να δούμε ότι το ύψος και η διάρκεια κάθε νότας αντικατοπτρίζονται με σαφήνεια στα ανιχνευθέντα ίχνη. Κάθε νότα διαρκεί περίπου 0.7-0.8sec. Τα αρμονικά ίχνη εκτείνονται από τη θεμελιώδη συχνότητα, που είναι περίπου 225-400Hz, μέχρι τα 5000Hz, και έχουν κυματοειδή μορφή. Το Σχήμα 25 αφορά μια γυναικεία φωνή που συνοδεύεται στο υπόβαθρο με μουσική και άλλους θορύβους. Εν τούτοις, το σήμα της φωνής είναι δεσπόζον στο φασματογράφημα, και τα ίχνη των φασματικών κορυφών ανιχνεύονται πολύ καλά παρά την παρεμβολή της μουσικής και του θορύβου. Τα αρμονικά ίχνη στην περιοχή των 150-250Hz είναι εδώ βραχύτερα σε σχέση με αυτά του τμήματος της μουσικής.

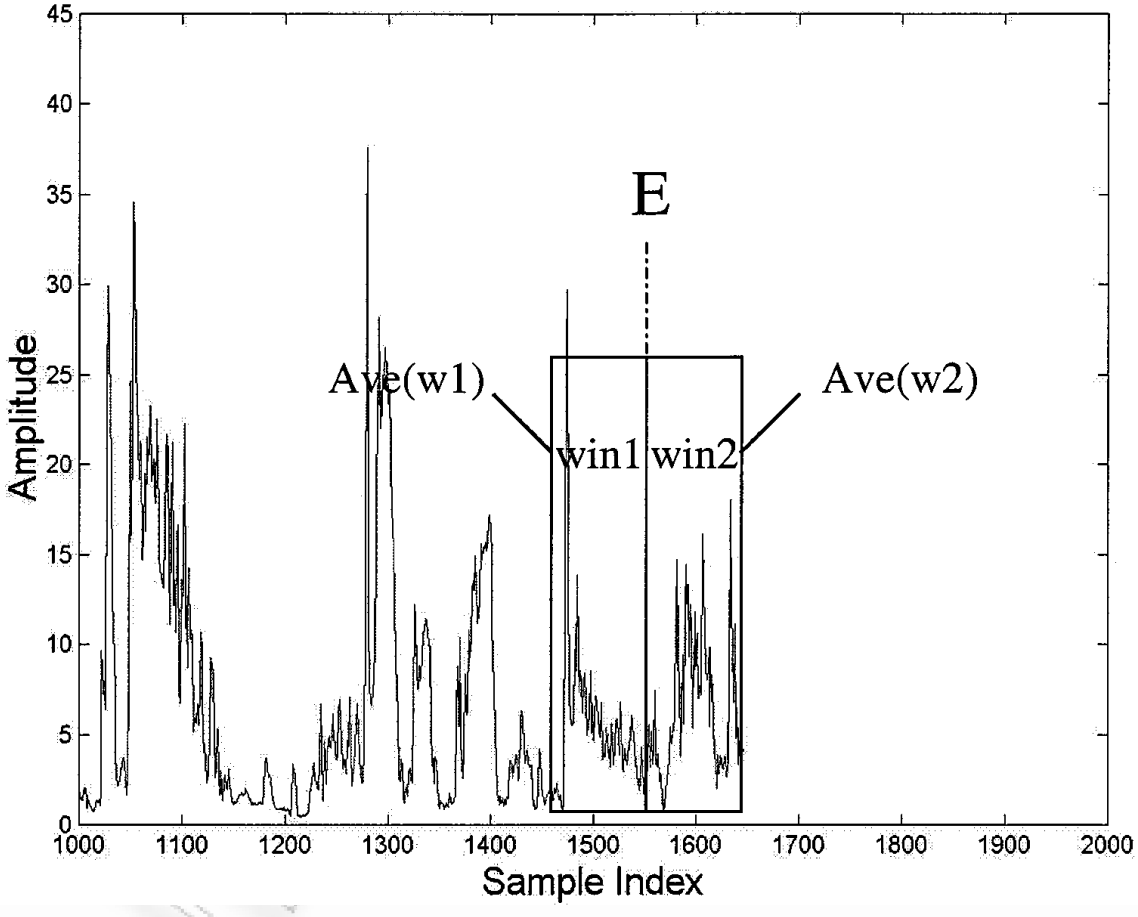
Σημείωση*: Η ιχνηλάτηση των αρμονικών κορυφών χρησιμοποιείται στη διάκριση των κατηγοριών που περιγράφονται στις παραγράφους 6.2.5 και 6.2.6. Όμως, λόγω των ασαφειών που υπάρχουν στην πρωτότυπη εργασία, όσον αφορά την ακριβή περιγραφή της διαδικασίας ταξινόμησης αυτών των κατηγοριών, δεν υλοποιήσαμε το παρόν τμήμα της ιχνηλάτησης των αρμονικών κορυφών διότι δεν θα μπορούσε να αξιοποιηθεί παρακάτω.*

6. ΚΑΤΑΤΜΗΣΗ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΤΗΣ ΗΧΗΤΙΚΗΣ ΡΟΗΣ

6.1. Ανίχνευση των ορίων των τμημάτων

Για την κατάτμηση των οπτικοακουστικών δεδομένων σε πραγματικό χρόνο, υπολογίζονται στον αέρα (on the fly) οι τιμές βραχέως χρόνου της ενέργειας, του ρυθμού διέλευσης του μηδενός και της θεμελιώδους συχνότητας των εισερχομένων ηχητικών δεδομένων. Οσάκις ανιχνευθεί μια απότομη αλλαγή σε οποιοδήποτε από αυτά τα χαρακτηριστικά, τίθεται η αρχή ενός νέου τμήματος. Πιο αναλυτικά:

Στη χρονική καμπύλη καθενός χαρακτηριστικού υπάρχουν δύο εφαπτόμενα ολισθαίνοντα παράθυρα, σε καθένα από τα οποία υπολογίζεται η μέση τιμή του χαρακτηριστικού, όπως φαίνεται στο Σχήμα 26.



Σχήμα 26: Τοποθέτηση των ολισθαινόντων παραθύρων στη χρονική καμπύλη του ηχητικού χαρακτηριστικού για την ανίχνευση των ορίων.

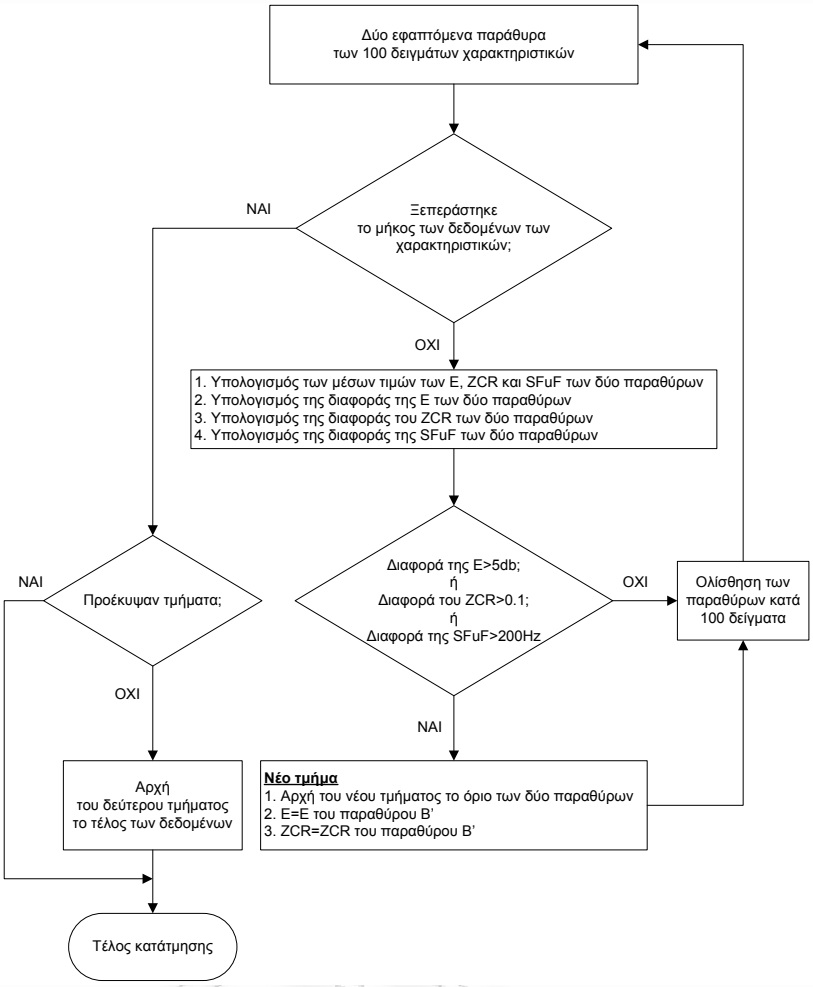
Τα παράθυρα ολισθαίνουν μαζί και κάθε φορά που υπολογίζεται μια νέα τιμή χαρακτηριστικού ενημερώνονται οι αντίστοιχες μέσες τιμές Ave(w1) και Ave(w2).

Αυτές οι δύο τιμές συγκρίνονται. Οσάκις υπάρχει μια μεγάλη διαφορά μεταξύ τους, θεωρείται ότι ανιχνεύθηκε μια απότομη αλλαγή στο κοινό τους όριο (π.χ. το σημείο Ε στο Σχήμα 26). Επιλέγουμε το μέγεθος κάθε παραθύρου να είναι 100 δείγματα χαρακτηριστικού, που αντιστοιχεί σε χρόνο περίπου 1sec για συχνότητα δειγματοληψίας 11.025Hz. Καθώς τα πρότυπα της χρονικής εξέλιξης και οι περιοχές τιμών αυτών των χαρακτηριστικών είναι διαφορετικά για τη φωνή, τη μουσική, τους περιβαλλοντικούς ήχους κ.λ.π., είναι δυνατόν, εφαρμόζοντας στατιστική ανάλυση σε αυτά τα χαρακτηριστικά, να ανιχνευθούν δραματικές αλλαγές στα όρια των περιοχών μεταξύ δύο διαφορετικών ηχητικών τύπων.

Σημείωση 1*: Το λογισμικό που υλοποιήσαμε δεν ταξινομεί ροές δεδομένων σε πραγματικό χρόνο αλλά αρχεία. Έτσι, πριν ξεκινήσει η επεξεργασία των δεδομένων γίνεται η εξής προετοιμασία τους:*

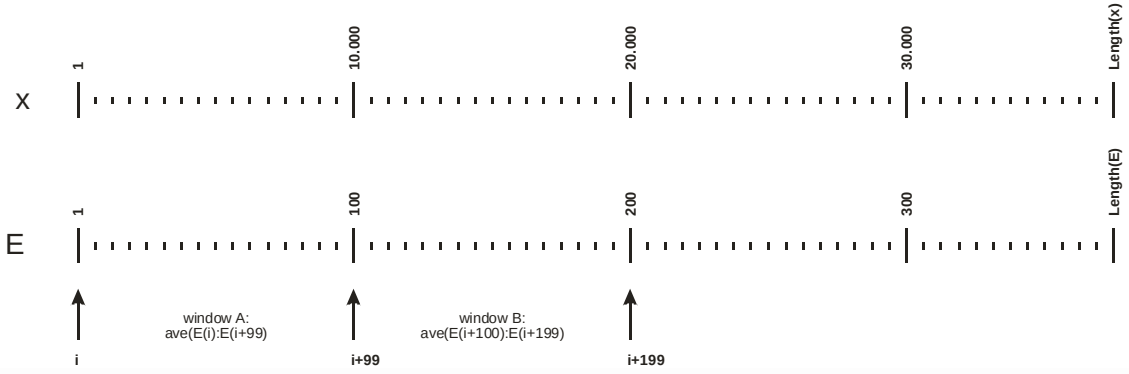
i. *Μετατροπή του στερεοφωνικού ήχου σε μονοφωνικό. Αν το αρχείο περιείχε πληροφορία για δύο κανάλια ήχου, τότε τα δύο κανάλια μετατρέπονται σε ένα ημιαθροίζοντας τις αντίστοιχες τιμές τους.*
ii. *Επαναδειγματοληψία στα 11025Hz. Αν τα δεδομένα του υπό εξέταση αρχείου έχουν συχνότητα δειγματοληψίας διαφορετική από 11025Hz, επαναδειγματοληπτούνται στα 11025Hz για να υπάρχει συμφωνία με την πρωτότυπη εργασία στην οποία αναφέρεται επανειλημένα ότι η συχνότητα δειγματοληψίας είναι 11025Hz.*
iii. *Κανονικοποίηση στη μονάδα. Αναπροσαρμόζονται οι τιμές των δειγμάτων έτσι ώστε η τιμή του μεγαλύτερου δείγματος να γίνει 1.*

Σημείωση 2*: Ο αλγόριθμος κατάτμησης που υλοποιήσαμε φαίνεται στο Σχήμα 27 και αναλύεται κατωτέρω:*



Σχήμα 27: Ο υλοποιηθείς αλγόριθμος κατάτμησης

Η ανίχνευση των ορίων γίνεται με δύο εφαπτόμενα ολισθαίνοντα παράθυρα. Τα δύο παράθυρα, όπως φαίνεται στο Σχήμα 28, έχουν μήκος 100 δείγματα χαρακτηριστικού (E, ή ZCR,ή SFuF), που αντιστοιχεί σε 10.000 δείγματα δεδομένων, δηλαδή σε περίπου 1sec. Τα παράθυρα ολισθαίνουν κατά 100 δείγματα χαρακτηριστικού, δηλαδή κατά 1sec περίπου, άρα και η διακριτική ικανότητα του συστήματός μας θα είναι περίπου 1sec που είναι κατά πολύ ανώτερη από αυτή που επιζητούμε.



Σχήμα 28: Τα ολισθαίνοντα παράθυρα

Για κάθε παράθυρο υπολογίζεται η μέση τιμή της ενέργειας, του ZCR και της SFuF, υπολογίζεται η διαφορά των αντιστοίχων τιμών των τριών χαρακτηριστικών για τα δύο παράθυρα και αν η διαφορά είναι μεγαλύτερη από μία τιμή κατωφλίου, τότε θεωρείται ότι ανιχνεύθηκε μια μεγάλη αλλαγή στο σήμα και τίθεται το κοινό όριο των δύο παραθύρων ως η αρχή ενός νέου τμήματος. ***Οι τιμές κατωφλίου βρέθηκαν πειραματικά να είναι: ΔZCR>0.1, ΔE>5 και ΔSFuF>200.***

Σημείωση 3*: Στην πρωτότυπη εργασία αναφέρεται ότι οι μέσες τιμές των χαρακτηριστικών των δύο παραθύρων επαναϋπολογίζονται και συγκρίνονται κάθε φορά που υπολογίζεται μια νέα τιμή χαρακτηριστικού. Αν αυτή η διαδικασία υλοποιηθεί επακριβώς τότε ο υπολογιστικός χρόνος γίνεται πολύ μεγαλύτερος από τον πραγματικό χρόνο και η μέθοδος παύει να λειτουργεί σε πραγματικό χρόνο (τουλάχιστον για έναν μέσο σημερινό υπολογιστή), αντίθετα από αυτό που επαγγέλλεται το πρωτότυπο άρθρο. Εμείς επιλέξαμε επαναϋπολογισμό των μέσων τιμών των δύο παραθύρων για κάθε 100 τιμές χαρακτηριστικών και η μέθοδός μας καταφέρνει να ολοκληρώνει το έργο της στο μισό περίπου από τον πραγματικό χρόνο, άρα μπορεί με μια μικρή αλλαγή να ταξινομεί ροές ηχητικών δεδομένων σε πραγματικό χρόνο.*

6.2. Ταξινόμηση κάθε τμήματος

Αφού ανιχνευθούν τα όρια των τμημάτων, κάθε τμήμα ταξινομείται σε έναν από τους βασικούς τύπους ήχου σύμφωνα με τη διαδικασία που φαίνεται για τη μεν πρωτότυπη εργασία στο Σχήμα 1 για δε την παρούσα πτυχιακή εργασία στο Σχήμα 2. Οι λεπτομέρειες κάθε βήματος της ταξινομητικής διαδικασίας περιγράφονται παρακάτω.

