ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Αυτόματη Ανίχνευση Σημαντικών Ηχητικών Γεγονότων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Γεώργιου Ν. Αναστασίου

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

Αθήνα, Ιούλιος 2015

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

Αυτόματη Ανίχνευση Σημαντικών Ηχητικών Γεγονότων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Γεώργιου Ν. Αναστασίου

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την Δευτέρα 20 Ιουλίου 2015.

.....	.....	.....
Πέτρος Μαραγκός Καθηγητής Ε.Μ.Π.	Γεράσιμος Ποταμιάνος Αναπληρωτής Καθηγητής Πανεπιστημίου Θεσσαλίας	Ευίτα-Σταυρούλα Φωτεινέα Ερευνήτρια Α Ινστιτούτου Επεξεργασίας Λόγου

Αθήνα, Ιούλιος 2015

.....
Γεώργιος Ν. Αναστασίου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright© Γεώργιος Ν. Αναστασίου, 2015.
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Στην παρούσα διπλωματική εργασία εξετάζεται η υπολογιστική προσέγγιση της ακουστικής προσοχής του ανθρώπου, και συγκεκριμένα η αυτόματη ανίχνευση ήχων που ενεργοποιούν τον κάτωθεν μηχανισμό της προσοχής (*bottom-up*). Ενεργοποίηση του κάτωθεν μηχανισμού της προσοχής παρατηρείται όταν οι ήχοι γίνονται αυθόρμητα αντιληπτοί από τους ανθρώπους, ανεξάρτητα από την βούληση τους. Ήχοι που κατέχουν αυτή την ιδιότητα θα ονομάζονται σημαντικοί (*salient*). Στόχος, επομένως, αυτής της εργασίας είναι η αυτόματη ανίχνευση σημαντικών ήχων (γεγονότων) σε αρχεία ήχου.

Προς επίτευξη αυτού του σκοπού, αρχικά παρουσιάζεται το μοντέλο των Kayser et al. το οποίο ανιχνεύει σημαντικά ηχητικά γεγονότα μέσω της επεξεργασίας του φασματογραφήματος του ήχου. Η έξοδος του μοντέλου είναι ένας διδιάστατος χάρτης σημαντικότητας, από τον οποίο υπολογίζεται καμπύλη σημαντικότητας και πραγματοποιείται ταξινόμηση των ηχητικών σκηνών. Επίσης, χρησιμοποιείται η έννοια του *gist* μιας σκηνής από τη βιβλιογραφία αντίληψης εικόνων και δομούνται διανύσματα από τον χάρτη τα οποία ταξινομούνται με τον αλγόριθμο kNN. Παρατηρείται συσχέτιση της εξόδου του μοντέλου με βασικούς μηχανισμούς της ακουστικής αντίληψης.

Στη συνέχεια προτείνεται μία τροποποίηση του μοντέλου των Kayser et al, όπου το φασματογράφημα αντικαθίσταται από μονοδιάστατα χαρακτηριστικά που εξάγονται σε πλαίσιο βραχέως χρόνου από το ηχητικό σήμα. Γίνεται προσαρμογή κάθε σταδίου του μοντέλου για το χειρισμό μονοδιάστατων καμπυλών. Η έξοδος του μοντέλου είναι μία καμπύλη σημαντικότητας με βάση την οποία χαρακτηρίζονται οι σκηνές ως σημαντικές ή μη. Με χρήση των χαρακτηριστικών, δημιουργούνται ιστογράμματα σε αναλογία με τη μέθοδο *bag-of-words* στην Όραση Υπολογιστών, και χειριζόμενα αυτά ως διανύσματα πραγματοποιείται ταξινόμηση των ηχητικών σκηνών με χρήση SVM. Το τροποποιημένο μοντέλο υπερβαίνει σε απόδοση το αρχικό των Kayser et al.

Επίσης δοκιμάζονται τα κλασσικά χαρακτηριστικά της βιβλιογραφίας, MFCC και AM-FM, στο πρόβλημα ανίχνευσης σημαντικών γεγονότων. Επιπλέον, πραγματοποιείται μια υψηλότερου επιπέδου προσέγγιση και εξάγονται διαφορετικά χαρακτηριστικά για τα σημεία του ηχητικού σήματος που εμφανίζεται φωνή από αυτά στα οποία δεν εμφανίζεται. Τέλος προτείνονται μελλοντικές κατευθύνσεις για έρευνα και επέκταση αυτής της εργασίας.

Τα πειράματα γίνονται σε ηχητικά σήματα που προέρχονται από βάση δεδομένων που περιέχει αποσπάσματα από κινηματογραφικές ταινίες. Ως βάση αναφοράς χρησιμοποιούνται ανθρώπινες επισημειώσεις της σημαντικότητας. Δηλαδή, άτομα που άκουσαν τα ηχητικά αρχεία, σημείωσαν ποια μέρη τους φάνηκαν σημαντικά.

Λέξεις Κλειδιά: Ψηφιακή επεξεργασία σήματος, επεξεργασία ήχου, ακουστική προσοχή, σημαντικά γεγονότα, επεξεργασία εικόνας, φασματογράφημα, χάρτης σημαντικότητας, καμπύλη σημαντικότητας, χαρακτηριστικά βραχέως χρόνου, μηχανική μάθηση, ομαδοποίηση δεδομένων, ιστόγραμμα.

Abstract

The present diploma thesis makes a computational approach of human auditory attention, and specifically the automatic detection of sounds that activate the bottom-up mechanism of auditory attention. The bottom-up mechanism of attention is activated when sounds are perceived unconsciously and involuntarily by humans. Such sounds will be called *salient*. The goal of this thesis is the automatic detection of salient sounds (events) in audio files.

To achieve this goal, initially, it is presented a model developed by Kayser et al, that detects salient sound events by processing the spectrogram of the sound. The output of Kayser's model is a two-dimensional saliency map from which it is computed a saliency curve in order to classify sound scenes. Also, it is used the meaning of the *gist* of a scene from the image perception bibliography, and vectors are created using the maps, that are classified by the kNN algorithm. It is observed correlation between the output of the model and basic mechanisms of auditory perception.

Next, it is proposed a modification of Kayser's model, where the spectrogram is replaced by one dimensional features that are extracted in short time framework from the sound signal. Every stage of the model is modified properly in order to handle one-dimensional curves. The output of the model is a saliency curve by which the auditory scenes are characterized as salient or not. Using the sort time features, histograms are created in correspondence to the bag-of-words method in Computer Vision, and they are classified by SVM. The modified model outperforms the initial model from Kayser et al.

Also, the performance of the classic features of the bibliography, MFCC and AM-FM, is tested on the problem of salient event detection. Furthermore, it is made a higher level approach and different features are extracted at points in time that human speech is present, than points that speech is not present. Finally, directions for future research and further development of this thesis are proposed.

The experiments are performed on a database that contains sound files from movies excerps. Human annotations of saliency are used as ground truth data. That is, people listened to the sound files and annotated which parts of them, they considered as salient.