6.2.1. Ανίχνευση σιγής

Το πρώτο βήμα είναι ο έλεγχος για το εάν το τμήμα του ήχου είναι σιγή ή όχι. Ορίζουμε ως «σιγή» ένα τμήμα μη αντιληπτού ήχου, συμπεριλαμβανομένου του μη αντιληπτού θορύβου και των πολύ σύντομων κλικ. Ο συνηθισμένος τρόπος ανίχνευσης της σιγής είναι με τη χρήση ενός κατωφλίου για την ενέργεια. Ωστόσο,, έχει βρεθεί ότι το επίπεδο της ενέργειας μερικών τμημάτων θορύβου δεν είναι χαμηλότερο από αυτό ορισμένων μουσικών κομματιών. Ο λόγος που σε κάποιες περιπτώσεις είναι δυνατόν να ακούμε τη μουσική ενώ δεν παρατηρούμε το θόρυβο

είναι ότι η συχνότητα του θορύβου μπορεί να είναι πολύ χαμηλότερη από αυτήν της μουσικής. Έτσι, για να ανιχνεύσουμε τη σιγή χρησιμοποιούμε τόσο την ενέργεια όσο και το ρυθμό διέλευσης του μηδενός. Αν η ενέργεια βραχέως χρόνου είναι διαρκώς χαμηλότερη από ένα σύνολο κατωφλίων (μπορεί να υπάρχουν διάρκειες κατά τις οποίες η ενέργεια να είναι υψηλότερη από το κατώφλι, αλλά αυτές πρέπει να είναι αρκούντως βραχείες και μακριά η μια από την άλλη), ή αν οι περισσότερες τιμές του ρυθμού διέλευσης του μηδενός είναι χαμηλότερες από ένα συγκεκριμένο σύνολο κατωφλίων, τότε αυτό το τμήμα ταξινομείται ως «σιγή».

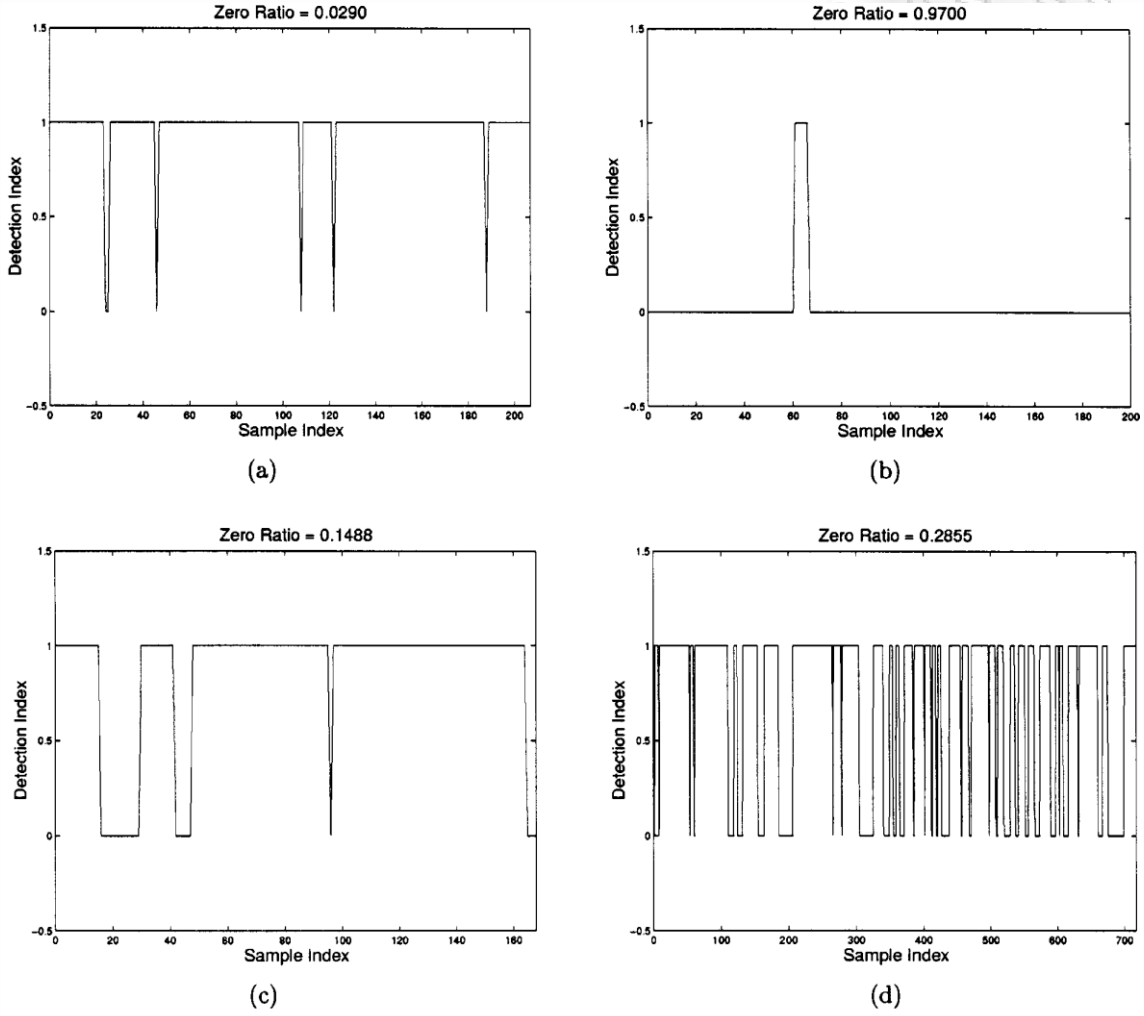
***Σημείωση**: Στην πρωτότυπη εργασία δεν αναφέρεται ποιο είναι αυτό το σύνολο κατωφλίων, τόσο για την ενέργεια όσο και για το ρυθμό διέλευσης του μηδενός. Στην εργασία μας, βρήκαμε πειραματικά ότι για να ταξινομηθεί ένα τμήμα ως σιγή πρέπει ο μέσος όρος της ενέργειάς του να είναι **κάτω από τα -40db** ή ο μέσος όρος του ρυθμού διέλευσης του μηδενός του να είναι **κάτω από το 0.01 (**τιμή που αντιστοιχεί στο ρυθμό διέλευσης του μηδενός ενός απλού σήματος συχνότητας περίπου 60Hz ή ενός σύνθετου σήματος τριών αρμονικών με θεμελιώδη συχνότητα μικρότερη των 30Hz).*

6.2.2. Διαχωρισμός των ήχων με ή δίχως μουσικές συνιστώσες

Όπως έχει παρατηρηθεί από τις κινηματογραφικές ταινίες και τα τηλεοπτικά προγράμματα, η μουσική είναι ένας σημαντικός ηχητικός τύπος που εμφανίζεται συχνά, είτε μόνη της ή σαν υπόκρουση στην ομιλία ή τους περιβαλλοντικούς ήχους. Έτσι, τα μη σιωπηλά ηχητικά τμήματα διαχωρίζονται πρώτα σε δύο κατηγορίες: με ή χωρίς μουσικές συνιστώσες, ανιχνεύοντας την ύπαρξη ή μη συνεχών και σταθερών κορυφών στο φάσμα ισχύος.

Το φάσμα δημιουργείται από τις παραμέτρους του AR μοντέλου τάξης 40, και υπολογίζεται για κάθε 400 δείγματα εισόδου. Κάθε πλαίσιο για τον υπολογισμό του φάσματος περιέχει 512 δείγματα. Αν υπάρχουν κορυφές που ανιχνεύονται σε διαδοχικά φάσματα, και που παραμένουν στην ίδια περίπου συχνότητα για συγκεκριμένη χρονική περίοδο, τότε αυτή η χρονική περίοδος ταξινομείται ως περιέχουσα μουσικές συνιστώσες. Για να αποφευχθεί η επίδραση των αρμονικών κορυφών της φωνής ή του χαμηλής συχνότητας θορύβου, **λαμβάνονται υπόψη μόνο οι κορυφές με συχνότητα πάνω από 500Hz** καθώς οι περισσότερες μουσικές συνιστώσες βρίσκονται σε αυτή την περιοχή. Αγνοούνται επίσης και τα πλαίσια του σήματος που έχουν ενέργεια κάτω από ένα επίπεδο. Δημιουργείται μια σειρά δεικτών για κάθε τμήμα ήχου, όπου η τιμή του δείκτη τίθεται στην τιμή 1 αν τη συγκεκριμένη στιγμή ανιχνεύονται μουσικές συνιστώσες ενώ διαφορετικά τίθεται στην τιμή 0. Ο λόγος του αριθμού των μηδενικών μιας σειράς δεικτών προς το συνολικό αριθμό των

δεικτών της σειράς (που καλείται «λόγος του μηδενός») μπορεί επομένως να αποτελέσει ένα μέτρο για το αν το ηχητικό τμήμα περιέχει μουσικές συνιστώσες ή όχι. Όσο υψηλότερος είναι αυτός ο λόγος, τόσο λιγότερες μουσικές συνιστώσες περιέχονται στον ήχο. Στο Σχήμα 29 φαίνονται οι σειρές των δεικτών για ηχητικά τμήματα διαφόρων ήχων.



Σχήμα 29: Σειρές δεικτών ανίχνευσης μουσικών συνιστωσών σε τμήματα ήχου: (α) καθαρή μουσική, (β) καθαρή φωνή, (γ) φωνή με μουσική υπόκρουση, και (δ) τραγούδι.

Εξετάζοντας τους λόγους του μηδενός για διάφορους τύπους ήχου, εξάγουμε τις εξής παρατηρήσεις:

- **.2.2.1.** ***Φωνή:*** Παρόλο που το σήμα της φωνής περιέχει μουσικές συνιστώσες, οι συχνοτικές κορυφές του αλλάζουν γρηγορότερα και διαρκούν λιγότερο σε σχέση με τις κορυφές της μουσικής. Οι λόγοι του μηδενός για τα τμήματα της φωνής είναι συνήθως άνω του 0.95.

- **.2.2.2.** ***Περιβαλλοντικοί ήχοι:*** Όλοι οι αρμονικοί και σταθεροί περιβαλλοντικοί ήχοι ταξινομούνται ως έχοντες μουσικές συνιστώσες, ενώ όλοι οι μη αρμονικοί ήχοι ταξινομούνται ως μη έχοντες μουσικές συνιστώσες. Όμως, υπάρχουν κατ΄ εξαίρεση μερικές ενδιάμεσες περιπτώσεις όπως είναι μερικά μίγματα αρμονικών και μη αρμονικών ήχων, για τους οποίους πρέπει να προσθέσουμε κάποιους κανόνες στο λογισμικό ώστε να τους ταξινομεί σωστά.

- **.2.2.3.** ***Καθαρή μουσική:*** Οι λόγοι του μηδενός για όλα τα μουσικά τμήματα είναι κάτω του 0.3. Τα σφάλματα ταξινόμησης προέρχονται από βραχείες νότες, από περάσματα χαμηλής έντασης ή χαμηλής συχνότητας, από μη αρμονικές συνιστώσες, και από τις παύσεις μεταξύ των νότων.

- **.2.2.4.** ***Τραγούδι:*** Τα περισσότερα τμήματα με τραγούδι έχουν λόγους μηδενός κάτω από 0.5. Τα τμήματα τραγουδιού στα οποία ανιχνεύεται εσφαλμένα ότι δεν έχουν μουσικές συνιστώσες προκύπτουν: από νότες μεγάλης διάρκειας όπου τα φασματικά ίχνη τους έχουν κυματοειδή αντί για ευθύγραμμη μορφή, από τις παύσεις μεταξύ των νότων, και από τους χαμηλής έντασης ή χαμηλής συχνότητας ήχους. Όταν τα κυματοειδή ίχνη των κορυφών ανιχνεύονται και ταξινομούνται ως μουσικές συνιστώσες, τότε μειώνονται σημαντικά οι λόγοι μηδενός των τμημάτων τραγουδιού.

- **.2.2.5.** ***Φωνή με μουσική υπόκρουση:*** Όταν το σήμα της φωνής είναι ισχυρό τότε η μουσική υπόκρουση είναι συνήθως κρυμμένη και δεν μπορεί να ανιχνευθεί. Όμως, οι μουσικές συνιστώσες μπορεί και πάλι να ανιχνευθούν κατά τη διάρκεια των παύσεων της ομιλίας ή όταν το μουσικό σήμα γίνει ισχυρότερο. Διακρίνουμε επομένως τις επόμενες δύο περιπτώσεις. Στην πρώτη περίπτωση, η μουσική ή είναι ισχυρότερη από τη φωνή ή υπάρχουν πολλές παύσεις της ομιλίας έτσι ώστε η μουσική να είναι προεξάρχουσα. Τότε οι λόγοι του μηδενός είναι κάτω από 0.6. Στη δεύτερη περίπτωση, η μουσική είναι αδύναμη ενώ η φωνή είναι ισχυρή και συνεχής, έτσι ώστε η φωνή να είναι η κύρια συνιστώσα και η μουσική να μπορεί να αγνοηθεί. Σε αυτή την περίπτωση, οι λόγοι του μηδενός είναι υψηλότεροι από 0.8.

Έτσι, βασιζόμενοι σε ένα κατώφλι για το λόγο του μηδενός περίπου στο 0.7 μαζί με μερικούς άλλους κανόνες, μπορούμε να διαχωρίσουμε τα ηχητικά τμήματα σε δύο κατηγορίες. Η πρώτη κατηγορία περιλαμβάνει τους αρμονικούς και σταθερούς περιβαλλοντικούς ήχους, την καθαρή μουσική, το τραγούδι, τη φωνή με μουσική υπόκρουση και τους περιβαλλοντικούς ήχους με μουσική υπόκρουση. Στη δεύτερη κατηγορία, υπάρχει η καθαρή φωνή και οι μη αρμονικοί περιβαλλοντικοί ήχοι. Για κάθε μία κατηγορία μπορεί να γίνει περαιτέρω ταξινόμηση.

Σημείωση: *Στην εργασία μας δεν δημιουργήσαμε τη σειρά δεικτών «0» και «1» ώστε να καταμετρήσουμε έπειτα τα μηδενικά για να υπολογιστεί ο λόγος του μηδενός, διότι αυτό θα αύξανε άσκοπα τον υπολογιστικό χρόνο και την πολυπλοκότητα του λογισμικού. Πιο απλά, καταμετρήσαμε απευθείας τις μηδενικές τιμές της SFuF.*

6.2.3. Ανίχνευση των αρμονικών περιβαλλοντικών ήχων

Από τους ήχους της πρώτης κατηγορίας, διαχωρίζονται πρώτα οι περιβαλλοντικοί ήχοι που είναι αρμονικοί και σταθεροί. Εξετάζεται η χρονική καμπύλη της θεμελιώδους συχνότητας βραχέως χρόνου. Αν τα περισσότερα μέρη της καμπύλης είναι αρμονικά, και η θεμελιώδης συχνότητα παραμένει σταθερή σε μία συγκεκριμένη τιμή, το τμήμα ταξινομείται ως «*αρμονικός δίχως αλλαγές*» (harmonic unchanged). Ένα τυπικό παράδειγμα αυτού του τύπου είναι οι ήχοι του τονικού τηλεφωνικού συστήματος (touch tone). Αν η θεμελιώδης συχνότητα ενός ηχητικού τμήματος μεταβάλλεται χρονικά μεταξύ ορισμένων μόνο τιμών, τότε αυτό ταξινομείται ως «*αρμονικός και σταθερός*» (harmonic and stable). Παραδείγματα αυτού του τύπου είναι ο ήχος του κουδουνιού της εξώπορτας (doorbell) και ο βομβητής (pager). Αυτό το βήμα ταξινόμησης πραγματοποιείται σε αυτό το σημείο σαν μια απομονωτική διαδικασία (screening process) των αρμονικών περιβαλλοντικών ήχων, έτσι ώστε αυτοί να μην παρεμβάλλονται κατά τη διάκριση της μουσικής.

6.2.4. Διάκριση της καθαρής μουσικής

Η καθαρή μουσική διακρίνεται με βάση τη στατιστική ανάλυση των καμπυλών του ρυθμού διέλευσης του μηδενός και της θεμελιώδους συχνότητας. Για το σκοπό αυτό, εξετάζονται τέσσερα χαρακτηριστικά: ο βαθμός αρμονικότητας, ο βαθμός συγκέντρωσης της θεμελιώδους συχνότητας σε συγκεκριμένες τιμές κατά τη διάρκεια μιας χρονικής περιόδου, η διακύμανση του ρυθμού διέλευσης του μηδενός, και η περιοχή του πλάτους του ρυθμού διέλευσης του μηδενός. Για κάθε χαρακτηριστικό, έχει καθοριστεί ένα εμπειρικό σύνολο κατωφλίων και μια τιμή απόφασης. Αν ξεπεραστεί το κατώφλι, η τιμή απόφασης τίθεται στο 1, διαφορετικά τίθεται σε μία κλασματική τιμή μεταξύ 0 και 1 αναλόγως της απόστασης από το κατώφλι. Στη συνέχεια, υπολογίζεται με προκαθορισμένα βάρη ο μέσος όρος των τεσσάρων τιμών απόφασης για να εξαχθεί έτσι η συνολική πιθανότητα του τμήματος να είναι καθαρή μουσική. Για να ταξινομηθεί ένα ηχητικό τμήμα ως «*καθαρή μουσική*», αυτή η πιθανότητα πρέπει να είναι πάνω από ένα συγκεκριμένο κατώφλι και επίσης τουλάχιστον τρεις από τις τιμές απόφασης να είναι πάνω από 0.5.

Σημείωση: *Λόγω της αόριστης περιγραφής του καθορισμού των εμπειρικών κατωφλίων και των βαρών εξαγωγής του μέσου όρου ήταν αδύνατον να αναπαράξουμε αυτό το τμήμα της πρωτότυπης εργασίας, και γι αυτό διακρίναμε την ύπαρξη μουσικού περιεχομένου μόνο από το κατώφλι του λόγου μηδενός και από το πλήθος των διαφορετικών τιμών της θεμελιώδους συχνότητας του τμήματος. Συγκεκριμένα, αν* ***η τιμή του λόγου του μηδενός είναι κάτω από 0.4*** *και* ***η τιμή της θεμελιώδους συχνότητας λαμβάνει πάνω από 15 διαφορετικές τιμές****, το τμήμα χαρακτηρίζεται ως «μουσική». Η τιμή του κατωφλίου και το πλήθος των τιμών της θεμελιώδους συχνότητας βρέθηκαν εμπειρικά και κατόπιν εκτεταμένων δοκιμών.*

6.2.5. Διάκριση του τραγουδιού

Μέχρι εδώ, έχουν απομείνει για ταξινόμηση από την πρώτη κατηγορία, το τραγούδι, η φωνή με μουσική υπόκρουση και οι περιβαλλοντικοί ήχοι με μουσική υπόκρουση. Για το σκοπό αυτό, εξάγουμε τα ίχνη των φασματικών κορυφών αυτών των τμημάτων, και διαχωρίζουμε αυτούς τους τρεις ηχητικούς τύπους βασιζόμενοι στη μορφολογική ανάλυση αυτών των ιχνών. Τα τμήματα με τραγούδι χαρακτηρίζονται από ένα από τα τρία χαρακτηριστικά γνωρίσματα: την κυματοειδή μορφή των ιχνών των αρμονικών κορυφών (λόγω της δόνησης των φωνητικών χορδών), τη μεγαλύτερη διάρκεια των ιχνών σε σχέση με τα ίχνη της φωνής, και τη θεμελιώδη συχνότητα των ιχνών που είναι μεγαλύτερη από 300Hz. Οι ομάδες των ιχνών εξετάζονται για το αν τους ταιριάζει κάποιο από αυτά τα τρία χαρακτηριστικά. Το τμήμα θα ταξινομηθεί ως «*τραγούδι*» είτε αν το άθροισμα των διαρκειών κατά τις οποίες τα ίχνη των αρμονικών κορυφών ικανοποιούν ένα από τα χαρακτηριστικά ισούται με μια ορισμένη τιμή ή αν σύγκρισή του με το συνολικό μήκος του τμήματος φτάσει σε ένα συγκεκριμένο λόγο. Τα κυματοειδή ίχνη ανιχνεύονται λαμβάνοντας τη διαφορά πρώτης τάξης (first-order difference) του ίχνους και εξετάζοντας το πρότυπο της σειράς που προκύπτει. Ένα πράγμα που πρέπει να επισημάνουμε είναι το ότι αν και υπάρχουν μερικά μουσικά όργανα, όπως το βιολί, που μπορούν να δημιουργούν κυματοειδή ίχνη κορυφών, αυτά τα ίχνη βρίσκονται κανονικά σε μεγαλύτερες περιοχές συχνοτήτων.

Σημείωση*: Λόγω της ελλιπούς και αόριστης περιγραφής της ανωτέρω διαδικασίας ήταν αδύνατον να αναπαράξουμε αυτό το τμήμα της πρωτότυπης εργασίας και έτσι δεν διακρίναμε την κατηγορία «τραγούδι». Αν σε ένα τμήμα υπάρχει τραγούδι, αυτό το τμήμα θα ταξινομηθεί είτε ως «φωνή», αν δεν υπάρχει*

συνοδεία μουσικής, ή ως «μουσική» αν υπάρχει μουσική υπόκρουση οπότε σε αυτή την περίπτωση ανιχνεύονται αρμονικές κορυφές.

6.2.6. Διάκριση Φωνής/Περιβαλλοντικών ήχων με μουσική υπόκρουση

Σύμφωνα με την αναφορά [21], «*όταν αναμειγνύονται ήχοι με φάσματα που περιέχουν κορυφές, σε κάθε κανάλι κυριαρχεί γενικά η ενέργεια από τη μία ή από την άλλη πηγή*». Συνεπώς, παρόλο που υπάρχει μουσική στο υπόβαθρο, όταν η φωνή είναι δυνατή, μπορεί να ανιχνευθούν τα ίχνη των αρμονικών κορυφών της φωνής παρά την ύπαρξη των μουσικών συνιστωσών. Για το σκοπό αυτό, εξετάζουμε τις ομάδες των ιχνών για να διαπιστώσουμε αν αυτές συγκεντρώνονται στις χαμηλές προς μεσαίες περιοχές συχνοτήτων (με θεμελιώδεις συχνότητες μεταξύ 100 και 300Hz) και αν έχουν μήκη μέσα σε μια συγκεκριμένη περιοχή. Αν υπάρχουν διάρκειες στις οποίες τα ίχνη των φασματικών κορυφών ικανοποιούν αυτά τα κριτήρια, τότε το τμήμα αυτό χαρακτηρίζεται ως «*φωνή με μουσική υπόκρουση*». Τελικά, ότι απομένει για ταξινόμηση από την πρώτη κατηγορία είναι τα τμήματα που έχουν μουσικές συνιστώσες αλλά δεν ικανοποιούν τα κριτήρια κανενός από τους παραπάνω ηχητικούς τύπους. Αυτά τα τμήματα ταξινομούνται ως «*περιβαλλοντικοί ήχοι με μουσική υπόκρουση».*

Σημείωση*: Λόγω της ελλιπούς περιγραφής της ανωτέρω διαδικασίας ήταν αδύνατον να αναπαράξουμε αυτό το τμήμα της πρωτότυπης εργασίας και έτσι η φωνή με μουσική υπόκρουση ή οι περιβαλλοντικοί ήχοι με μουσική υπόκρουση ταξινομούνται ως «μουσική», αφού ανιχνεύονται αρμονικές κορυφές που προέρχονται από τη μουσική υπόκρουση.*

6.2.7. Διάκριση της καθαρής φωνής

Από τους ήχους της δεύτερης κατηγορίας, διακρίνεται πρώτα η καθαρή φωνή και εξετάζονται πέντε χαρακτηριστικές συνθήκες. Η πρώτη συνθήκη είναι η σχέση μεταξύ των χρονικών καμπυλών του ρυθμού διέλευσης του μηδενός και της ενέργειας. Στα τμήματα με φωνή, η καμπύλη του ρυθμού διέλευσης του μηδενός έχει κορυφές για τις άφωνες συνιστώσες και κοιλάδες για τις έμφωνες συνιστώσες, ενώ η καμπύλη της ενέργειας έχει ακριβώς το αντίθετο, δηλαδή κορυφές για τις έμφωνες συνιστώσες και κοιλάδες για τις άφωνες. Έτσι, υπάρχει μεταξύ τους μια αντισταθμιστική σχέση. Για το σκοπό αυτό, ψαλιδίζουμε την καμπύλη του ρυθμού διέλευσης του μηδενός και την καμπύλη της ενέργειας στο ένα τρίτο του μέγιστου ύψους τους και απομακρύνουμε το χαμηλό τμήμα έτσι ώστε να απομείνουν μονάχα οι κορυφές των δύο καμπυλών. Έπειτα, υπολογίζουμε το εσωτερικό

γινόμενο των δύο εναπομεινάντων καμπυλών. Το γινόμενο αυτό, για τα τμήματα της φωνής έχει τιμή κανονικά κοντά στο μηδέν διότι οι κορυφές εμφανίζονται στις δύο καμπύλες σε διαφορετικές χρονικές στιγμές, ενώ έχει πολύ μεγαλύτερη τιμή για τους άλλους ηχητικούς τύπους. Η δεύτερη πτυχή είναι το σχήμα της καμπύλης του ρυθμού διέλευσης του μηδενός. Για τη φωνή, η καμπύλη αυτή έχει μια σταθερή και χαμηλή γραμμή βάσης με κορυφές άνωθέν της. Ορίζουμε ως γραμμή βάσης τη συνδετική γραμμή των χαμηλότερων σημείων των κοιλάδων της καμπύλης του ρυθμού διέλευσης του μηδενός. Έπειτα, υπολογίζεται η μέση τιμή και η διακύμανση της γραμμής βάσης. Λαμβάνονται επίσης υπόψη οι παράμετροι των κορυφών (ύψος, πλάτος, και οξύτητα) και η συχνότητα εμφάνισής τους. Η Τρίτη και η τέταρτη πτυχή είναι η διακύμανση και η περιοχή του ύψους της καμπύλης του ρυθμού διέλευσης του μηδενός, αντίστοιχα. Σε αντίθεση με τα μουσικά τμήματα όπου η διακύμανση και η περιοχή του ύψους είναι κάτω από κάποια συγκεκριμένα κατώφλια, ένα τυπικό τμήμα μουσικής έχει διακύμανση και περιοχή ύψους πάνω από κάποια συγκεκριμένα κατώφλια. Η Πέμπτη πτυχή είναι η ιδιότητα της θεμελιώδους συχνότητας. Καθώς η φωνή είναι ένα μίγμα αρμονικών και μη αρμονικών ήχων, έχει ένα αρμονικό ποσοστό μέσα σε μια συγκεκριμένη περιοχή. Υπάρχει επίσης μια σχέση μεταξύ της καμπύλης της θεμελιώδους συχνότητας και της καμπύλης της ενέργειας, καθώς τα αρμονικά μέρη της καμπύλης της θεμελιώδους συχνότητας αντιστοιχούν στις κορυφές της καμπύλης της ενέργειας ενώ τα μηδενικά μέρη της καμπύλης της θεμελιώδους συχνότητας αντιστοιχούν στις κοιλάδες της καμπύλης της ενέργειας. Για κάθε μία από τις πέντε πτυχές, καθορίζεται μια κλασματική τιμή απόφασης μεταξύ 0 και 1. Η σταθμισμένη μέση τιμή (weighted average) αυτών των τιμών απόφασης αντιπροσωπεύει την πιθανότητα του τμήματος να είναι φωνή. Όταν η πιθανότητα είναι πάνω από κάποιο συγκεκριμένο κατώφλι και επίσης τουλάχιστον τρεις από τις τιμές απόφασης είναι πάνω από 0.5, το τμήμα ταξινομείται ως «καθαρή μουσική».

***Σημείωση**: Λόγω της ελλιπούς και αόριστης περιγραφής της ανωτέρω διαδικασίας καθορισμού εμπειρικών κατωφλίων και βαρών εξαγωγής μέσων όρων ήταν αδύνατον να αναπαράξουμε αυτό το τμήμα της πρωτότυπης εργασίας, και γι αυτό διακρίναμε την ύπαρξη φωνής μόνο από το εσωτερικό γινόμενο του ρυθμού διέλευσης του μηδενός και της ενέργειας. Συγκεκριμένα, αν το γινόμενο είναι **κάτω από 0.05** το τμήμα χαρακτηρίζεται ως «φωνή». Αυτή η τιμή κατωφλίου βρέθηκε εμπειρικά και κατόπιν εκτεταμένων δοκιμών.*

6.2.8. Ταξινόμηση των μη αρμονικών περιβαλλοντικών ήχων

Το τελευταίο βήμα είναι η ταξινόμηση των ήχων που απέμειναν στη δεύτερη κατηγορία ως τύπων μη αρμονικών περιβαλλοντικών ήχων. Η ταξινόμηση αυτή γίνεται όπως παρακάτω:

□.2.□.1. Αν οι γειτονικές κορυφές της καμπύλης της ενέργειας ή της καμπύλης του ρυθμού διέλευσης του μηδενός ισαπέχουν προσεγγιστικά, το τμήμα ταξινομείται ως ***«περιοδικός ή ψευδοπεριοδικός ήχος»***. Παραδείγματα αυτού του τύπου είναι ο κτύπος του ρολογιού και ο ήχος των βημάτων.

□.2.□.2. Αν το ποσοστό των αρμονικών τμημάτων στην καμπύλη της θεμελιώδους συχνότητας είναι εντός μιας συγκεκριμένης περιοχής (κάτω από το κατώφλι για τη μουσική αλλά πάνω από το κατώφλι για το μη αρμονικό ήχο), το τμήμα ταξινομείται ως ***«μίγμα αρμονικού και μη αρμονικού ήχου»***. Για παράδειγμα, ο ήχος της κόρνας του τρένου που είναι αρμονικός εμφανίζεται να έχει ένα μη αρμονικό υπόβαθρο. Επίσης, ο ήχος του βήχα αποτελείται και από αρμονικές και από μη αρμονικές συνιστώσες.

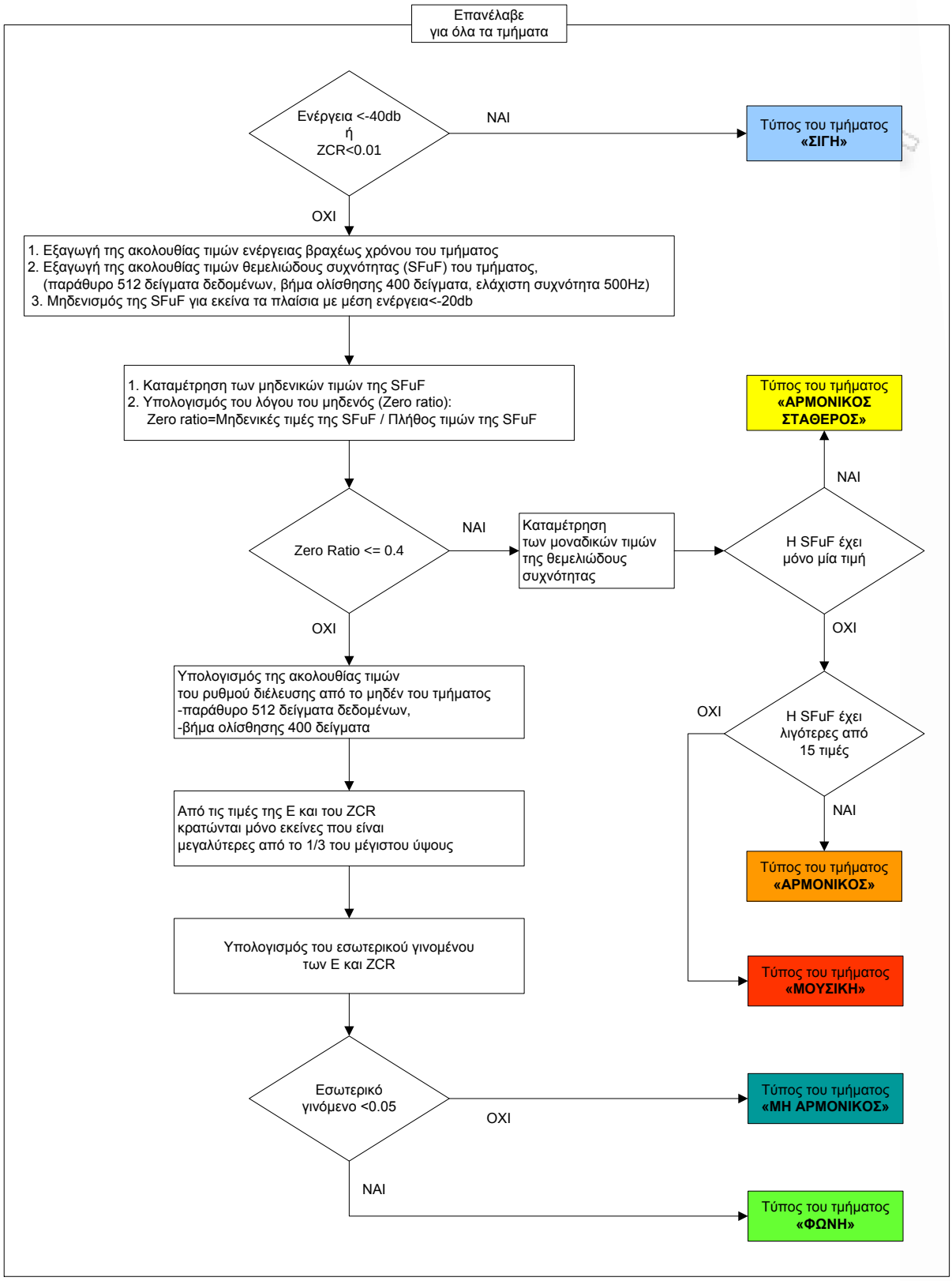
□.2.□.3. Αν οι τιμές του ρυθμού διέλευσης του μηδενός είναι μέσα σε μια σχετικά μικρή περιοχή συγκριτικά με την απόλυτη περιοχή της κατανομής της συχνότητας, το τμήμα ταξινομείται ως ***«μη αρμονικός και σταθερός ήχος».*** Ένα παράδειγμα είναι ο ήχος της κραυγής των πουλιών, που είναι μη αρμονικός αλλά η καμπύλη του ρυθμού διέλευσης του μηδενός είναι συγκεντρωμένη στην περιοχή του 80-120 με μια απόλυτη περιοχή του 150.