Keywords: Digital signal processing, sound processing, auditory attention, salient events, image processing, spectrogram, saliency map, saliency curve, short-time features, machine learning, data clustering, histogram.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή Πέτρο Μαραγκό, επιβλέποντα της διπλωματικής, που μου έδωσε την ευκαιρία πραγματοποίησης αυτής της εργασίας και της ένταξης μου στο εργαστήριο. Οι ερευνητικές συναντήσεις που πραγματοποιήθηκαν κατά την διάρκεια αυτής της εργασίας, στις οποίες συμμετείχαν άτομα από διάφορους επιστημονικούς κλάδους, με βοήθησαν να αποκτήσω μια πιο σφαιρική εικόνα του προβλήματος με το οποίο ασχολήθηκα, καθώς και του τρόπου με τον οποίο πραγματοποιείται η έρευνα στα πλαίσια ερευνητικών ομάδων.

Θέλω, επίσης, να ευχαριστήσω τον μεταδιδακτορικό ερευνητή Αθανάσιο Κατσαμάνη για τις άκρως εποικοδομητικές συζητήσεις που είχαμε, και τις συστάσεις προς την πιο θεμελιώδη αντιμετώπιση του προβλήματος(η οποία τελικά εγκαταλείφθηκε λόγω χρονικών περιορισμών και δεν αναφέρεται εδώ), που οδήγησαν στην βαθύτερη κατανόηση του.

Θα ήθελα να ευχαριστήσω την μεταδιδακτορική ερευνήτρια Αθανασία Ζλατίντση, για την παροχή πολύτιμου υλικού που αφορά την βάση δεδομένων στην οποία εκτελέσθηκαν τα πειράματα, καθώς και μέρος του κώδικα που χρησιμοποιήθηκε.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Με τον όρο προσοχή (*attention*) νοείται η διαδικασία της επιλογής ενός υποσυνόλου των ερεθισμάτων που εμφανίζονται στο περιβάλλον για περαιτέρω επεξεργασία, και απόρριψη των υπολοίπων. Ο όγκος πληροφορίας που υπάρχει στα περιβάλλοντα όπου ζει ο άνθρωπος και άλλοι οργανισμοί είναι αρκετά μεγάλος και απαιτεί σημαντικό πλήθος υπολογιστικών πόρων για την επεξεργασία του. Επιπλέον το μεγαλύτερο μέρος αυτού δεν επηρεάζει άμεσα τη δράση και την επιβίωση των οργανισμών. Ως συνέπεια μέσω της εξέλιξης και της προσαρμογής των ειδών στα περιβάλλοντα που ζουν, έχουν αναπτυχθεί μηχανισμοί που τους επιτρέπουν να φιλτράρουν την εισερχόμενη πληροφορία και να επεξεργάζονται μόνο την πιο σχετική για την επιβίωση και την εκπλήρωση των στόχων τους [16, 36]. Το σύνολο αυτών των μηχανισμών αποτελούν την προσοχή.

Η μελέτη της ανθρώπινης προσοχής έχει αποτελέσει αντικείμενο έρευνας των γνωστικών επιστημών κατά τη διάρκεια του 20ου αιώνα και συνεχίζεται με αμείωτη ένταση έως και σήμερα. Στόχος των ερευνητών είναι να εξηγήσουν πώς ο άνθρωπος αλληλεπιδρά με τα διάφορα ερεθίσματα που δέχεται από το περιβάλλον του, όπως ηχητικά και οπτικά, και να προσδιορίσουν ποια είναι η επεξεργασία που πραγματοποιείται από τα περιφερειακά αισθητήρια όργανα έως την αναπαράσταση των ερεθισμάτων στον ανθρώπινο εγκέφαλο που οδηγεί τελικά στον τρόπο αντίληψης τους [59]. Ερευνητικοί κλάδοι που ενδιαφέροντα άμεσα για την μελέτη της προσοχής του ανθρώπου περιλαμβάνουν αυτούς της γνωστικής ψυχολογίας, ιατρικής, νευρο-επιστήμης, νευρο-βιολογίας, και κοινωνιολογίας.

Στην παρούσα εργασία θα ασχοληθούμε συγκεκριμένα με την ακουστική προσοχή του ανθρώπου. Ο όρος ακουστική προσοχή αναφέρεται στο μέρος των μηχανισμών της προσοχής που είναι υπεύθυνοι για την αντίληψη της ηχητικής πληροφορίας. Στόχος της μελέτης της ακουστικής αντίληψης είναι να γίνουν γνωστά τα στάδια επεξεργασίας της ηχητικής πληροφορίας, από τα περιφερειακά ακουστικά όργανα (όπως το αυτί), έως πιο κεντρικές διαδικασίες στον ανθρώπινο εγκέφαλο. Ζητούμενο είναι η κατανόηση σε ποιο στάδιο πραγματοποιείται το φιλτράρισμα της πληροφορίας, η αποκοπή της μη-σημαντικής και η περαιτέρω επεξεργασία της σημαντικής. Επίσης, πως καθορίζεται εάν η πληροφορία είναι σημαντική ή όχι. Τελικά ποια είναι η αναπαράσταση στον ανθρώπινο εγκέφαλο και πως λαμβάνεται απόφαση για την παρακολούθηση των ηχητικών πηγών που υπάρχουν στα περιβάλλοντα που κινείται ο άνθρωπος [20].

Δύο μηχανισμοί της προσοχής γενικότερα, και της ακουστικής προσοχής πιο συγκεκριμένα, που έχουν παρατηρηθεί σε πειραματικές διαδικασίες, είναι ο άνωθεν μηχανισμός (*top-down*) και ο κάτωθεν μηχανισμός της προσοχής (*bottom-up*). Στον άνωθεν μηχανισμό, η προσοχή κατευθύνεται από τη βούληση και τους στόχους του ανθρώπου και αναφέρεται επίσης ως ενδογενής μηχανισμός. Η αναζήτηση ενός προσώπου μέσα σε ένα πλήθος ανθρώπων είναι ένα παράδειγμα όπου ο άνωθεν μηχανισμός της προσοχής λαμβάνει χώρα. Ο παρατηρητής γνωρίζει τα χαρακτηριστικά του προσώπου που αναζητά, και πρέπει να ψάξει μέσα στο πλήθος για την εύρεση του. Η αναζήτηση πιθανώς πρέπει να γίνει σειριακά, δηλαδή να εξετασθούν τα πρόσωπα ένα-ένα μέχρι να βρεθεί αυτό με τα επιθυμητά χαρακτηριστικά. Ως συνέπεια, ο άνωθεν μηχανισμός της προσοχής είναι συνήθως αργός. Αντίστοιχα στην ακουστική προσοχή ο παρατηρητής προσπαθεί να ανιχνεύσει τη φωνή του ατόμου που αναζητά μέσα στο υπόλοιπο βουητό από φωνές.

Στον κάτωθεν μηχανισμό της προσοχής (bottom-up), ο άνθρωπος δεν ελέγχει που θα στρέψει την προσοχή. Είναι ένας μηχανισμός που εξαρτάται από τις ιδιότητες του ερεθίσματος με το οποίο θα έλθει σε επαφή ο άνθρωπος, και συνεπώς συχνά καλείται εξωγενής μηχανισμός. Η μετάβαση της προσοχής προς ένα αντικείμενο γίνεται χωρίς την βούληση του ανθρώπου και με μεγάλη ταχύτητα. Ο κάτωθεν μηχανισμός είναι πολύ σημαντικό στοιχείο για την επιβίωση των οργανισμών καθώς τους βοηθά να αντιληφθούν άμεσα μεταβολές του περιβάλλοντος οι οποίες πιθανώς σημαίνουν ότι πρέπει να αλλάξουν τον άμεσο στόχο τους λόγω ύπαρξης κινδύνου ή ωφέλειας.