□.2.□.4. Τέλος, αν το τμήμα δεν ικανοποιεί καμία από τις παραπάνω συνθήκες, ταξινομείται ως ***«μη αρμονικός και ακανόνιστος ήχος».*** Οι περισσότεροι περιβαλλοντικοί ήχοι είναι αυτού του τύπου, όπως ο ήχος της βροντής, του σεισμού, κα της φωτιάς.

<u>Σημείωση</u>*: Λόγω της αόριστης περιγραφής των διαφόρων περιοχών τιμών, ήταν αδύνατον να αναπαράξουμε αυτό το τμήμα της πρωτότυπης εργασίας, και γι αυτό ταξινομήσαμε ως «μη αρμονικό ήχο» οποιονδήποτε μη αρμονικό ήχο που δεν είναι φωνή. Εξάλλου, θεωρούμε ότι έχει μικρή αξία σε μια ταξινομητική διαδικασία η τόσο λεπτομερής ανάλυση των περιβαλλοντικών ήχων τη στιγμή μάλιστα που ο καθένας ξεχωριστός και αυτόνομος περιβαλλοντικός ήχος δεν διαρκεί αρκούντως πολύ έτσι ώστε να διατηρηθεί ταξινομημένος μετά από το στάδιο της μεταεπεξεργασίας, κατά το οποίο τα πολύ μικρά τμήματα επαναταξινομούνται και συνενώνονται με μεγαλύτερα.*

6.3. Ο υλοποιηθείς ταξινομητικός αλγόριθμος

Κατόπιν όλων των παραπάνω σημειώσεων, κάθε τμήμα που προκύπτει από το στάδιο της κατάτμησης ταξινομείται ως ένας από έξι συνολικά τύπους ήχου: **«Σιγή», «Αρμονικός & σταθερός», «Αρμονικός», «Μουσική», «Μη αρμονικός», και «Φωνή».** Η ταξινόμηση γίνεται σύμφωνα με τον αλγόριθμο που φαίνεται στο Σχήμα 30 και περιγράφεται παρακάτω:



Σχήμα 30: Ο αλγόριθμος ταξινόμησης

.3.1. Πρώτα διαχωρίζονται τα σιωπηλά τμήματα με κριτήριο αν η μέση τιμή της ενέργειας ή η μέση τιμή του ρυθμού διέλευσης του μηδενός είναι κάτω από ένα κατώφλι. Τότε, ο ήχος του τμήματος χαρακτηρίζεται ως **«Σιγή»** υπό την έννοια ότι ή δεν υπάρχει καθόλου ήχος ή η έντασή του είναι τόσο χαμηλή που αυτός δεν γίνεται αντιληπτός ή είναι απλά θόρυβος χαμηλής συχνότητας. **Τα κατώφλια αυτά βρέθηκε πειραματικά να είναι για μεν την ενέργεια: Eave<-40db για δε το ρυθμό διέλευσης του μηδενός: ZCR<0.01.**

.3.2. Για κάθε ένα από τα υπόλοιπα τμήματα εξάγεται η ακολουθία τιμών της ενέργειας βραχέως χρόνου και η ακολουθία τιμών της θεμελιώδους συχνότητας SFuF, με παράθυρο 512 δειγμάτων δεδομένων που ολισθαίνει κατά 400 δείγματα. Για τον υπολογισμό της SFuF επιλέχθηκε ένας FFT 512 σημείων και τάξη αυτοπαλίνδρομου φίλτρο 40. Μετά την εξαγωγή της ακολουθίας της SFuF μηδενίζονται εκείνες οι τιμές της που αντιστοιχούν σε πλαίσια που έχουν μέση ενέργεια κάτω από ένα κατώφλι. **Το κατώφλι βρέθηκε πειραματικά να είναι: E<-20db**.

.3.3. Αν δεν προκύψει ακολουθία τιμών SFuF τότε τίθεται ως τιμή της SFuF το μηδέν, αλλιώς καταμετράται ο αριθμός των μηδενικών τιμών της SFuF κάθε τμήματος και υπολογίζεται ο λόγος του μηδενός (Zero ratio): πλήθος των μηδενικών τιμών της SFuF προς το συνολικό πλήθος των τιμών της.

.3.4. Στη συνέχεια διαχωρίζονται τα τμήματα που περιέχουν **αρμονικούς ήχους** από αυτά που περιέχουν **μη αρμονικούς** αφού τεθεί ένα κατώφλι απόφασης για το λόγο του μηδενός (zero ratio). **Το κατώφλι αυτό βρέθηκε πειραματικά να έχει τιμή 0.4.**

3.4.1. Αν ο λόγος του μηδενός του τμήματος είναι <=0.4 τότε ο ήχος χαρακτηρίζεται ότι ανήκει στην ομάδα των αρμονικών ήχων και από την εξαγχθείσα ακολουθία τιμών της SFuF απαριθμείται το πλήθος των μοναδικών διαφορετικών τιμών της θεμελιώδους συχνότητας. Στη συνέχεια:

3.4.1.1. Αν το πλήθος αυτό βρεθεί να ίσο με ένα, που σημαίνει ότι η θεμελιώδης συχνότητα παραμένει σταθερά σε μία τιμή, τότε ο ήχος του τμήματος χαρακτηρίζεται ως **«Αρμονικός και σταθερός»**.

3.4.1.2. Αλλιώς, αν βρεθεί να έχει τιμή μικρότερη από δεκαπέντε, που σημαίνει ότι η θεμελιώδης συχνότητα παίρνει δεκατέσσερις το πολύ τιμές, τότε ο ήχος του τμήματος χαρακτηρίζεται ως «**Αρμονικός**».

3.4.1.3. Διαφορετικά, δηλαδή αν η θεμελιώδης συχνότητα παίρνει περισσότερες από δεκαπέντε τιμές, ο ήχος του τμήματος χαρακτηρίζεται ως **«Μουσική».**

3.4.2. Αν ο λόγος του μηδενός του τμήματος είναι >0.4, ο ήχος χαρακτηρίζεται ότι ανήκει στην ομάδα των μη αρμονικών ήχων και τότε υπολογίζεται ακόμη ο ρυθμός διέλευσης του μηδενός του τμήματος. Στη συνέχεια επιχειρείται ο διαχωρισμός των μη αρμονικών ήχων σε «φωνή» και σε μη δυνάμενο να ταξινομηθεί διαφορετικά «μη αρμονικό ήχο». Ο διαχωρισμός γίνεται παρατηρώντας ότι για το σήμα της φωνής η καμπύλη της ενέργειας παρουσιάζει κορυφές στα έμφωνα μέρη της και κοιλάδες στα άφωνα, ενώ η καμπύλη του ZCR παρουσιάζει την ακριβώς αντίθετη εικόνα. Έτσι, αν υπολογίσουμε το εσωτερικό γινόμενο μόνο των κορυφών και των κοιλάδων των δύο καμπυλών, αυτό αναμένεται να έχει για το σήμα της φωνής τιμή πολύ κοντά στο μηδέν. Για το σκοπό αυτό ενεργούμε ως εξής:

3.4.2.1. Πρώτα ψαλιδίζουμε την καμπύλη της ενέργειας και την καμπύλη του ZCR στο 1/3 της μέγιστης τιμής τους και απορρίπτουμε τα χαμηλά τμήματα κρατώντας μόνο τις κορυφές τους. Απορρίπτουμε έτσι τις πολύ μικρές τιμές των δύο ακολουθιών που είναι επιμολυσμένες με θόρυβο.

3.4.2.2. Στη συνέχεια, υπολογίζουμε το εσωτερικό γινόμενο των δύο καμπυλών και αν αυτό βρεθεί μικρότερο από ένα κατώφλι, ο ήχος χαρακτηρίζεται ως **«Φωνή»** αλλιώς χαρακτηρίζεται ως **«Μη αρμονικός»**. **Το κατώφλι αυτό βρέθηκε πειραματικά να έχει τιμή 0.05**

6.4. Μεταεπεξεργασία (postprocessing)

Στο στάδιο της μεταεπεξεργασίας μειώνονται τα πιθανά σφάλματα της κατάτμησης και της ταξινόμησης. Έχουμε ρυθμίσει τον αλγόριθμο της κατάτμησης να είναι αρκετά ευαίσθητος έτσι ώστε να ανιχνεύει όλες τις απότομες μεταβολές. Έτσι όμως, είναι πιθανόν μια συνεχόμενη σκηνή να διασπαστεί σε αρκετά τμήματα. Για παράδειγμα, ένα μουσικό κομμάτι μπορεί να διασπαστεί σε αρκετά τμήματα λόγω των απότομων μεταβολών της καμπύλης της ενέργειας, και μερικά μικρά τμήματα να ταξινομηθούν ακόμη και εσφαλμένα ως «αρμονικός και σταθερός περιβαλλοντικός ήχος» λόγω του ότι ο τόνος της μουσικής παραμένει σταθερός στη διάρκεια αυτού του τμήματος. Μέσω της μεταεπεξεργασίας, αυτά τα τμήματα συνενώνονται με άλλα τμήματα και επαναταξινομούνται βάσει των συναφών σχέσεών τους. Ακολουθούν μερικά παραδείγματα ευριστικών κανόνων που χρησιμοποιούνται σε αυτό το στάδιο:

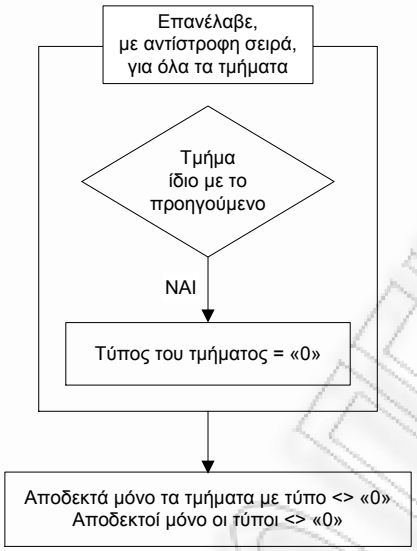
- Αν ένα τμήμα «σιγής» είναι βραχύτερο των 2sec και τα δύο τμήματα πριν και μετά από αυτό είναι του ιδίου τύπου, τότε τα τρία τμήματα συνενώνονται σε ένα και ταξινομούνται όπως το πρώτο τμήμα.
- Αν ένα «αρμονικό χωρίς αλλαγές» ή «σταθερά αρμονικό» περιβαλλοντικό ηχητικό τμήμα είναι βραχύτερο των 5sec και έπεται ενός τμήματος «καθαρής μουσικής» ή «τραγουδιού», τότε αυτό συνενώνεται σε ένα τμήμα μαζί με το προηγούμενο.
- Αν ένας «μικτός αρμονικός και μη αρμονικός» περιβαλλοντικός ήχος είναι βραχύτερος των 2sec και βρίσκεται μεταξύ δύο τμημάτων «φωνής με μουσική υπόκρουση» ή «περιβαλλοντικού ήχου με μουσική υπόκρουση», τα τρία τμήματα συνενώνονται σε ένα και ταξινομούνται σύμφωνα με το πρώτο τμήμα.

Σημείωση: *Λόγω του ότι δεν υλοποιήσαμε όλες τις κατηγορίες της πρωτότυπης εργασίας, για τους λόγους που έχουμε τμηματικά αναφέρει μέχρι τώρα, και λόγω του ότι ο αλγόριθμος μεταεπεξεργασίας δεν περιγράφεται αναλυτικά παρά δίνονται μόνο ορισμένοι ευριστικοί κανόνες που εφαρμόζονται σε αυτόν, υλοποιήσαμε τον ακόλουθο αλγόριθμο μεταεπεξεργασίας ο οποίος περιλαμβάνει τα εξής τρία στάδια:*

1. *Τη **συνένωση** (linking) όπου τα όμοια τμήματα συνενώνονται σε μεγαλύτερα,*
2. *Την **επαναταξινόμηση** (reindexing) όπου τα τμήματα διάρκειας μικρότερης από περίπου 3sec επαναταξινομούνται στον τύπο του προηγούμενου τμήματος και*
3. *Την **τελική συνένωση** (final linking) όπου τα όμοια τμήματα συνενώνονται για να σχηματίσουν τα τελικά τμήματα.*

Στα σχήματα 31, 32 και 33 φαίνονται τα διαγράμματα ροής των τριών σταδίων και δίνεται μία σύντομη περιγραφή των αντιστοίχων λειτουργιών.

Στάδιο 1: Συνένωση

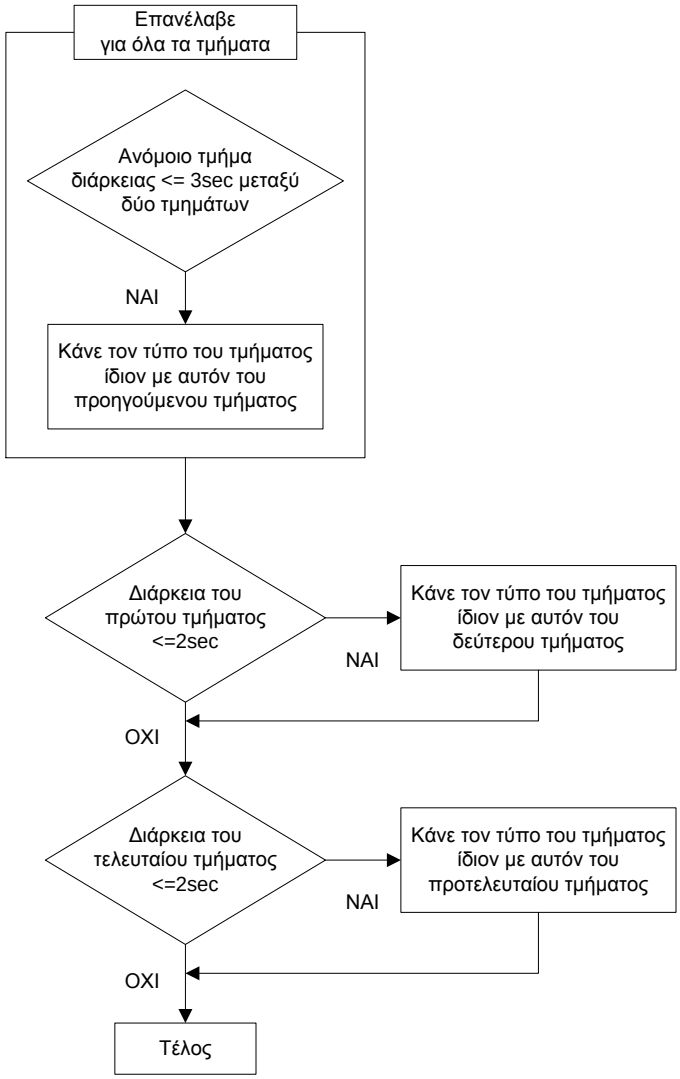


Σχήμα 31: Συνένωση ομοίων διαδοχικών τμημάτων

1. *Εξετάζουμε τα τμήματα με αντίστροφη σειρά, δηλαδή από το τελευταίο στο πρώτο και αν το τμήμα είναι ίδιο με το προηγούμενό του μετατρέπεται σε τύπο «0».*
2. *Κρατούνται μόνο τα τμήματα με τύπο διαφορετικό από «0».*
3. *Κρατούνται μόνο οι τύποι που είναι διαφορετικοί από «0».*

Με αυτόν τον τρόπο κρατούνται μόνο οι αρχές των σειρών των ομοίων τμημάτων

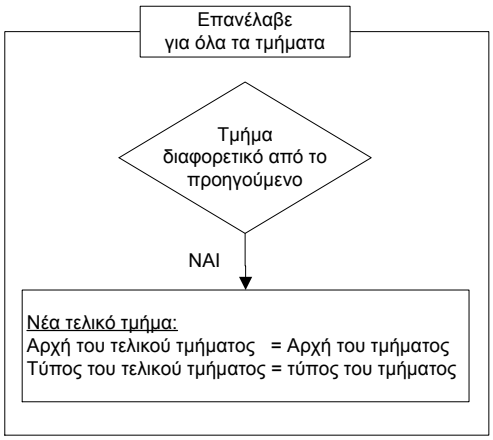
Στάδιο 2: Επαναταξινόμηση.



Σχήμα 32: Επαναταξινόμηση και συνένωση μικρών τμημάτων με μεγαλύτερα

1. *Εξετάζεται **αν το τμήμα έχει διάρκεια μικρότερη από 3sec**. Αυτό το χρονικό διάστημα επιλέχθηκε διότι τα τμήματα μικρότερης διάρκειας, ακόμη και αν έχουν ταξινομηθεί ορθά, δεν έχουν κανένα πρακτικό ενδιαφέρον. Ίσως μάλιστα να μην μας ενδιαφέρουν και τμήματα ακόμη μεγαλύτερης διάρκειας.*
2. *Αν το τμήμα είναι ανόμοιο με το προηγούμενο και το επόμενό του, τότε επαναταξινομείται σε τύπο ίδιο με τον τύπο του προηγουμένου του τμήματος.*
3. *Εξετάζεται αν το πρώτο τμήμα έχει διάρκεια μικρότερη από 2sec. Αν ναι, μετατρέπεται στον τύπο του δεύτερου τμήματος.*
4. *Εξετάζεται αν το τελευταίο τμήμα έχει διάρκεια μικρότερη από 2sec. Αν ναι, μετατρέπεται στον τύπο του προτελευταίου τμήματος.*

Στάδιο 3: Τελική συνένωση.



Σχήμα 33: Τελική συνένωση ομοίων τμημάτων

Σε αυτό το βήμα τα όμοια τμήματα συνενώνονται σε μεγαλύτερα και προκύπτουν τα τελικά τμήματα.

7. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

7.1. Η ηχητική βάση δεδομένων

Η επίδοση του συστήματος δεν είναι δυνατόν να μετρηθεί αντικειμενικά και με απόλυτη ακρίβεια λόγω της υποκειμενικής επιλογής των αρχείων ελέγχου βάσει των οποίων θα γίνει η αξιολόγηση του συστήματος. Στα πλαίσια αυτής της εργασίας, και για το σκοπό της ρύθμισης των παραμέτρων και της μέτρησης της απόδοσης του συστήματος, δημιουργήθηκε μια ηχητική βάση δεδομένων αποτελούμενη από έξι ομάδες αρχείων ήχου:

1. Αρχεία μουσικής που καλύπτουν τα εξής είδη:
   i. Κλασσική μουσική (16 αρχεία)
   ii. Μουσική Rock (15 αρχεία)
   iii. Μουσική Jazz (11 αρχεία)
   iv. Ελληνική Λαϊκή μουσική (12 αρχεία)
   v. Ελληνική Δημοτική μουσική (10 αρχεία)
   vi. Ελληνική μουσική (5 αρχεία)
2. Κατασκευασμένα αρχεία ρυθμίσεων και ελέγχου (73 αρχεία)
3. Ηχογραφήσεις ραδιοφωνικών και τηλεοπτικών εκπομπών (16 αρχεία)
4. Αρχεία ομιλιών σε διάφορες γλώσσες (8 αρχεία)
5. Δείγματα ήχων μουσικών οργάνων (8 αρχεία)
6. Ηχητικά εφέ (7 αρχεία)

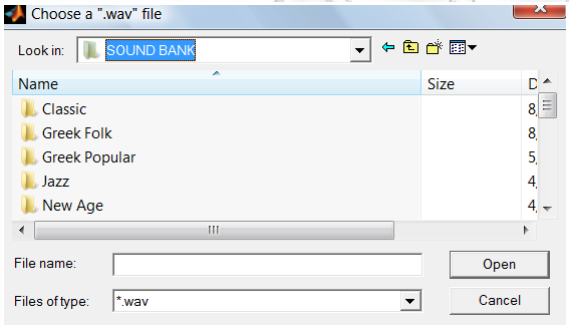
Έγινε μεγάλη προσπάθεια ώστε τα παραπάνω αρχεία αφενός μεν να είναι αντιπροσωπευτικά της κατηγορίας τους, αφετέρου δε να παρουσιάζουν μεταξύ τους όσο το δυνατόν μικρότερη ομοιομορφία ώστε το συνολικό δείγμα να είναι γενικό και αντιπροσωπευτικό.

7.2. Η γραφική διεπαφή

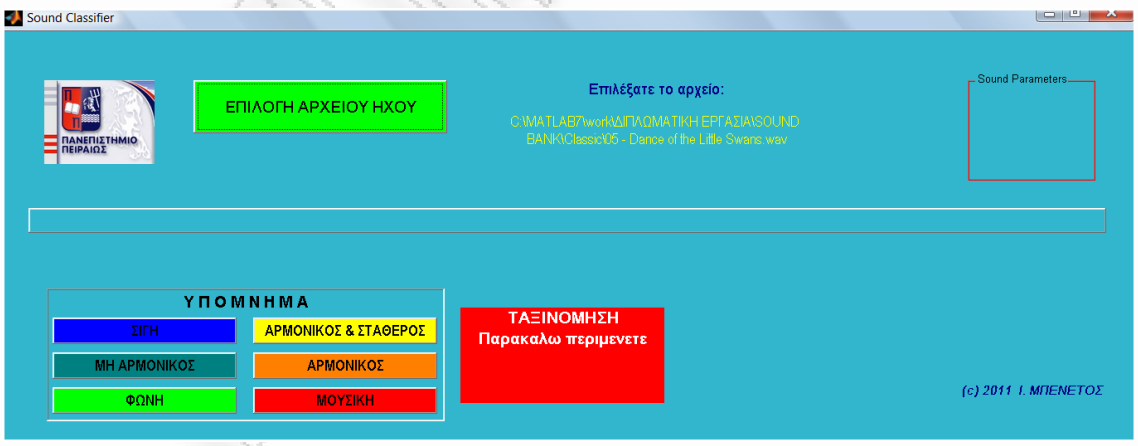
Κατασκευάστηκε η γραφική διεπαφή του συστήματος που διευκολύνει την επιλογή του προς εξέταση αρχείου και την ανάγνωση των αποτελεσμάτων. Η γραφική διεπαφή φαίνεται στην Εικόνα 1.

Εικόνα 1: Η γραφική διεπαφή πριν την ταξινόμηση

Πατώντας το πλήκτρο «ΕΠΙΛΟΓΗ ΑΡΧΕΙΟΥ ΗΧΟΥ» ανοίγει ένα παράθυρο με το οποίο επιλέγουμε το υπό εξέταση αρχείο:



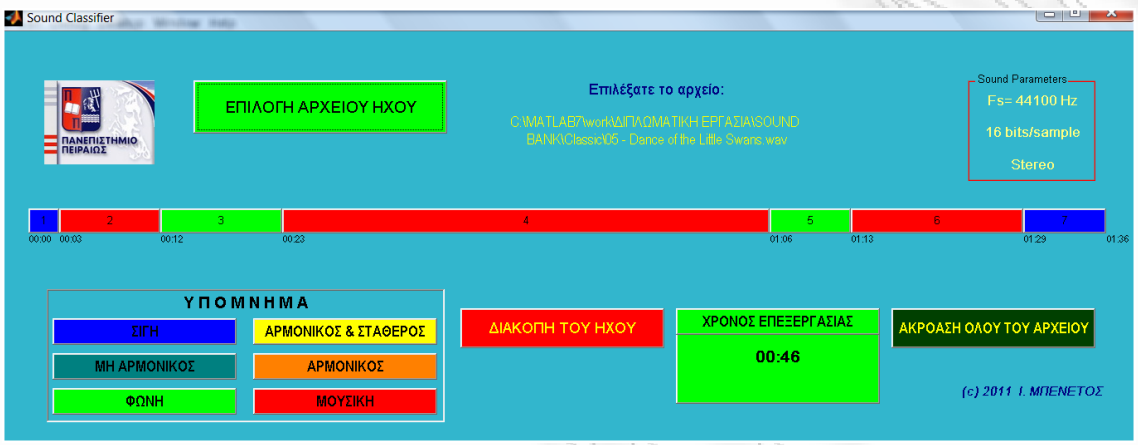
Στη συνέχεια αναμένουμε μέχρι να ολοκληρωθεί η ταξινόμηση:



Εικόνα 2: Η γραφική διεπαφή κατά τη διάρκεια της ταξινόμησης

Τελικά εμφανίζεται το αποτέλεσμα της ταξινόμησης ως μια έγχρωμη λωρίδα στην οποία τα ταξινομηθέντα τμήματα εμφαίνονται με τα χρώματα του υπομνήματος.

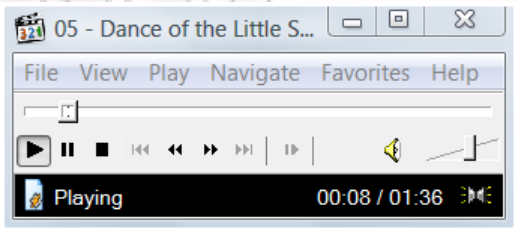
Εμφανίζονται επίσης οι παράμετροι του αρχείου, δηλαδή ο ρυθμός δειγματοληψίας σε Hz (πριν την επαναδειγματολήπτηση στα 11025Hz), η ανάλυση των δειγμάτων σε bits/sample, και ο αριθμός των καναλιών (Mono – Stereo). Επίσης, εμφανίζεται και ο συνολικός χρόνος επεξεργασίας του αρχείου.



Εικόνα 3: Η γραφική διεπαφή στο τέλος της ταξινόμησης

Κάνοντας κλικ πάνω σε οποιοδήποτε τμήμα ακούμε τον ήχο αυτού του τμήματος, ενώ πατώντας το πλήκτρο «ΔΙΑΚΟΠΗ ΤΟΥ ΗΧΟΥ» διακόπτουμε την αναπαραγωγή του ηχητικού τμήματος.

Επίσης, πατώντας το πλήκτρο «ΑΚΡΟΑΣΗ ΟΛΟΥ ΤΟΥ ΑΡΧΕΙΟΥ» καλείται και εκτελείται μια εξωτερική εφαρμογή, συγκεκριμένα ο "Media Player Classic, με την οποία μπορούμε να περιηγηθούμε σε ολόκληρο το αρχείο ήχου:



Εικόνα 4: Η εξωτερική εφαρμογή "Media Player Classic"

7.3. Ρυθμίσεις και έλεγχοι

Ο καθορισμός των κατωφλίων που χρησιμοποιήθηκαν στις ευριστικές διαδικασίες του συστήματος έγινε με τη βοήθεια 73 τεχνητών αρχείων ρυθμίσεων και ελέγχου που δημιουργήθηκαν από άλλα απλούστερα αρχεία. Τα αρχικά απλά αρχεία περιείχαν δείγματα όλων των υπό εξέταση ηχητικών τύπων καθώς επίσης και τεχνητά

κατασκευασμένα απλά ημιτονικά, σύνθετα ημιτονικά και τετραγωνικά σήματα διαφόρων συχνοτήτων. Τα αρχεία ρυθμίσεων και ελέγχου κατασκευάστηκαν με παράθεση ή με μίξη, και σε διάφορους συνδυασμούς, τμημάτων των αρχικών απλών αρχείων. Σκοπός της κατασκευής των εν λόγω αρχείων ήταν η δημιουργία μιας ποικιλίας πρότυπων ηχητικών αλλαγών βάσει των οποίων θα γινόταν η ρύθμιση των κατωφλίων του συστήματος.

Οι τιμές των κατωφλίων καθορίστηκαν βήμα-βήμα, σύμφωνα με τη διαδικασία κατάτμησης και ταξινόμησης που περιγράφεται στην ενότητα 6. Σε κάθε βήμα εκτελούταν μια επαναληπτική διαδικασία αλλαγών και ελέγχου του αποτελέσματος έως ότου επιτυγχανόταν ένα βέλτιστο αποτέλεσμα. Στη συνέχεια χρησιμοποιήθηκε όλο το σύνολο των ηχητικών δεδομένων για τις τελευταίες μικρορυθμίσεις και την τελική αξιολόγηση του συστήματος.

Αρχικά έγινε η επιβεβαίωση της ορθής λειτουργίας του συστήματος με 22 κατασκευασμένα αρχεία ελέγχου. Συγκεκριμένα, δημιουργήθηκαν δύο ομάδες αρχείων που περιείχαν:

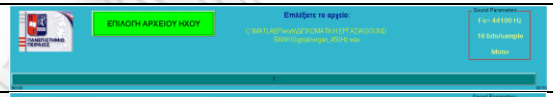
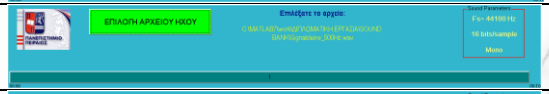
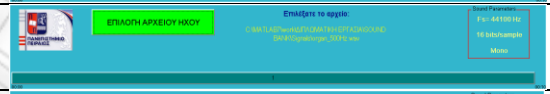
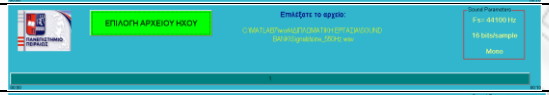
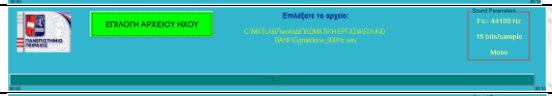
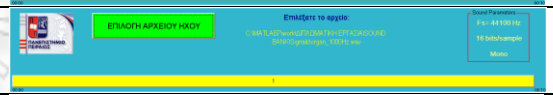
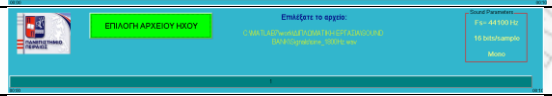
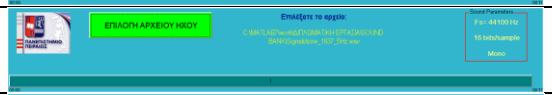
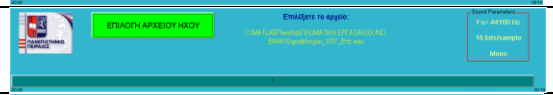
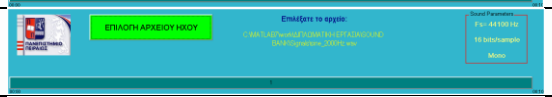
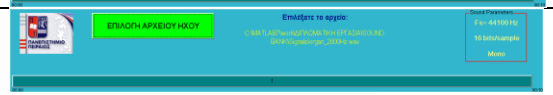
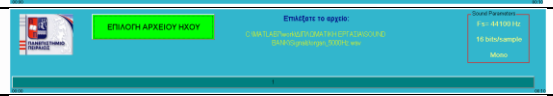
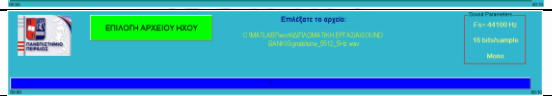
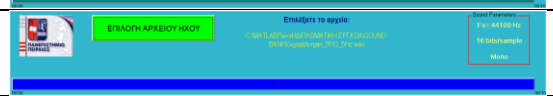
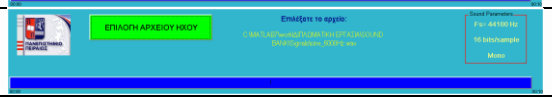
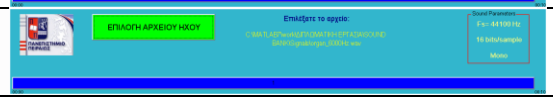
- Απλά ημιτονικά σήματα
- Σύνθετα ημιτονικά σήματα αποτελούμενα από μια θεμελιώδη συχνότητα f1 συνοδευόμενη από τις τρεις πρώτες αρμονικές της: f2=2*f1, f3=3*f1 και f4=4*f1 με πλάτη f1: -8db, f2: -10db, f3: -10db, f4: -14db.

Κάθε ομάδα περιλάμβανε έντεκα αρχεία σημάτων με θεμελιώδεις συχνότητες: 450, 500, 550, 800, 1000, 1800, 1837.5, 2000, 5000, 5512.5 και 6000Hz.

Τα αποτελέσματα των ταξινομήσεων φαίνονται στον Πίνακα 1 από τον οποίο αφενός τεκμηριώνεται η ορθή λειτουργία του συστήματος και αφετέρου εντοπίζονται μερικές αδυναμίες που θα αποτελέσουν πηγές σφαλμάτων κατά τη λειτουργία του συστήματος με πραγματικά αρχεία. Από αυτόν τον πίνακα εξάγουμε τις εξής παρατηρήσεις:

1. Το απλό ημιτονικό σήμα ταξινομήθηκε σε όλες τις περιπτώσεις ορθά ως «μη αρμονικός» ήχος διότι πράγματι δεν περιέχει αρμονικές συχνότητες.
2. Το σύνθετο ημιτονικό σήμα αποτελείται από τη θεμελιώδη και τις τρεις πρώτες αρμονικές του. Μέχρι και τα 500Hz ταξινομήθηκε ως «μη αρμονικός» ήχος διότι ο αλγόριθμος ταξινόμησης απορρίπτει τις συχνότητες κάτω των 500Hz. Σε αυτή την περίπτωση, ο αλγόριθμος προσπάθησε να εκτιμήσει ως θεμελιώδη τη δεύτερη αρμονική f2=1000Hz που θα έπρεπε να έχει αρμονικές 2*1000=2000Hz, 3*1000=3000Hz, 4*1000=4000Hz και 5*1000=5000Hz. Το σύνθετο σήμα των 500Hz έχει αρμονικές **2*500=1000Hz**, 3*500=1500Hz και 4*500=2000Hz. Έτσι, δεν μπορούν να ανιχνευθούν τουλάχιστον δύο αρμονικές, όπως απαιτεί ο αλγόριθμος ταξινόμησης, και το σήμα αναγνωρίζεται ορθά ως «μη αρμονικός».