Η ύπαρξη ενός συνόλου κόκκινων λουλουδιών μέσα σε ένα πράσινο λιβάδι θα γίνει άμεσα αντιληπτή όταν το βλέμμα του θεατή στραφεί προς την μεριά τους. Επίσης, η οπτική προσοχή τείνει να επικεντρώνεται στα περιγράμματα των αντικειμένων στον κόσμο. Στην ακουστική προσοχή η εμφάνιση μιας σειρήνας σε έναν δρόμο με έντονη κίνηση θα γίνει άμεσα αντιληπτή λόγω των ιδιοτήτων του ήχου της σειρήνας που την κάνουν να ξεχωρίζει από τον υπόλοιπο θόρυβο. Σε αυτές τις περιπτώσεις ενεργοποιείται ο κάτωθεν μηχανισμός της προσοχής.

Ένα ευρέως χρησιμοποιούμενο παράδειγμα στην βιβλιογραφία, όπου λαμβάνουν μέρος και οι δύο μηχανισμοί της προσοχής είναι το φαινόμενο του κοκτέιλ πάρτι (*cocktail party phenomenon*) [10, 7]. Άτομα παρευρισκόμενα σε μία αίθουσα με πολύ κόσμο, έχουνε την ικανότητα να στρέφουν ηθελημένα την προσοχή τους προς τις διάφορες συζητήσεις αποκόπτοντας τους υπόλοιπους ήχους. Σε αυτή την περίπτωση ο άνωθεν μηχανισμός της προσοχής λαμβάνει χώρα καθώς η προσοχή καθοδηγείται από τους στόχους του κάθε ατόμου. Εάν, όμως, σπάσει κάποιο ποτήρι η προσοχή όλων στρέφεται στιγμιαία και αυθόρμητα προς αυτό, ενώ ήταν επικεντρωμένη αλλού. Παρατηρείται εμφάνιση του κάτωθεν μηχανισμού της προσοχής.

Παρά τους περιορισμούς που έχει η ανθρώπινη αντίληψη, ο άνθρωπος ξεπερνά την επίδοση των μηχανών στην πραγματοποίηση πολλών διεργασιών που αφορούν την επεξεργασία και κατανόηση της εισερχόμενης πληροφορίας. Για παράδειγμα είναι αρκετά εύκολο για τον άνθρωπο να αναγνωρίσει αντικείμενα του κόσμου υπό διαφορετικές όψεις, να αναγνωρίσει την φωνή ενός οικείου ατόμου υπό συνθήκες έντονου θορύβου, καθώς και την διάκριση του είδους του μουσικού οργάνου που παράγει έναν ήχο υπό την παρουσία άλλων. Ως συνέπεια ο τρόπος που λειτουργεί η ανθρώπινη αντίληψη έχει τραβήξει το εν-

διαφέρον επιστημονικών κλάδων όπως αυτών της επιστήμης των υπολογιστών και των μηχανικών. Σκοπός αυτών των κοινοτήτων είναι να λάβουν στοιχεία και να εμπνευστούν από λειτουργίες της ανθρώπινης αντίληψης ώστε να τα ενσωματώσουν στα μηχανικά τους συστήματα και να βελτιώσουν την απόδοσή τους.

Μία άλλη προσέγγιση που ακολουθείται σε υπολογιστικούς κλάδους είναι η προσπάθεια αναπαραγωγής πειραματικών δεδομένων που έχουν ληφθεί από ανθρώπους και αφορούν τον τρόπο που κατευθύνουν την προσοχή τους προς τα διάφορα αντικείμενα, με καθαρά υπολογιστικά μέσα χωρίς να λαμβάνονται σε μεγάλο βαθμό υπόψη θεωρίες της προσοχής. Η κατεύθυνση αυτή μπορεί να αποτελέσει ανάδραση για την περαιτέρω κατανόηση του τρόπου που λειτουργεί η ανθρώπινη αντίληψη καθώς η εύρεση ενός μηχανισμού που αναπαράγει σε μεγάλο βαθμό την ανθρώπινη συμπεριφορά, μπορεί να είναι μια πολύ καλή προσέγγιση του τρόπου που πραγματικά δουλεύει η ανθρώπινη αντίληψη. Αυτή η προσέγγιση είναι που ακολουθείται σε αυτή την εργασία και περιγράφεται συνοπτικά στην επόμενη ενότητα.

1.1 Αντικείμενο της Διπλωματικής Εργασίας

Στόχος της παρούσας εργασίας είναι με χρήση υπολογιστικών μεθόδων η αυτόματη ανίχνευση σε ηχητικά δεδομένα χρονικών στιγμών όπου λαμβάνει χώρα ο κάτωθεν μηχανισμός της ακουστικής προσοχής, όπως αυτός ορίστηκε προηγουμένως. Οι χρονικές στιγμές αυτές θα καλούνται στο εξής *σημαντικές* ή *προεξέχουσες* (*salient*). Τα δεδομένα των σημαντικών ηχητικών στιγμών θα καλούνται αντίστοιχα σημαντικά ή προεξέχοντα, ή θα λέμε ότι υπάρχει εμφάνιση *σημαντικών ηχητικών γεγονότων (salient sound event)*. Οι χρονικές στιγμές που δεν λαμβάνει χώρα ο κάτωθεν μηχανισμός της προσοχής θα καλούνται μη-σημαντικές ή μη-προεξέχουσες και τα αντίστοιχα δεδομένα θα λέγονται μη-σημαντικά. Βάσει των προηγούμενων ορισμών τα δεδομένα διαμερίζονται σε δύο σύνολα, όπου το ένα αποτελείται από τα σημαντικά γεγονότα, και το άλλο από τα μη-σημαντικά. Τα δύο σύνολα θα αναφέρονται στο εξής ως *κλάσεις σημαντικότητας (salient classes)*. Η παρούσα εργασία έχει επομένως ως στόχο την αυτόματη διαμέριση των ηχητικών δεδομένων σε κλάσεις σημαντικότητας με χρήση υπολογιστικών μεθόδων.

Για την μοντελοποίηση της ακουστικής σημαντικότητας χρησιμοποιούνται τεχνικές ψηφιακής επεξεργασίας σήματος, ψηφιακής επεξεργασίας εικόνας, αναγνώρισης προτύπων, καθώς και στοιχεία από την μελέτη της ακουστικής αντίληψης του ανθρώπου.

Αρχικά γίνεται υπολογισμός της σημαντικότητας των ηχητικών σκηνών με τεχνικές επεξεργασίας εικόνας. Το ηχητικό σήμα χωρίζεται σε χρονικά παράθυρα της τάξης του ενός δευτερολέπτου και σε κάθε ένα από αυτά υπολογίζεται το φασματογράφημα του (*spectrogram*). Το φασματογράφημα χειρίζεται ως να ήταν εικόνα, και φιλτράρεται με χρήση τριών Gabor φίλτρων, που τονίζουν σημεία όπου υπάρχει υψηλή ενέργεια, μεταβολή κατά μήκος του χρονικού άξονα, και μεταβολή κατά μήκος του συχνοτικού άξονα. Με επεξεργασία του φασματογραφήματος, καταλήγουμε σε μία δισδιάστατη απεικόνιση που ονομάζουμε χάρτη σημαντικότητας (*saliency map*), ο οποίος έχει υψηλές τιμές στα σημεία χρόνου συχνότητας που είναι υποψήφια να έλκουν αυθόρμητα την ανθρώπινη προσοχή. Ακολουθούμε δύο προσεγγίσεις στη συνέχεια, όπου στην πρώτη από τον χάρτη ση-

μαντικότητας καταλήγουμε σε καμπύλη σημαντικότητας και με χρήση κατωφλίου πραγματοποιούμαι ταξινόμηση των δεδομένων. Στην δεύτερη προσέγγιση, εισάγουμε την έννοια του *gist* μιας σκηνής η οποία προέρχεται από μελέτες της οπτικής αντίληψης του ανθρώπου. Σε αυτή την προσέγγιση από τον χάρτη δημιουργούνται *gist* διανύσματα, και πραγματοποιείται ταξινόμηση τους με μεθόδους μηχανικής μάθησης (*machine learning*).

Στην συνέχεια προσαρμόζουμε το μοντέλο του χάρτη σημαντικότητας για τον χειρισμό μονοδιάστατων σημάτων. Αντί για το φασματογράφημα, ως είσοδος δίνονται καμπύλες σε μία διάσταση που έχουν προέλθει από εξαγωγή χαρακτηριστικών από το ηχητικό σήμα. Η εξαγωγή χαρακτηριστικών γίνεται στο πεδίο του χρόνου και της συχνότητας σε πλαίσιο βραχέως χρόνου. Προτείνουμε την χρήση τεσσάρων χαρακτηριστικών. Όλα τα επιμέρους στάδια προσαρμόζονται για τον χειρισμό μονοδιάστατων καμπυλών. Η έξοδος του μοντέλου είναι μια καμπύλη σημαντικότητας για κάθε χαρακτηριστικό που δίνεται ως είσοδος. Για την ταξινόμηση χρησιμοποιούνται και πάλι τεχνικές κατωφλίου και μηχανικής μάθησης. Δείχνουμε ότι το προσαρμοσμένο μοντέλο ξεπερνά σε απόδοση το αρχικό.

Από τα χαρακτηριστικά που εξήχθησαν σε πλαίσιο βραχέως χρόνου γίνεται υπολογισμός ιστογραμμάτων με τρόπο παρόμοιο των μεθόδων *bag of words* που χρησιμοποιούνται στην επεξεργασία φυσικής γλώσσας και πιο πρόσφατα στην όραση υπολογιστών. Πραγματοποιείται ταξινόμηση σε κλάσεις σημαντικότητας με χρήση των ιστογραμμάτων ως διανύσματα και μεθόδων μηχανικής μάθησης. Να σημειωθεί, ότι αυτός ο τρόπος χειρισμού των χαρακτηριστικών δεν έχει χρησιμοποιηθεί ξανά για την αντιμετώπιση του προβλήματος ανίχνευσης σημαντικών γεγονότων και αποτελεί μία από τις καινοτομίες αυτής της εργασίας.

Έπειτα πραγματοποιείται εξαγωγή των ευρέως γνωστών στη βιβλιογραφία αναγνώρισης φωνής, MFCC χαρακτηριστικών, και συγκρίνεται η απόδοση τους με τα υπόλοιπα χαρακτηριστικά. Γίνεται επίσης, υπολογισμός των AM-FM χαρακτηριστικών και ελέγχεται η δυνατότητα τους να διακρίνουν τις δύο κλάσεις σημαντικότητας.

Τέλος, ακολουθείται μια υψηλότερου επιπέδου προσέγγιση. Αντί να πραγματοποιείται ανίχνευση σημαντικών γεγονότων σε ολόκληρο το ηχητικό σήμα, αυτό χωρίζεται σε δύο σύνολα και γίνεται ανίχνευση σε κάθε ένα ξεχωριστά. Στο ένα σύνολο ανήκουν όλα τα σημεία στα οποία υπάρχει ανθρώπινη ομιλία, ενώ στο άλλο δεν υπάρχει ανθρώπινη ομιλία. Με χρήση των χαρακτηριστικών που εξήχθησαν στο υπόλοιπο της εργασίας πραγματοποιείται ταξινόμηση σε κάθε σύνολο ξεχωριστά. Ο διαχωρισμός στα σύνολα φωνής-μη φωνής αποτελεί επίσης μια καινοτομία της εργασίας.

Τα πειράματα πραγματοποιήθηκαν σε μία βάση ηχητικών δεδομένων που αποτελείται από αποσπάσματα ταινιών του Hollywood. Υπήρχαν αποσπάσματα από έξι διαφορετικές ταινίες, συνολικής διάρκειας περίπου 180 λεπτών (3 ωρών). Οι ταινίες περιείχαν ποικιλία ηχητικών δεδομένων που καλύπτουν μεγάλο εύρος ηχητικών ερεθισμάτων με τα οποία έρχεται άνθρωπος καθημερινά σε επαφή. Για αυτά τα δεδομένα υπάρχουν ανθρώπινες επισημειώσεις της σημαντικότητας οι οποίες θα θεωρηθούν ως δεδομένα αλήθειας (ground truth).

1.2 Διάρθρωση της Εργασίας

Στο Κεφάλαιο 2 αρχικά πραγματοποιείται μια σύντομη βιβλιογραφική ανασκόπηση των μεθόδων για ανίχνευση σημαντικών γεγονότων. Στη συνέχεια παρουσιάζεται ένα υπολογιστικό μοντέλο το οποίο εφαρμόζει τεχνικές της όρασης υπολογιστών για την ανίχνευση προεξεχόντων ηχητικών γεγονότων. Ελέγχεται η δυνατότητα του μοντέλου να ανιχνεύει σημαντικά γεγονότα. Περιγράφεται ένα υπολογιστικό μοντέλο το οποίο προσεγγίζει τον τρόπο που πραγματοποιείται επεξεργασία της ηχητικής πληροφορίας στο ακουστικό σύστημα του ανθρώπου, και δίνονται παραδείγματα από την έξοδο του. Τέλος, παρουσιάζονται υπάρχουσες μέθοδοι αξιολόγησης της σημαντικότητας και η μέθοδος που ακολουθείται σε αυτή την εργασία.

Στο Κεφάλαιο 3 πραγματοποιείται συνδυασμός των μοντέλων του προηγούμενου κεφαλαίου για ανίχνευση σημαντικών γεγονότων σε ηχητικά ερεθίσματα από ταινίες. Δοκιμάζεται ταξινόμηση των ηχητικών ερεθισμάτων με χρήση κατωφλίου και μεθόδων μηχανικής μάθησης.

Στο Κεφάλαιο 4 αναπτύσσεται μοντέλο το οποίο προσαρμόζει το μοντέλο ανίχνευσης σημαντικών γεγονότων της ενότητας 2 σε μία διάσταση. Η είσοδος και τα επιμέρους στάδια του μοντέλου προσαρμόζονται κατάλληλα για τον χειρισμό μονοδιάστατων σημάτων, και δείχνονται παραδείγματα εξόδου από κάθε στάδιο. Επίσης, περιγράφονται τα χαρακτηριστικά που δίνονται ως είσοδος στο μοντέλο, και υπολογίζεται η μεταξύ τους συσχέτιση.