3. Το σύνθετο ημιτονικό σήμα με συχνότητα μεγαλύτερη ή ίση των 500Hz και μικρότερη των 1837,5Hz ταξινομήθηκε ορθά ως «αρμονικός και σταθερός» ήχος διότι η θεμελιώδης συχνότητά του παραμένει σταθερή και συνοδεύεται από δύο τουλάχιστον αρμονικές συχνότητες.
4. Το σύνθετο ημιτονικό σήμα με συχνότητα μεγαλύτερη ή ίση των 1837,5 και μικρότερη των 5512,5Hz ταξινομήθηκε ως «μη αρμονικός» διότι δεν μπόρεσαν να ανιχνευθούν τουλάχιστον δύο αρμονικές του με συχνότητες: 2ης αρμονικής 2*1837,5=3675Hz και άνω, και 3ης αρμονικής 3*1837,5=5512,5Hz και άνω λόγω της απόσβεσής τους από το φίλτρο αντιαλλοίωσης (anti-alias filter)
5. Όλα τα σήματα με θεμελιώδη συχνότητα μεγαλύτερη ή ίση των 5512,5Hz ταξινομήθηκαν ως «σιγή» λόγω της απόσβεσής τους από το φίλτρο αντιαλλοίωσης.

Θεμελιώδης Συχνότητα (Hz)	Απλό ημιτονικό σήμα	Σύνθετο ημιτονικό σήμα
450		
500		
550		
800		
1000		
1800		
1837,5		
2000		
5000		
5512,5		
6000		

Πίνακας 1: Τα αποτελέσματα της ταξινόμησης των αρχείων ελέγχου

7.4. Τα αποτελέσματα της ταξινόμησης

Όλες οι ταξινομήσεις έγιναν με έναν φορητό υπολογιστή Pentium Dual Core T4200 2.00GHz που εκτελούσε το λειτουργικό σύστημα Microsoft Windows Vista. Παρατηρήσαμε ότι οι χρόνοι επεξεργασίας των αρχείων κυμάνθηκαν περίπου στο ήμισυ του πραγματικού χρόνου, και έτσι συμπεραίνουμε ότι το υλοποιηθέν σύστημα είναι ικανό για επεξεργασία σε πραγματικό χρόνο.

Αξιολογήθηκαν συνολικά 559 ηχητικά τμήματα ακολουθώντας την εξής διαδικασία:

1. Μετά το τέλος της ταξινόμησης κάθε αρχείου γινόταν σύλληψη (capture) της εικόνας της γραφικής διεπαφής και αποθηκευόταν ως αρχείο εικόνας.
2. Δημιουργούσαμε ένα αντίγραφο της έγχρωμης λωρίδας πάνω στην οποία γινόντουσαν οι διορθώσεις των ηχητικών τύπων κατά την ακρόαση του αρχείου, όπως στην επόμενη εικόνα

3. Με σύγκριση των δύο έγχρωμων λωρίδων προέκυπταν τα αποτελέσματα της αξιολόγησης του συστήματος.

Τα αποτελέσματα της αξιολόγησης της κάθε ηχητικής ομάδας φαίνονται στους επόμενους πίνακες σύγχυσης (confusion matrices)

ROCK	ΣΙΓΗ	ΜΗ ΑΡΜΟΝΙΚΟΣ	ΦΩΝΗ	ΑΡΜΟΝΙΚΟΣ	ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	ΜΟΥΣΙΚΗ
ΣΙΓΗ	13					
ΜΗ ΑΡΜΟΝΙΚΟΣ						
ΦΩΝΗ			2			
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ						
ΑΡΜΟΝΙΚΟΣ					1	
ΜΟΥΣΙΚΗ	1	2	22		3	24

ΚΛΑΣΣΙΚΗ	ΣΙΓΗ	ΜΗ ΑΡΜΟΝΙΚΟΣ	ΦΩΝΗ	ΑΡΜΟΝΙΚΟΣ	ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	ΜΟΥΣΙΚΗ
ΣΙΓΗ	19					
ΜΗ ΑΡΜΟΝΙΚΟΣ						
ΦΩΝΗ						
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ						
ΑΡΜΟΝΙΚΟΣ					20	
ΜΟΥΣΙΚΗ	2		33			51

ΔΗΜΟΤΙΚΑ	ΣΙΓΗ	ΜΗ ΑΡΜΟΝΙΚΟΣ	ΦΩΝΗ	ΑΡΜΟΝΙΚΟΣ	ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	ΜΟΥΣΙΚΗ
ΣΙΓΗ	10					
ΜΗ ΑΡΜΟΝΙΚΟΣ						
ΦΩΝΗ						
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ						
ΑΡΜΟΝΙΚΟΣ					1	
ΜΟΥΣΙΚΗ						11

JAZZ	ΣΙΓΗ	ΜΗ ΑΡΜΟΝΙΚΟΣ	ΦΩΝΗ	ΑΡΜΟΝΙΚΟΣ	ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	ΜΟΥΣΙΚΗ
ΣΙΓΗ	2					
ΜΗ ΑΡΜΟΝΙΚΟΣ						
ΦΩΝΗ			5			
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ						
ΑΡΜΟΝΙΚΟΣ					16	
ΜΟΥΣΙΚΗ		1	20			23

ΛΑΪΚΑ	ΣΙΓΗ	ΜΗ ΑΡΜΟΝΙΚΟΣ	ΦΩΝΗ	ΑΡΜΟΝΙΚΟΣ	ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	ΜΟΥΣΙΚΗ
ΣΙΓΗ	12					
ΜΗ ΑΡΜΟΝΙΚΟΣ						
ΦΩΝΗ			1			
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ						
ΑΡΜΟΝΙΚΟΣ					3	
ΜΟΥΣΙΚΗ	4		4		1	15

RADIO-TV	ΣΙΓΗ	ΜΗ ΑΡΜΟΝΙΚΟΣ	ΦΩΝΗ	ΑΡΜΟΝΙΚΟΣ	ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	ΜΟΥΣΙΚΗ
ΣΙΓΗ	3					
ΜΗ ΑΡΜΟΝΙΚΟΣ						
ΦΩΝΗ	2		40			6
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ						
ΑΡΜΟΝΙΚΟΣ					5	
ΜΟΥΣΙΚΗ	1	2	33			60

ΔΙΑΦΟΡΑ	ΣΙΓΗ	ΜΗ ΑΡΜΟΝΙΚΟΣ	ΦΩΝΗ	ΑΡΜΟΝΙΚΟΣ	ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	ΜΟΥΣΙΚΗ
ΣΙΓΗ	8					
ΜΗ ΑΡΜΟΝΙΚΟΣ		1				2
ΦΩΝΗ			4			
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ						
ΑΡΜΟΝΙΚΟΣ			1		4	
ΜΟΥΣΙΚΗ			1			6

ΦΩΝΗ	ΣΙΓΗ	ΜΗ ΑΡΜΟΝΙΚΟΣ	ΦΩΝΗ	ΑΡΜΟΝΙΚΟΣ	ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	ΜΟΥΣΙΚΗ
ΣΙΓΗ	11					
ΜΗ ΑΡΜΟΝΙΚΟΣ						3
ΦΩΝΗ		1	24			12
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ						
ΑΡΜΟΝΙΚΟΣ		1			4	
ΜΟΥΣΙΚΗ						4

Τα αποτελέσματα της συνολικής αξιολόγησης καταχωρήθηκαν στον επόμενο πίνακα σύγχυσης (confusion matrix) (Πίνακας 2), από τον οποίο κατασκευάστηκαν οι Πίνακες αποτελεσμάτων 3 και 4.

	ΣΙΓΗ	ΜΗ ΑΡΜΟΝΙΚΟΣ	ΦΩΝΗ	ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	ΑΡΜΟΝΙΚΟΣ	ΜΟΥΣΙΚΗ	ΣΥΝΟΛΟ
ΣΙΓΗ	78						78
ΜΗ ΑΡΜΟΝΙΚΟΣ		1				5	6
ΦΩΝΗ	2		76			18	96
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ							
ΑΡΜΟΝΙΚΟΣ			1		54		55
ΜΟΥΣΙΚΗ	8	5	113		4	194	324
ΣΥΝΟΛΟ	88	6	190		58	217	559

Πίνακας 2: Ο πίνακας σύγχυσης (confusion matrix) των αποτελεσμάτων

		ΟΡΘΑ ΔΕΙΓΜΑΤΑ		ΛΑΘΟΣ ΕΚΤΙΜΗΣΕΙΣ	
ΚΑΤΗΓΟΡΙΑ	**ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ**	**ΑΡΙΘΜΟΣ**	**ΕΥΑΙΣΘΗΣΙΑ (Sensitivity)**	**ΑΡΙΘΜΟΣ**	**ΑΝΑΚΛΗΣΗ (Recall)**
ΣΙΓΗ	78	78	100%	0	89%
ΜΕ ΜΟΥΣΙΚΕΣ ΣΥΝΙΣΤΩΣΕΣ	379	248	65%	131	90%
ΔΙΧΩΣ ΜΟΥΣΙΚΕΣ ΣΥΝΙΣΤΩΣΕΣ	102	77	75%	25	39%

Πίνακας 3: Αποτελέσματα ταξινόμησης για τις ηχητικές κατηγορίες

		ΟΡΘΑ ΔΕΙΓΜΑΤΑ		ΛΑΘΟΣ ΕΚΤΙΜΗΣΕΙΣ	
ΚΑΤΗΓΟΡΙΑ	**ΑΡΙΘΜΟΣ ΔΕΙΓΜΑΤΩΝ**	**ΑΡΙΘΜΟΣ**	**ΕΥΑΙΣΘΗΣΙΑ (Sensitivity)**	**ΑΡΙΘΜΟΣ**	**ΑΝΑΚΛΗΣΗ (Recall)**
ΜΗ ΑΡΜΟΝΙΚΟΣ	6	1	17%	5	17%
ΦΩΝΗ	96	76	79%	20	40%
ΑΡΜΟΝΙΚΟΣ & ΣΤΑΘΕΡΟΣ	0	0		0	
ΑΡΜΟΝΙΚΟΣ	55	54	98%	1	93%
ΜΟΥΣΙΚΗ	324	194	60%	130	89%

Πίνακας 4: Αποτελέσματα ταξινόμησης για τους βασικούς ηχητικούς τύπους

Στους Πίνακες 3 και 4 φαίνονται τα αποτελέσματα για τις δύο ιεραρχίες ταξινόμησης, δηλαδή τις ηχητικές κατηγορίες και τους βασικούς ηχητικούς τύπους αντίστοιχα, όπου ως «ευαισθησία» (Sensitivity) ορίζεται ο λόγος του αριθμού των ορθά ταξινομηθέντων δειγμάτων της κατηγορίας προς τον πραγματικό αριθμό των δειγμάτων της κατηγορίας, και ως «ανάκληση» (Recall) ορίζεται ο λόγος των ορθά ταξινομηθέντων δειγμάτων της κατηγορίας προς το συνολικό αριθμό των ταξινομηθέντων δειγμάτων της κατηγορίας.

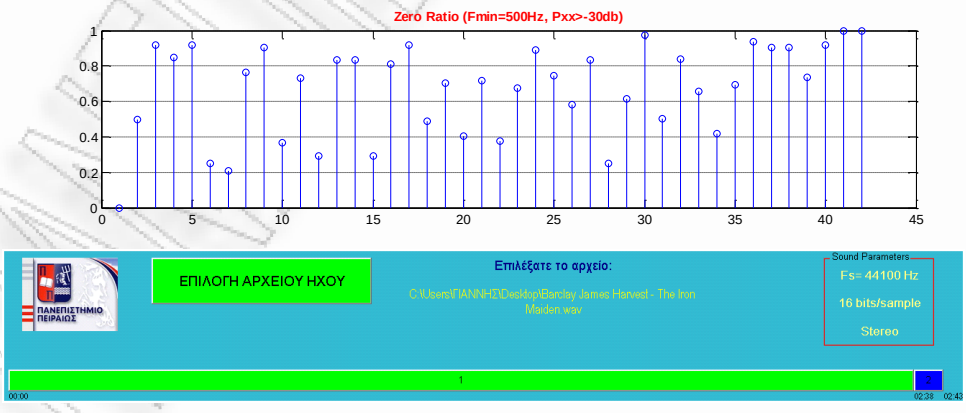
Από τον Πίνακα 3 παρατηρούμε τα εξής:

1. Τα τμήματα της σιγής αναγνωρίζονται με μεγάλη επιτυχία καθώς είναι εύκολη η ανίχνευσή τους.
2. Τα τμήματα με μουσικές συνιστώσες παρουσιάζουν υψηλή ανάκληση που σημαίνει ότι αν κάποιο τμήμα ταξινομηθεί ως έχον μουσικές συνιστώσες είναι πολύ πιθανόν να περιέχει όντως μουσικές συνιστώσες. Επίσης, αυτά τα τμήματα παρουσιάζουν μέτρια ευαισθησία που σημαίνει ότι αρκετά τμήματα που περιέχουν μουσικές συνιστώσες δεν ταξινομούνται ορθά.
3. Τα τμήματα δίχως μουσικές συνιστώσες παρουσιάζουν πολύ χαμηλή ανάκληση που σημαίνει ότι αν κάποιο τμήμα ταξινομηθεί ως μη έχον μουσικές συνιστώσες είναι πολύ πιθανόν στην πραγματικότητα να περιέχει μουσικές συνιστώσες. Επίσης, αυτά τα τμήματα παρουσιάζουν αρκετά καλή ευαισθησία που σημαίνει ότι τα περισσότερα τμήματα που περιέχουν μουσικές συνιστώσες ταξινομούνται ορθά.

Τα σφάλματα ταξινόμησης στις δύο βασικές κατηγορίες, δηλαδή σε ήχους με ή δίχως μουσικό περιεχόμενο, οφείλονται στο απλό κριτήριο του λόγου του μηδενός, σύμφωνα με το οποίο ένας ήχος με χαμηλό λόγο μηδενός περιέχει μουσικές συνιστώσες ενώ ένας με υψηλό λόγο δεν περιέχει μουσικές συνιστώσες. Αυτή η πρώτη αδρή ταξινόμηση οδηγεί ορισμένους ήχους σε λάθος ταξινομητικό μονοπάτι και έτσι αυτοί χαρακτηρίζονται εντελώς αντίθετα απ΄ ότι θα έπρεπε. Οι ανεπιτυχείς ταξινομήσεις οφείλονται στους εξής λόγους:

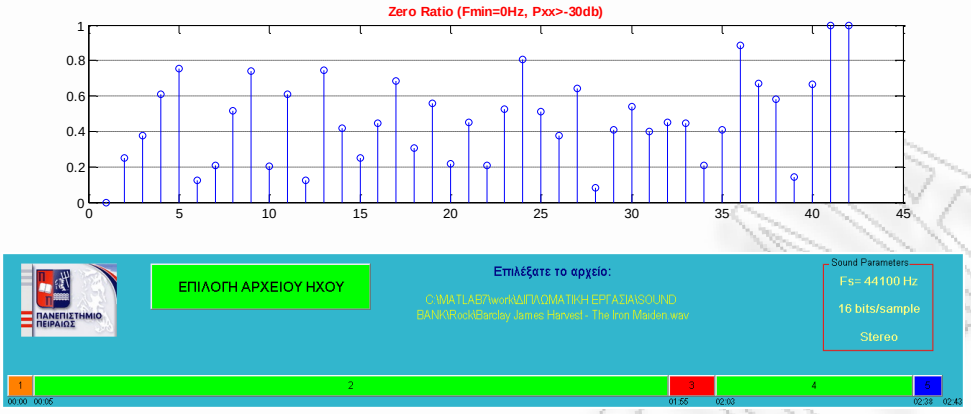
1. Η τιμές των κατωφλίων και κυρίως του λόγου του μηδενός τίθενται εν πολλοίς αυθαίρετα διότι βασίζονται στις παρατηρήσεις που έγιναν με τα αρχεία αναφοράς.
2. Όταν η ένταση του μουσικού περιεχομένου είναι πολύ χαμηλή σε σχέση με το φωνητικό περιεχόμενο.
3. Όταν στη μουσική κυριαρχούν τα κρουστά ή τα νυκτά έγχορδα τα οποία παρουσιάζουν υψηλούς λόγους μηδενός, παρόμοιους με αυτούς της φωνής. Σε αυτή την περίπτωση η μουσική ταξινομείται εσφαλμένα ως φωνή.
4. Η θεμελιώδης συχνότητα, πάνω στην οποία στηρίχτηκε ο υπολογισμός του λόγου του μηδενός, εξήχθη λαμβάνοντας υπόψη μόνο τις φασματικές κορυφές με συχνότητα άνω των 500Hz για να μην υπάρξει επηρεασμός από το συνυφασμένο σήμα της φωνής. Επίσης, δεν λαμβάνονται υπόψη οι φασματικές κορυφές με χαμηλό ύψος. Οι δύο αυτοί περιορισμοί οδηγούν σε εσφαλμένη εκτίμηση της θεμελιώδους συχνότητας, και συνακόλουθα του λόγου του μηδενός, στην περίπτωση που ένα μπάσο μουσικό όργανο (π.χ. ένα κοντραμπάσο) είτε εκτελώντας ένα σόλο ή κυριαρχώντας έναντι των υπολοίπων μουσικών οργάνων, παράγει νότες με θεμελιώδεις συχνότητες κάτω από 500Hz. Σε αυτή την περίπτωση η μουσική ταξινομείται εσφαλμένα ως φωνή.

   Στην Εικόνα 5 φαίνεται η περίπτωση ενός μουσικού κομματιού στο οποίο κυριαρχεί το ηλεκτρικό μπάσο έναντι των υπολοίπων μουσικών οργάνων. Στο παράδειγμα αυτό, δεν ανιχνεύεται η θεμελιώδης συχνότητα λόγω της ύπαρξης του κατωφλίου Fmin=500Hz, ο λόγος του μηδενός λαμβάνει υψηλές τιμές και ολόκληρο το κομμάτι ταξινομείται εσφαλμένα ως «φωνή».



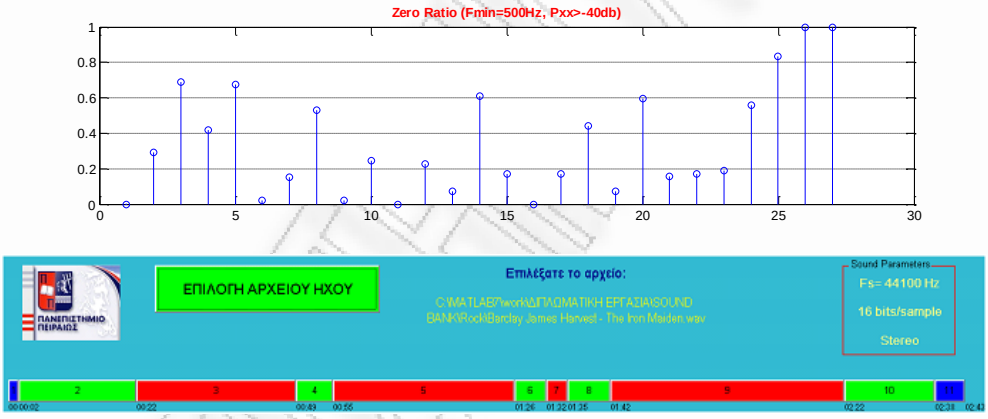
Εικόνα 5: Ανίχνευση θεμελιωδών με Fmin>500Hz και κορυφές >-30db

Στην Εικόνα 6 φαίνονται τα αποτελέσματα για το ίδιο αρχείο αλλά δίχως να έχει τεθεί κατώφλι ελάχιστης θεμελιώδους συχνότητας.



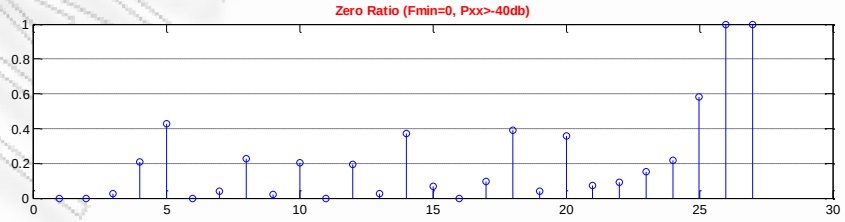
Εικόνα 6: Ανίχνευση όλων των θεμελιωδών συχνοτήτων με κορυφές > -30db

Στην Εικόνα 7 φαίνονται τα αποτελέσματα για το ίδιο αρχείο έχοντας θέσει κατώφλι ελάχιστης θεμελιώδους συχνότητας Fmin=500Hz και ανιχνεύοντας φασματικές κορυφές με ύψος -40db.



Εικόνα 7: Ανίχνευση θεμελιωδών με Fmin>500Hz και κορυφές > -40db

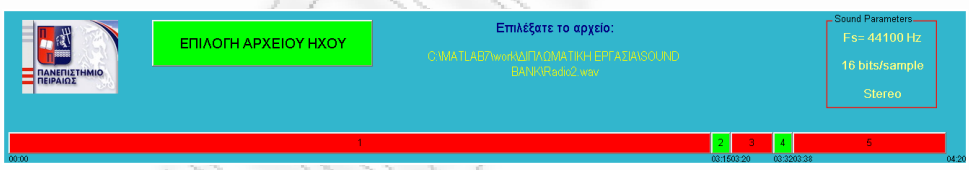
Και τέλος, στην Εικόνα 8 φαίνονται τα αποτελέσματα για το ίδιο αρχείο δίχως κατώφλι ελάχιστης θεμελιώδους συχνότητας και ανιχνεύοντας φασματικές κορυφές με ύψος -40db.



Εικόνα 8: Ανίχνευση όλων των θεμελιωδών συχνοτήτων με κορυφές > -40db

Ύστερα από τις παραπάνω παρατηρήσεις θα μπορούσαμε να υποθέσουμε ότι η καλλίτερη επιλογή είναι αυτή της εικόνας 8 δηλαδή να μην θέσουμε κατώφλι για τη θεμελιώδη συχνότητα και να ανιχνεύουμε κορυφές με ύψος >-40db. Αν εφαρμόσουμε λοιπόν τα κατώφλια αυτά για την περίπτωση ενός άλλου αρχείου, συγκεκριμένα μιας ηχογράφησης ραδιοφωνικής εκπομπής, το αποτέλεσμα είναι να έχουμε τώρα εσφαλμένη ταξινόμηση ενώ με τις αρχικές τιμές κατωφλίων, βάσει των οποίων και έγιναν όλες οι μετρήσεις της εργασίας μας, η ταξινόμηση ήταν ορθή. Το αρχικό ορθό αποτέλεσμα φαίνεται στην Εικόνα 9, ενώ το εσφαλμένο που προέκυψε με τις νέες τιμές φαίνεται στην Εικόνα 10.

Εικόνα 9: Ορθή ταξινόμηση με τις αρχικές τιμές κατωφλίων



Εικόνα 10: Εσφαλμένη ταξινόμηση με τις νέες τιμές κατωφλίων

Από τα ανωτέρω παραδείγματα και από άλλες πολλές παρόμοιες περιπτώσεις που συναντήσαμε, εξάγεται το συμπέρασμα ότι το σύστημα είναι πολύ ευαίσθητο στην επιλογή των κατωφλίων, όπως επίσης ότι υπάρχει μια αλληλεξάρτηση μεταξύ των διαφόρων κατωφλίων έτσι ώστε το αποτέλεσμα που επέρχεται με την αλλαγή ενός από αυτά μπορεί να αναιρεθεί με μια αλλαγή σε κάποιο άλλο.

5. Η συχνότητα δειγματοληψίας περιορίζεται στα 11025Hz που σημαίνει ότι η μέγιστη συχνότητα που μπορεί να υπάρξει στο δειγματοληπτημένο σήμα είναι 11025/2=5512,5Hz. Χαρακτηρίζουμε έναν ήχο ως αρμονικό αν έχει τουλάχιστον δύο αρμονικές, στην καλλίτερη περίπτωση τη δεύτερη και την τρίτη αρμονική με συχνότητες $2*f_{fund}$ και $3*f_{fund}$ αντίστοιχα. Άρα οι θεμελιώδεις

συχνότητες των αρμονικών ήχων που δύνανται να ανιχνευθούν περιορίζονται μέχρι τα 5512,5/3= 1837,5Hz. Συμπερασματικά, το σύστημά είναι σε θέση να ανιχνεύσει αρμονικούς ήχους με θεμελιώδεις συχνότητες από 500Hz, λόγω του κατωφλίου της ελάχιστης θεμελιώδους συχνότητας που σχολιάστηκε προηγουμένως, μέχρι 1837,5Hz. Έξω από αυτά τα όρια, τα σήματα ταξινομούνται είτε ως «σιγή» είτε ως «μη αρμονικοί ήχοι», όπως εξάλλου σχολιάστηκε προηγουμένως από τον πίνακα 1.

Από τους πίνακες 5 και 6 φαίνεται ότι αυτές οι συχνότητες αντιστοιχούν στις νότες C5 έως A6, αποκλείοντας έτσι το μεγαλύτερο μέρος της γκάμας των συχνοτήτων που μπορούν να παράγουν τα μουσικά όργανα.

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

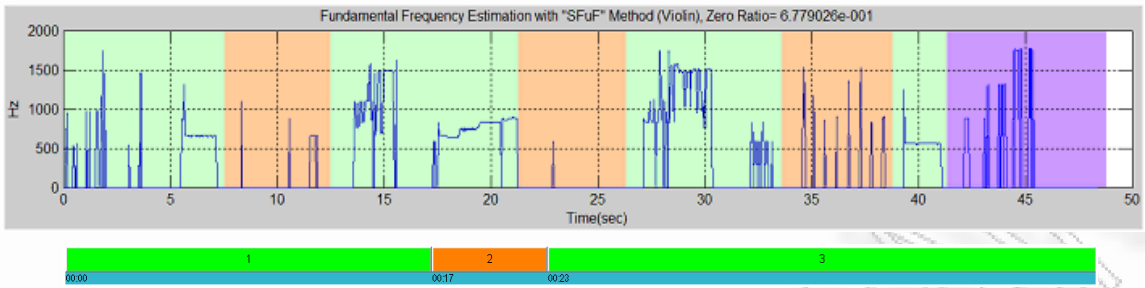
Πίνακας 5 Οι νότες και οι αντίστοιχες συχνότητές τους

Πιάνο	A0=27.50Hz έως C8=4186Hz
Χορδές κιθάρας	E2=82.41Hz, A2=110Hz, D3=146.8Hz, G3=196Hz, B3=246.9Hz, E4=329.6Hz
Χορδές μπάσου	5η χορδή: B0=30.87Hz, 4η χορδή: E1=41.20Hz, A1=55Hz, D2=73.42Hz, G2=98Hz
Χορδές μαντολίνου και βιολιού	G3=196Hz, D4=293.7Hz, A4=440Hz, E5=659.3Hz
Χορδές βιόλας και τενόρου μπάντζο	C3=130.8Hz, G3=196Hz, D4=293.7Hz, A4=440Hz
Χορδές τσέλου	C2=65.41Hz, G2=98Hz, D3=146.8Hz, A3=220Hz

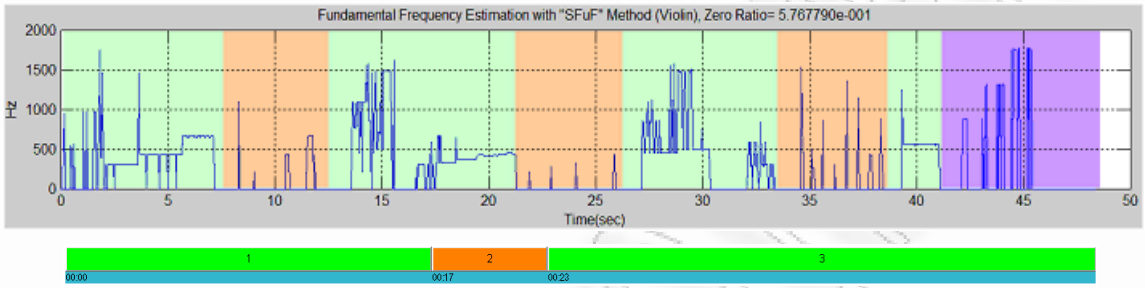
Πίνακας 6: Θεμελιώδεις συχνότητες διαφόρων έγχορδων

Το αποτέλεσμα αυτής της πηγής σφαλμάτων ταξινόμησης φαίνεται χαρακτηριστικά στο επόμενο παράδειγμα, όπου ένα βιολί παίζεται εναλλάξ με δοξάρι και «πιτσικάτο», και ο ήχος του ταξινομείται συνεχώς εσφαλμένα.

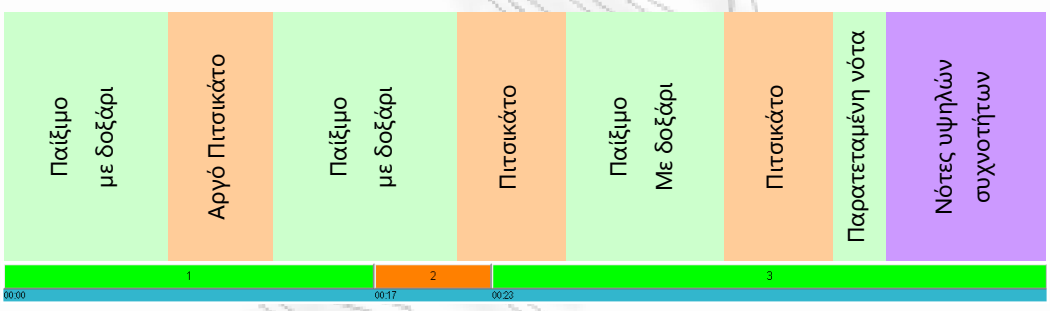
[Το Pizzicato είναι τεχνική παιξίματος εγχόρδων μουσικών οργάνων κυρίως του ηλεκτρικού μπάσου, του κοντραμπάσου και των υπόλοιπων οργάνων της οικογένειας του βιολιού, οι χορδές των οποίων κανονικά παίζονται με το δοξάρι. Pizzicato θα πει «τσιμπητά» δηλαδή τα παραπάνω όργανα μπορούν να παιχτούν τσιμπώντας τις χορδές τους με τα δάχτυλα δίνοντας έτσι έναν μονοκόμματο και πιο μπάσο ήχο στις χορδές τους. (Πηγές: Βικιπαίδεια και Wikipedia)]



Σχήμα 34: Εκτίμηση της θεμελιώδους συχνότητας ενός βιολιού με Fmin=500Hz



Σχήμα 35: Εκτίμηση της θεμελιώδους συχνότητας ενός βιολιού με Fmin=0Hz

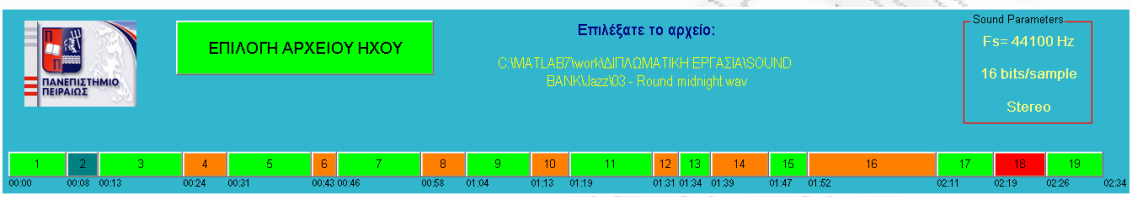


Σχήμα 36: Η διαδοχή της τεχνικής της εκτέλεσης των νότων του βιολιού και η ταξινόμησή του από το σύστημα.

Από τα σχήματα 34, 35 και 36 εξάγονται οι εξής παρατηρήσεις:

1. Όταν το βιολί παίζεται κανονικά, με το δοξάρι, η θεμελιώδης συχνότητα εκτιμάται σωστά
2. Όταν το βιολί παίζεται «πιτσικάτο», η θεμελιώδης συχνότητα εκτιμάται σωστά μόνο όταν αυτή βρίσκεται εντός των ανιχνευτικών ορίων, αλλά και πάλι ο λόγος του μηδενός παραμένει υψηλός λόγω του ότι ο παραγόμενος ήχος έχει κρουστικό χαρακτήρα και μικρή διάρκεια. Έτσι το τμήμα χαρακτηρίζεται εσφαλμένα ως «φωνή».
3. Όταν το βιολί παράγει υψηλές νότες, με θεμελιώδη συχνότητα πάνω από 1837,5Hz, δεν μπορούν πλέον να ανιχνευθούν οι αρμονικές της και η θεμελιώδης συχνότητα (SFuF) τίθεται στο μηδέν με αποτέλεσμα ο λόγος του μηδενός να παραμένει υψηλός και το τμήμα να χαρακτηρίζεται και πάλι εσφαλμένα ως «φωνή».

4. Όταν για τον υπολογισμό της SFuF, δεν τεθεί το κατώφλι ελάχιστης συχνότητας Fmin=500Hz, ανιχνεύονται μεν περισσότερες νότες, κυρίως κατά τη διάρκεια παιξίματος «πιτσικάτο», με αποτέλεσμα τη μείωση του λόγου του μηδενός, αλλά η μείωση αυτή δεν είναι αρκετή για να ταξινομηθεί το τμήμα ορθά.
5. Ο ήχος ορισμένων μουσικών οργάνων, ιδίως όταν εκτελούν σόλο ή κυριαρχούν στη μουσική σκηνή, επιφέρει σύγχυση στον αλγόριθμο όπως χαρακτηριστικά φαίνεται στο επόμενο παράδειγμα:



Εικόνα 11: Σύγχυση ταξινόμησης

Στην Εικόνα 11 φαίνεται η αναποφασιστικότητα του αλγορίθμου να οδηγήσει τον ήχο του αρχείου στο σωστό ταξινομητικό μονοπάτι. Συγκεκριμένα, στο αρχείο αυτό ακούγονται κατά κύριο λόγο ένα τύμπανο (Εικόνα 12) και μία τρομπέτα (Εικόνα 13).