Στο Κεφάλαιο 5 γίνεται έλεγχος του μοντέλου του προηγούμενου κεφαλαίου για την ανίχνευση σημαντικών γεγονότων. Αρχικά ελέγχεται μέσω ενός πειράματος σύγκρισης σημαντικότητας σκηνών και έπειτα στην ταξινόμηση δεδομένων σε κλάσεις σημαντικότητας. Πραγματοποιείται εξαγωγή ιστογραμμάτων από τα χαρακτηριστικά.

Στο Κεφάλαιο 6 γίνεται εξαγωγή καθιερωμένων χαρακτηριστικών της βιβλιογραφίας και πραγματοποιείται σύγκριση της απόδοσης τους με τα υπόλοιπα. Επίσης, πραγματοποιείται διαχωρισμός του σήματος σε σύνολα φωνής και μη-φωνής, και ταξινόμηση σε κάθε ένα ξεχωριστά.

Τέλος στο Κεφάλαιο 7 συνοψίζεται το πρόβλημα που εξετάσθηκε και προτείνονται μελλοντικές κατευθύνσεις για έρευνα, και επέκταση αυτής της εργασίας.

Κεφάλαιο 2

Βιβλιογραφική Ανασκόπηση

Η ανάπτυξη μοντέλων που ανιχνεύουν αυτόματα σημαντικούς ήχους αλλά και γενικότερα που θα κατευθυνθεί η ανθρώπινη προσοχή είναι αντικείμενο έρευνας την τελευταία δεκαετία και δεν έχει σημειώσει ιδιαίτερη πρόοδο. Τα μοντέλα που έχουν αναπτυχθεί είναι λιγοστά και έχουν αρκετές ομοιότητες μεταξύ τους. Στηρίζονται κυρίως στην χρονοσυχνοτική αναπαράσταση του φασματογραφήματος ενός ηχητικού σήματος το οποίο και επεξεργάζονται. Στην επόμενη ενότητα γίνεται μια συνοπτική περιγραφή των υπαρχόντων υπολογιστικών μοντέλων.

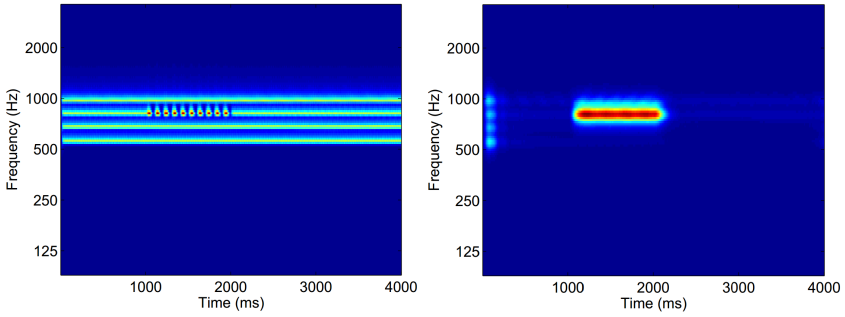
2.1 Υπάρχοντα Υπολογιστικά Μοντέλα

Ένα από τα πρώτα μοντέλα που εμφανίσθηκε για την ανίχνευση σημαντικών ηχητικών γεγονότων είναι το μοντέλο των Kayser et al [31]. Αναλυτική περιγραφή του μοντέλου πραγματοποιείται στην ενότητα 2 και δεν θα αναλυθεί εδώ.

Το μοντέλο των Kalinli και Narayanan [29] αναπτύχθηκε για την ανίχνευση προεξεχουσών συλλαβών σε ομιλία στην αγγλική γλώσσα (*prominent syllables detection*). Το μοντέλο περιλαμβάνει ένα bottom-up μέρος στο οποίο πραγματοποιείται εξαγωγή χαρακτηριστικών από την ηχητική σκηνή, και ένα top-down μέρος όπου χρησιμοποιείται λεκτική και συντακτική πληροφορία για την διαμόρφωση (*modulation*) των ακουστικών χαρακτηριστικών.

Η εξαγωγή ακουστικών χαρακτηριστικών είναι παρόμοια με αυτή του μοντέλου των Kayser et al [31]. Αρχικά υπολογίζεται το ακουστικό φάσμα[1] ενός ηχητικού σήματος, το οποίο είναι παρόμοιο με το φασματογράφημα και δειγματοληπτείται σε διάφορες κλίμακες. Κάθε κλίμακα φιλτράρεται με φίλτρα έντασης, χρονικής και συχνοτικής αντίθεσης, και αντίθεσης με προσανατολισμό 45° και 135° , για την ανίχνευση σταδιακών αυξήσεων και μειώσεων της συχνότητας του σήματος (*chirps*) για τα οποία υπάρχουν ενδείξεις ότι αποτελούν σημαντικούς ήχους σε αναλογία με τις ακμές μιας εικόνας με αντίστοιχους προσανατολισμούς [35]. Επίσης με βάση την υπόθεση του χρονικού pitch [58] εξάγεται πληροφορία για την μεταβολή του.

[1] Αναλυτική περιγραφή του ακουστικού φάσματος πραγματοποιείται στην ενότητα 2



(α΄) (β΄)

Σχήμα 2.1: Αριστερά: φασματογράφημα συστοιχίας αρμονικών όπου μία αρμονική διαμορφώνεται κατά πλάτος για κάποια χρονική διάρκεια. Δεξιά: χάρτης σημαντικότητας όπως υπολογίστηκε από το μοντέλο στο [15].

Ακολουθούν διαφορές κέντρου-περίγυρου για κάθε χαρακτηριστικό. Τέλος, κάθε χάρτης διαμερίζεται με τετραγωνικό πλέγμα και σε κάθε κελί του πλέγματος υπολογίζεται ο μέσος όρος των τιμών του δημιουργώντας ένα διάνυσμα χαρακτηριστικών για κάθε χάρτη, τα οποία έπειτα συνενώνονται σε ένα ενιαίο διάνυσμα. Από την εφαρμογή PCA στην συνένωση των διανυσμάτων προκύπτει το τελικό διάνυσμα χαρακτηριστικών.

Λεκτική πληροφορία ενσωματώνεται χρησιμοποιώντας ένα πιθανοτικό γλωσσικό μοντέλο με n-grams, και συντακτική κάνοντας χρήση τι μέρος του λόγου (*part of speech*) είναι κάθε λέξη. Η σημαντικότητα κάθε ακολουθίας συλλαβών επιλέγεται έτσι ώστε να μεγιστοποιείται η πιθανότητα εμφάνισης της, δοθείσας ακουστικής, λεκτικής και συντακτικής πληροφορίας. Το μοντέλο έδωσε ποσοστά ανίχνευσης γύρω στο 80% σε μια βάση με δεδομένα από ραδιοφωνικές εκπομπές cite..buradio, όπου η απόδοση σημαντικότητας σε κάθε συλλαβή καθορίζεται από την προφορά του pitch [49] για την συλλαβή όπως σημειώθηκε από ανθρώπους.