Εικόνα 12: Σκουπάκια

Εικόνα 13: Τρομπέτα

Στα τμήματα που έχουν ταξινομηθεί (εσφαλμένα) ως «φωνή» κυριαρχεί ο ήχος του τύμπανου, που προσομοιάζει με θόρυβο λόγω του είδους της μπαγκέτας που χρησιμοποιείται (σκουπάκια), ενώ στα τμήματα που έχουν ταξινομηθεί (ορθά) ως «αρμονικός ήχος» κυριαρχεί η τρομπέτα που εκτελεί μακρόσυρτες νότες και που πράγματι ο ήχος της είναι αρμονικός. Στο τμήμα που έχει ταξινομηθεί ως «μουσική» ακούγεται και δεύτερο πνευστό όργανο που προσθέτει θεμελιώδεις συχνότητες και ξεπερνιέται έτσι το κατώφλι των 15

θεμελιωδών. Ο ήχος τότε χαρακτηρίζεται ορθά, όχι πλέον ως «αρμονικός» αλλά ως «μουσική».

6. Η «ηχητική σκίαση». Όταν σε ένα πλαίσιο συνυπάρχουν δύο φασματικές κορυφές όπου η πρώτη έχει συχνότητα κάτω των 500Hz και η δεύτερη έχει ανιχνεύσιμη θεμελιώδη συχνότητα, τότε, αν η πρώτη κορυφή έχει πολύ μεγάλο ύψος σε σχέση με τη δεύτερη «σκιάζει» τη δεύτερη διότι η κανονικοποίηση των κορυφών γίνεται στη μέγιστη τιμή και έτσι η δεύτερη κορυφή παραμένει κάτω του ανιχνευτικού κατωφλίου Pxx>-30db. Σε αυτή την περίπτωση δεν ανιχνεύεται η θεμελιώδης συχνότητα με αποτέλεσμα να μειώνεται ο λόγος του μηδενός. Έγινε προσπάθεια να εκτιμηθεί η θεμελιώδης συχνότητα αγνοώντας τις κορυφές με συχνότητα κάτω των 500Hz, κανονικοποιώντας μόνο τις υπόλοιπες, αλλά αυτό οδήγησε σε χειρότερα σφάλματα λόγω του ότι ανιχνευόντουσαν τότε κορυφές που οφείλονταν στο θόρυβο υποβάθρου και που δεν είχαν καμία σχέση με το σήμα.
7. Όταν συνυπάρχουν οι φωνές δύο ή περισσοτέρων ομιλητών που ομιλούν φυσιολογικά και όχι συγχρονισμένα (δηλαδή δεν απαγγέλλουν κάτι ταυτόχρονα), τότε αυτοί σπανίως εκφέρουν σε συγχρονισμό τα έμφωνα και τα άφωνα τμήματα του λόγου. Ο λόγος του μηδενός παραμένει τότε χαμηλός και το τμήμα ταξινομείται εσφαλμένα ως «μουσική».
8. Τέλος, θα πρέπει να σημειώσουμε ότι η απόφαση περί του αν ένας ήχος είναι «μουσική» ή όχι βασίζεται πολλές φορές στα υποκειμενικά κριτήρια του ακροατή. Για παράδειγμα, ο ήχος που παράγεται από το συγκρότημα «Stomp», Εικόνα 14, μπορεί να χαρακτηριστεί ως «μουσική» με αντικειμενικά και μετρήσιμα κριτήρια;

Εικόνα 14: Το συγκρότημα "Stomp"

Για το λόγο αυτό οι μετρήσεις μας ενέχουν ένα μεγάλο βαθμό υποκειμενικότητας που αφήνει ένα περιθώριο συστηματικού πειραματικού σφάλματος δύσκολα εκτιμήσιμου.

8. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Σε αυτή την εργασία υλοποιήσαμε ένα σύστημα αυτόματης κατάτμησης και ταξινόμησης οπτικοακουστικών δεδομένων που βασίζεται στην ανάλυση του ηχητικού περιεχομένου. Οι προηγούμενες εργασίες πάνω στην κατάτμηση και ταξινόμηση έχουν εστιάσει κυρίως στην οπτική πληροφορία. Το σύνηθες πλαίσιο εργασίας είναι η ανίχνευση των αλλαγών κατά τη λήψη χρησιμοποιώντας τη διαφορά του ιστογράμματος και τα διανύσματα κίνησης, για να εξαχθούν έτσι καρέ-κλειδιά αντιπροσωπευτικά της κάθε λήψης. Όμως, αυτή η οπτικά βασισμένη επεξεργασία οδηγεί συχνά σε έναν πολύ λεπτομερή κατακερματισμό της οπτικοακουστικής ακολουθίας σε σχέση με το εννοιολογικό περιεχόμενο των δεδομένων. Για παράδειγμα, στην οπτική ακολουθία κατά την εκτέλεση ενός τραγουδιού, μπορεί να υπάρχουν λήψεις στις οποίες να εμφανίζεται διαδοχικά ο τραγουδιστής, το συγκρότημα, το ακροατήριο και μερικά άλλα πλάνα. Σύμφωνα με την οπτική πληροφορία, αυτές οι λήψεις θα ταξινομηθούν ξεχωριστά. Αλλά σύμφωνα με την ηχητική πληροφορία, ξέρουμε ότι όλες ενυπάρχουν σε μια μοναδική εκτέλεση ενός τραγουδιού. Έτσι, στην προσέγγισή μας για την ανάλυση του βίντεο, το πρώτο βήμα είναι η κατάτμηση της ακολουθίας του βίντεο σε εννοιολογικές σκηνές που βασίζονται σε ηχητικά γνωρίσματα και η ταξινόμηση των σκηνών με τους προτεινόμενους αλγορίθμους.

Ενώ οι τρέχουσες προσεγγίσεις για την ανάλυση του ηχητικού περιεχομένου έχουν συνήθως αναπτυχθεί για συγκεκριμένα σενάρια, σε αυτή την εργασία ερευνήθηκε ένα γενικό σχήμα που καλύπτει όλα τα είδη των ηχητικών σημάτων. Αναλύθηκαν τρία ηχητικά χαρακτηριστικά όπως η ενέργεια, ο ρυθμός διέλευσης του μηδενός, και η θεμελιώδης συχνότητα για να αποκαλυφθούν οι διαφορές μεταξύ των διαφόρων τύπων των ηχητικών δεδομένων. Προτάθηκαν επίσης οι μέθοδοι της εκτίμησης της θεμελιώδους συχνότητας και της εξαγωγής των φασματικών ιχνών από το φάσμα που παρήχθη από το AR μοντέλο. Υλοποιήθηκε μια διαδικασία κατάτμησης και ταξινόμησης του συνοδευτικού ηχητικού σήματος των οπτικοακουστικών δεδομένων σε πραγματικό χρόνο, βασισμένη στην ανάλυση των ηχητικών χαρακτηριστικών γνωρισμάτων.

Τονίζοντας ότι η αξιολόγηση ενός παρόμοιου συστήματος ταξινόμησης εξαρτάται έντονα από το σύνολο των δεδομένων που θα δοθούν για ταξινόμηση και από την υποκειμενικότητα του παρατηρητή-αξιολογητή, καταλήγουμε στο συμπέρασμα ότι το προτεινόμενο σύστημα κατάτμησης και ταξινόμησης επιτυγχάνει να θέσει με ακρίβεια τα όρια των τμημάτων και ταξινομεί ορθά περί το 72% αυτών. Όμως, εμφανίσθηκαν και σφάλματα ταξινόμησης που οφείλονται:

1. Στην αυθαίρετη επιλογή των κατωφλίων που γίνεται με βάση ένα σύνολο αρχείων αναφοράς που δεν εξασφαλίζουν καθολική επιτυχία σε πραγματικά δεδομένα.
2. Στο γεγονός ότι το αποτέλεσμα παρουσιάζει μεγάλη ευαισθησία στην επιλογή των κατωφλίων.
3. Στην παρατηρούμενη αλληλεξάρτηση των κατωφλίων.
4. Στις εγγενείς αδυναμίες της προσέγγισης που εφαρμόζεται στο σύστημα για τη σωστή εκτίμηση της θεμελιώδους συχνότητας και συνακόλουθα του λόγου του μηδενός.
5. Στη φύση του ηχητικού περιεχομένου που μπορεί να ξεγελάσει τον αλγόριθμο ταξινόμησης.
6. Στον πρώτο αδρό διαχωρισμό των ήχων σε έχοντες και μη έχοντες μουσικό περιεχόμενο.
7. Ίσως, στις ασαφείς περιγραφές της πρωτότυπης εργασίας που δεν επέτρεψαν την πλήρη και ακριβή αναπαραγωγή του συστήματος.

Θα πρέπει να γίνει προσπάθεια να βρεθούν τρόποι αντιμετώπισης των πηγών των σφαλμάτων αν και πολλές φορές αυτό δεν είναι πάντα εφικτό λόγω της φύσης τους. Ως μια σημαντική μελλοντική ερευνητική κατεύθυνση προτείνεται η πολυμορφική επεξεργασία των οπτικοακουστικών δεδομένων καθώς και η χρήση συνδυασμού ταξινομητών.

9. ΑΝΑΦΟΡΕΣ – ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] S. W. Smoliar and H. Zhang, "Content-based video indexing and retrieval," IEEE Multimedia, pp. 62–72, Summer 1994.

[2] M. Flickner, H. Sawhney, and W. Niblack et al., "Query by image and video content: The QBIC system," Computer, vol. 28, no. 9, pp. 23–32, 1995.

[3] S.-F. Chang,W. Chen, and H. J. Meng et al., "A fully automated content based video search engine supporting spatio-temporal queries," IEEE Trans. Circuits Syst. Video Technol., vol. 8, pp. 602–615, Sept. 1998.

[4] J. Saunders, "Real-time discrimination of broadcast speech/music," in Proc. Int. Conf. Acoustics, Speech, Signal Processing' 96, vol. 2, Atlanta, GA, May 1996, pp. 993–996.

[5] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in Proc. Int. Conf. Acoustics, Speech, Signal Processing' 97, Munich, Germany, Apr. 1997.

[6] L. Wyse and S. Smoliar, "Toward content-based audio indexing and retrieval and a new speaker discrimination technique," Inst. Syst. Sci., Nat. Univ. Singapore, http://www.iss.nus.sg/People/lwyse/lwyse.html, Dec. 1995.

[7] D. Kimber and L.Wilcox, "Acoustic segmentation for audio browsers," in Proc. Interface Conf., Sydney, Australia, July 1996.

[8] S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic audio content analysis," Praktische Informatik IV, Univ. Mannheim, Mannheim, Germany, http://www.informatik.uni-mannheim.de/pfeiffer/publications/, Apr. 1996.

[9] A. Ghias, J. Logan, and D. Chamberlin, "Query by humming-musical information retrieval in an audio database," in Proc. ACM Multimedia Conf., 1995, pp. 231–235.

[10] J. Foote, "Content-based retrieval of music and audio," Proc. SPIE, 1997.

[11] E. Wold, T. Blum, and D. Keislar et al., "Content-based classification, search, and retrieval of audio," IEEE Multimedia, pp. 27–36, Fall 1996.

[12] G. Smith, H. Murase, and K. Kashino, "Quick audio retrieval using active search," in Proc. Int. Conf. Acoustics, Speech, Signal Processing' 98, Seattle, WA, May 1998, pp. 3777–3780.

[13] Z. Liu, J. Huang, and Y.Wang et al., "Audio feature extraction and analysis for scene classification," in Proc. IEEE 1st Multimedia Workshop, 1997.

[14] Z. Liu, J. Huang, and Y. Wang, "Classification of TV programs based on audio information using hidden Markov model," in Proc. IEEE 2nd Workshop Multimedia Signal Processing, Redondo Beach, CA, Dec. 1998, pp. 27–32.

[15] Z. Liu and Q. Huang, "Classification of audio events in broadcast news," in Proc. IEEE 2nd Workshop Multimedia Signal Processing, Dec. 1998, pp. 364–369.

[16] N. Patel and I. Sethi, "Audio characterization for video indexing," in Proc. SPIE Conf. Storage Retrieval Still Image Video Databases, vol. 2670, San Jose, CA, 1996, pp. 373–384.

[17] K. Minami, A. Akutsu, and H. Hamada et al., "Video handling with music and speech detection," IEEE Multimedia, pp. 17–25, Fall 1998.

[18] J. Huang, Z. Liu, and Y.Wang, "Integration of audio and visual information for content-based video segmentation," in Proc. IEEE Conf. Image Processing, Oct. 1998.

[19] M. R. Naphade, T. Kristjansson, and B. Frey et al., "Probabilistic multimedia objects (MULTI-JECTS): A novel approach to video indexing and retrieval in multimedia systems," in Proc. IEEE Conf. Image Processing, Chicago, IL, Oct. 1998.

[20] J. S. Boreczky and L. D. Wilcox, "A hidden Markov model framework for video segmentation using audio and image features," in Proc. Int. Conf. Acoustics, Speech, Signal Processing' 98, May 1998, pp. 3741–3744.

[21] A. S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press, 1990.

[22] G. J. Brown and M. Cooke, "Computational auditory scene analysis," Comput. Speech Lang., vol. 8, no. 2, pp. 297–336, 1994.

[23] M.Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ. , Stanford, CA, 1985.

[24] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Mass. Inst. Technol., Cambridge, MA, 1996.

[25] B. L. Vercoe,W. G. Gardner, and E. D. Scheirer, "Structured audio: Creation, transmission, and rendering of parametric sound representations," Proc. IEEE, vol. 86, pp. 922–939, May 1998.

[26] L. Rabiner and R. Schafer, Digital Processing of Speech Signals. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[27] A. Choi, "Real-time fundamental frequency estimation by least-square fitting," IEEE Trans. Speech Audio Processing, vol. 5, pp. 201–205, Mar. 1997.

[28] B. Doval and X. Rodet, "Estimation of fundamental frequency of music sound signals," in Proc. Int. Conf. Acoustics, Speech, Signal Processing' 91, vol. 5, Toronto, ON, Canada, Apr. 1991, pp. 3657–3660.

[29] W. B. Kuhn, "A real-time pitch recognition algorithm for music applications," Comput. Music J., vol. 14, no. 3, pp. 60–71, Fall 1990.

[30] F. Everest, The Master Handbook of Acoustics. New York: McGraw- Hill, 1994.

10. ΤΑ ΑΡΧΕΙΑ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

10.1. Βασικές συναρτήσεις

1. Colorbar.m (*Ιωάννης Μπενέτος*)

   Η γραφική διεπαφή της εργασίας. Αλληλεπιδρά με το χρήστη, καλεί και εμφανίζει τα αποτελέσματα της συνάρτησης Extractor.m

2. Extractor.m (*Ιωάννης Μπενέτος*)

   Επιστρέφει τα τμήματα (segments), τους τύπους των τμημάτων, τον αριθμό των δειγμάτων, την Fs, τα bits/sample και τον αριθμό των καναλιών (mono-stereo).

10.2. Συναρτήσεις εξαγωγής χαρακτηριστικών γνωρισμάτων

1. stEnergy.m (*A. Pikrakis, S. Theodoridis, K. Koutroumbas, D. Cavouras*)

   Υπολογίζει την ακολουθία τιμών της ενέργειας βραχέως χρόνου.

2. stZeroCrossingRate.m (*A. Pikrakis, S. Theodoridis, K. Koutroumbas, D. Cavouras*)

   Υπολογίζει την ακολουθία τιμών του ρυθμού διέλευσης του μηδενός βραχέως χρόνου.

3. sfSFuF.m (*Ιωάννης Μπενέτος*)

   Υπολογίζει την τιμή της θεμελιώδους συχνότητας ενός πλαισίου αν το σήμα είναι αρμονικό, δηλαδή αν υπάρχουν αρμονικές συχνότητες, αλλιώς επιστρέφει τιμή μηδέν.

4. stSFuF.m (*Ιωάννης Μπενέτος*)

   Καλεί επαναληπτικά τη συνάρτηση sfSFuF.m και εξάγει την ακολουθία τιμών της θεμελιώδους συχνότητας μιας σειράς πλαισίων με βάση την αρμονικότητα.

10.3. Βοηθητικές συναρτήσεις

1. Count.m (*Richard Medlock*)

   Απαριθμεί τα στοιχεία του διανύσματος A που ικανοποιούν τα κριτήρια που ορίζονται στο όρισμα B.

2. sec2hms.m (*Ιωάννης Μπενέτος*)

   Μετατρέπει το χρόνο από δευτερόλεπτα σε χρόνο στη μορφή: hh:mm:ss