Το μοντέλο της Duangudom [15] εξάγει και αυτό χάρτη σημαντικότητας από το ακουστικό φάσμα. Το ακουστικό φάσμα φιλτράρεται με μία συστοιχία από 2-Δ Gabor φίλτρα για την ανίχνευση των χρονικών και φασματικών διακυμάνσεων (*modulations*). Με βάση την συχνότητα των Gabor φίλτρων στον χρονικό και συχνοτικό άξονα λαμβάνονται χάρτες που περιέχουν πληροφορία για την συνολική κατανομή ενέργειας, τις χρονικές διακυμάνσεις, τις συχνοτικές διακυμάνσεις, και τις χρονο-συχνοτικές διακυμάνσεις. Ακολουθεί ένα στάδιο κανονικοποίησης των χαρτών όμοιο με αυτό του Kayser. Οι χάρτες σε κάθε κατηγορία συνδυάζονται γραμμικά και στο αποτέλεσμα πραγματοποιείται και πάλι η ίδια κανονικοποίηση. Ο τελικός χάρτης σημαντικότητας προκύπτει από το άθροισμα των επιμέρους χαρτών.

Το μοντέλο έχει επιβεβαιωθεί ότι έχει συσχέτιση με ψυχο-ακουστικές παρατηρήσεις για την ικανότητα ήχων να έλκουν την ανθρώπινη προσοχή. Ένα παράδειγμα φαίνεται στο Σχήμα 2.1, όπου ένας διαμορφωμένος κατά πλάτος τόνος ξεχωρίζει ανάμεσα σε σταθερούς τόνους, το οποίο έχει παρατηρηθεί και σε πειράματα με ανθρώπους.

Το μοντέλο των Kaya και Elhilali [30] εξάγει χαρακτηριστικά από το ακουστικό φά-

σμα αλλά και από το ηχητικό σήμα άμεσα. Τα χαρακτηριστικά που εξάγονται από το ηχητικό σήμα είναι η περιβάλλουσα, για την παρακολούθηση της έντασης και του ηχοχρώματος (*tembre*), και το *pitch* για την ανίχνευση μεταβολών στην βασική συχνότητα του σήματος. Από το ακουστικό φάσμα εξάγονται ο ρυθμός (*rate*), το εύρος ζώνης (*bandwidth*), και το ακουστικό φάσμα (ταυτοτικός μετασχηματισμός). Όλα τα χαρακτηριστικά εξάγονται σε πολλαπλές κλίμακες. Ακολουθούν διαφορές κέντρου-περίγυρου μεταξύ των κλιμάκων και κλιμάκωση ώστε η μέγιστη τιμή του αθροίσματος κατά μήκος του συχνοτικού άξονα να είναι 1. Έπεται επαναληπτική μη-γραμμική κανονικοποίηση κάθε χάρτη ως εξής:

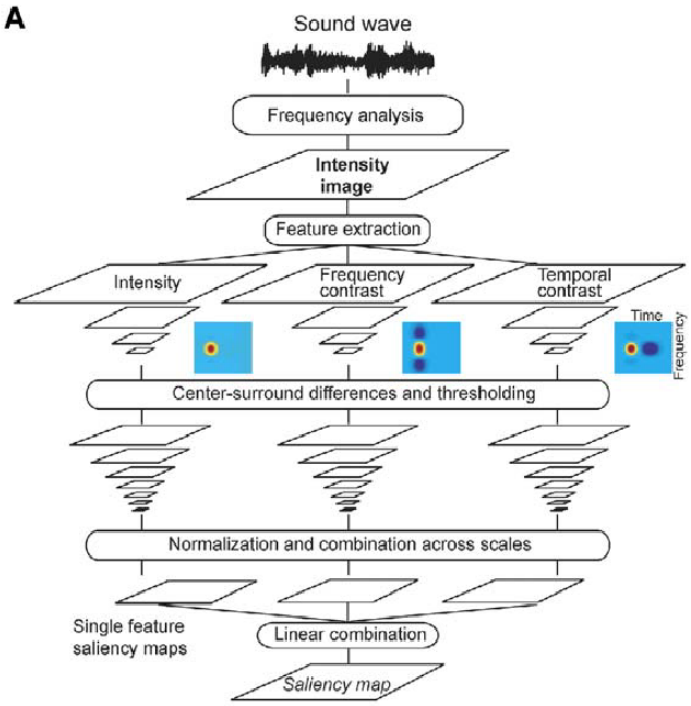
$$M \leftarrow |M + M * DoG - C_{inh}| \tag{2.1}$$

όπου M ο χάρτης σε κάθε επανάληψη, DoG μονοδιάστατη διαφορά από Gaussians και C_{inh} ο χάρτης στην πρώτη επανάληψη σταθμισμένος με μικρή σταθερά. Τέλος οι χάρτες για κάθε χαρακτηριστικό αθροίζονται κατά μήκος των κλιμάκων. Επίσης, οι δισδιάστατοι χάρτες απεικονίζονται σε μία διάσταση λαμβάνοντας την μέση τιμή κατά μήκος του συχνοτικού άξονα. Έτσι για κάθε χαρακτηριστικό προκύπτει μία καμπύλη σημαντικότητας. Η τελική καμπύλη σημαντικότητας υπολογίζεται από την μέση τιμή των επιμέρους καμπυλών.

Για τον έλεγχο του μοντέλου χρησιμοποιήθηκαν τόνοι από μουσικά όργανα. Μεταβάλλοντας κάθε φορά μόνο ένα από τα ηχόχρωμα, pitch και ένταση για μικρό χρονικό διάστημα χρησιμοποιώντας διαφορετικό κύριο όργανο, ελέγχονταν εάν το μοντέλο ανίχνευε την μεταβολή. Παρατηρήθηκε ότι στην πλειοψηφία των περιπτώσεων η καμπύλη σημαντικότητας είχε την μέγιστη τιμή στο σημείο της μεταβολής ή τιμή κοντά στην μέγιστη στο σημείο αυτό.

Το μοντέλο των Tsuchida και Cottrell [62] συνδυάζει το μοντέλο των Kayser et al [31] με το μοντέλο των Zhang et al [71] για ανίχνευση προεξεχόντων γεγονότων σε εικόνες. Στηρίζεται στην ιδέα ότι τα σημαντικά γεγονότα έχουν μικρή πιθανότητα εμφάνισης. Αρχικά το σήμα φιλτράρεται με συστοιχία από Gammatone φίλτρα για την δημιουργία διδιάστατης απεικόνισης. Έπειτα, ο συχνοτικός άξονας διαμερίζεται σε μπάντες διάστασης επτά και ο χρονικός άξονας σε παράθυρα μήκους οκτώ δημιουργώντας χρονο-συχνοτικές περιοχές στην απεικόνιση, και κάθε μία λαμβάνεται ως ένα διάνυσμα 56 διαστάσεων. Εφαρμόζοντας PCA σε κάθε μπάντα ξεχωριστά διατηρούνται δύο ή τρεις διαστάσεις στα διανύσματα. Ένα GMM εκπαιδεύεται για την εκτίμηση της πιθανότητας εμφάνισης των διανυσμάτων χαρακτηριστικών. Η σημαντικότητα κάθε διανύσματος ορίζεται ως ο αρνητικός λογάριθμος της πιθανότητας εμφάνισης του. Από την σημαντικότητα των διανυσμάτων προκύπτει άμεσα και η σημαντικότητα της δισδιάστατης σκηνής που συντίθεται από αυτά.

Το μοντέλο ελέγχθηκε σε πείραμα σύγκρισης σημαντικότητας των σκηνών, παρόμοιο με αυτό του Kayser, καθώς και σε γνωστά ψυχο-ακουστικά φαινόμενα [12]. Και στις δύο περιπτώσεις παρατηρήθηκε συσχέτιση της εξόδου του με τα πειραματικά δεδομένα.



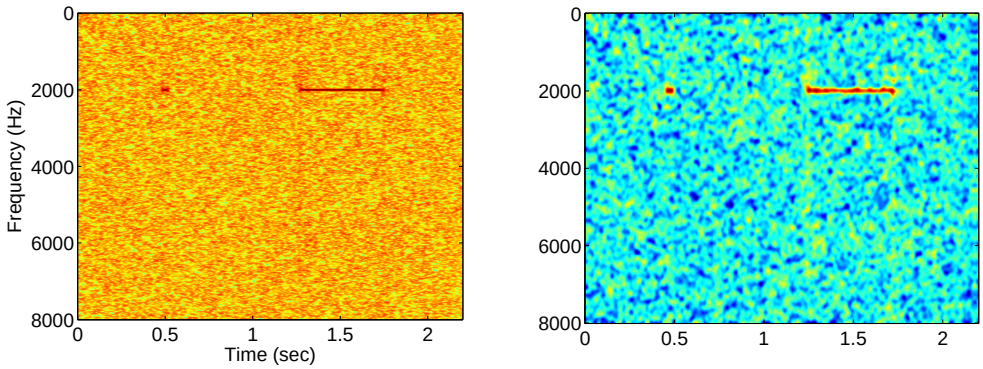
Σχήμα 2.2: Διάγραμμα με τα βασικά στάδια του αλγορίθμου εξαγωγής του χάρτη σημαντικότητας (Saliency Map), όπως λήφθηκε από το [31].

2.2 Μοντέλο των Kayser et al

Το μοντέλο των Kayser et al [31], βασίζεται στο μοντέλο των Itti και Koch που αναπτύχθηκε για την ανίχνευση οπτικά σημαντικών αντικειμένων (*visually salient events*) σε εικόνες [25, 27, 24] και στηρίζεται στην θεωρία ολοκλήρωσης χαρακτηριστικών (*feature integration*) της A. Treismann [61]. Οι Kayser et al χρησιμοποιούν ως εικόνα το φασματογράφημα (*spectrogram*) του σήματος και εφαρμόζουν σε αυτή μεθόδους της Όρασης Υπολογιστών για την επεξεργασία της, και ανίχνευση ακουστικά σημαντικών σημείων. Σημαντικά σημεία στο φασματογράφημα θεωρούνται εκείνα που διαφέρουν από το περιβάλλον τους, καθώς εκεί παρατηρείται κάποια μεταβολή στις ιδιότητες του σήματος που είναι πιθανώς ικανή να τραβήξει την ανθρώπινη προσοχή. Ένα παράδειγμα σημαντικού σημείου στο φασματογράφημα είναι η εμφάνιση μιας οριζόντιας μπάρας, όπως φαίνεται στο Σχήμα 2.3α΄, που οφείλεται στην εμφάνιση κάποιου τόνου στην συγκεκριμένη περιοχή συχνοτήτων.

2.2.1 Περιγραφή μοντέλου

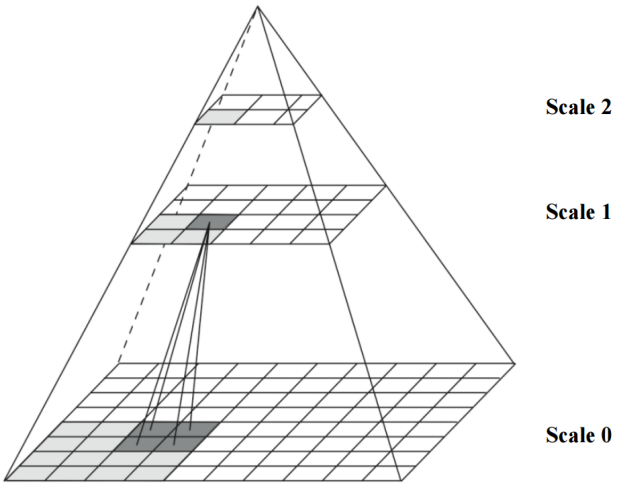
Στο Σχήμα 2.2 φαίνονται τα βασικά στάδια του αλγορίθμου για την εξαγωγή μιας δισδιάστατης εικόνας από το ηχητικό σήμα, που είναι η έξοδος του μοντέλου και οι συγγρα-



(α΄) (β΄)

(γ΄) (δ΄)

Σχήμα 2.3: Από αριστερά προς τα δεξιά και πάνω προς τα κάτω: φασματογράφημα δύο τόνων σε περιβάλλον λευκού θορύβου, και εικόνες έπειτα από φιλτράρισμα με φίλτρο έντασης, φίλτρο χρονικής αντίθεσης, φίλτρο συχνοτικής αντίθεσης.



Σχήμα 2.4: Μέθοδος λήψης πολλαπλών κλιμάκων από εικόνες.

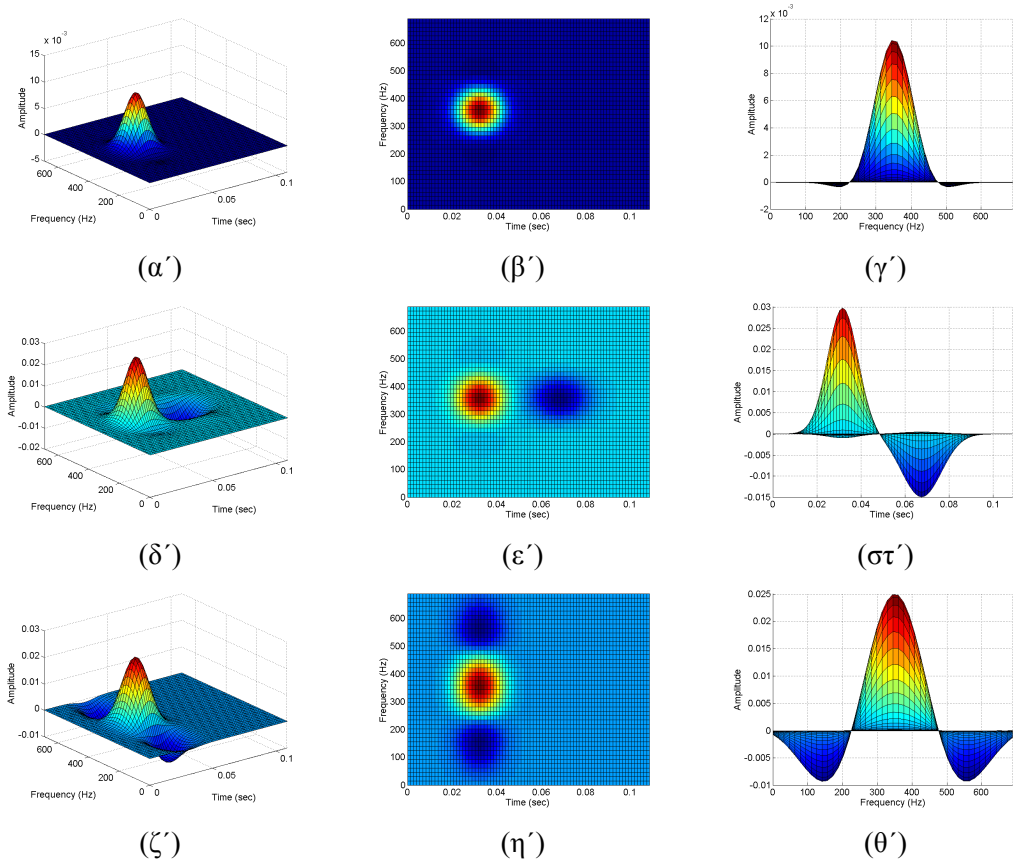
φείς ονομάζουν χάρτη σημαντικότητας (*saliency map*). Πιο συγκεκριμένα, σε πρώτο στάδιο υπολογίζεται από το ηχητικό σήμα ο λογάριθμος του φασματογραφήματος του, που θεωρείται η εικόνα προς ανάλυση. Αυτό στην συνέχεια υπο - δειγματοληπτείται ομοιόμορφα με παράγοντα $1/2$ σε πέντε κλίμακες, από $1/2^0$ έως $1/2^4$. Η εικόνα στην κλίμακα τάξης n προκύπτει από την κλίμακα τάξης $n-1$ κρατώντας κάθε δεύτερο pixel στις γραμμές και στήλες της εικόνας, και εικόνα στην κλίμακα τάξης μηδέν θεωρείται το αρχικό φασματογράφημα (Σχήμα 2.4). Στην συνέχεια κάθε εικόνα φιλτράρεται με τρία διαφορετικά φίλτρα για την εξαγωγή διαφορετικών χαρακτηριστικών. Τα φίλτρα αυτά φαίνονται στο Σχήμα 2.5 και παράγονται με χρήση Gabor φίλτρου, G, του οποίου η εξίσωση έχει την ακόλουθη μορφή:

$$G(t,f) = \exp\left(-\frac{1}{2}\left(\left(\frac{t-t_0}{Dur}\right)^2 + \left(\frac{f-f_0}{BW}\right)^2\right)\right)\cos(2\pi \cdot freq \cdot f) \tag{2.2}$$

(t_0, f_0) είναι το σημείο στο οποίο βρίσκεται το κέντρο του στο επίπεδο χρόνου - συχνότητας, (Dur, BW) τυπικές αποκλίσεις, και $freq$ η συχνότητα ταλάντωσης της περιβάλλουσας κατά μήκος του συχνοτικού άξονα.

Το φίλτρο της έντασης (*intensity filter*) είναι ένα Gabor φίλτρο της μορφής 2.2, στο οποίο η συχνότητα του συνημιτόνου είναι μικρή σχετικά με την διασπορά της Gaussian, και δεν παρατηρείται φάση αποκοπής (*inhibition phase*) με μεγάλο πλάτος. Όταν μια εικόνα φιλτράρεται με αυτό το φίλτρο, ενισχύονται περιοχές με την πιο έντονη φωτεινότητα στην εικόνα. Σημεία με μεγάλη φωτεινότητα στο φασματογράφημα αντιστοιχούν σε ήχους με μεγάλη ένταση οι οποίοι πιθανώς κυριαρχούν των υπόλοιπων ήχων στην ακουστική σκηνή, και είναι αυτοί που γίνονται αντιληπτοί.

Το φίλτρο της χρονικής αντίθεσης (*temporal contrast filter*), είναι η διαφορά δύο Gabor φίλτρων της μορφής 2.2, όπου το ένα έχει το μισό πλάτος του άλλου και έχουν



Σχήμα 2.5: Φίλτρα έντασης 2.5α΄-2.5γ΄, χρονικής αντίθεσης 2.5δ΄-2.5στ΄, και συχνοτικής αντίθεσης 2.5ζ΄-2.5θ΄.

διαφορετικά κέντρα στον χρονικό άξονα. Αυτό το φίλτρο ενισχύει σημεία στα οποία παρατηρείται μεταβολή κατά μήκος του χρόνου. Τέλος, το φίλτρο της συχνοτικής αντίθεσης (*frequency contrast*), είναι ένα Gabor φίλτρο με εξίσωση ίδιας μορφής με το φίλτρο της έντασης, αλλά η διακύμανση της Gaussian κατά μήκος του άξονα των συχνοτήτων είναι μεγαλύτερη του φίλτρου της έντασης, με αποτέλεσμα να υπάρχει έντονη φάση αποκοπής. Αυτό το φίλτρο ενισχύει περιοχές που παρατηρείται μεταβολή κατά μήκος του άξονα των συχνοτήτων. Τα δύο τελευταία φίλτρα χρησιμοποιούνται διότι έχει παρατηρηθεί πως οι έντονες μεταβολές σε αρκετές περιπτώσεις τραβούν την ανθρώπινη προσοχή, ωστόσο αυτό δεν ισχύει πάντοτε όπως θα δειχθεί αργότερα. Το φίλτρο της έντασης είναι ανάλογο του φίλτρου της φωτεινότητας (*luminance*) που χρησιμοποιείται στην Όραση Υπολογιστών, και τα φίλτρα της χρονικής και συχνοτικής αντίθεσης είναι ανάλογα των φίλτρων ανίχνευσης κάθετων και οριζόντιων ακμών (*edges*), αντίστοιχα. Από αυτό το στάδιο προκύπτουν 3 χάρτες χαρακτηριστικών (*feature maps*), για κάθε εικόνα σε κάθε κλίμακα, και συνεπώς 15 χάρτες χαρακτηριστικών συνολικά.

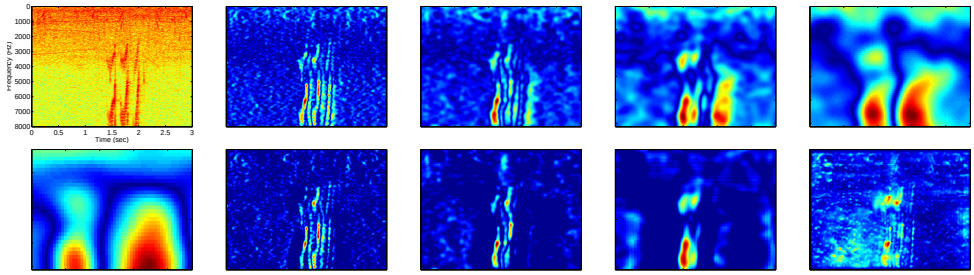
Στο Σχήμα 2.3 φαίνεται το αποτέλεσμα της εφαρμογής των φίλτρων στο φασματογράφημα δύο τόνων σε περιβάλλον λευκού θορύβου. Οι περιοχές εμφάνισης των τόνων ενισχύονται από το φίλτρο έντασης. Το φίλτρο χρονικής αντίθεσης έχει τονίσει την αρχή (*onset*) και το τέλος (*offset*) των τόνων, όπου υπάρχει μεταβολή κατά μήκος του χρονικού άξονα και είναι πιθανά σημεία έλξης της προσοχής. Τέλος, το φίλτρο συχνοτικής αντίθεσης έχει ενισχύει την περιοχή εμφάνισης των τόνων λόγω ύπαρξης μεταβολής κατά μήκος του συχνοτικού άξονα.

Το επόμενο στάδιο είναι αυτό που θα αναδείξει σημεία τα οποία είναι υποψήφια να θεωρηθούν σημαντικά. Λαμβάνονται διαφορές μεταξύ κλιμάκων του ίδιου χαρακτηριστικού ώστε να τονιστούν σημεία τα οποία διαφέρουν από τα γειτονικά τους. Οι διαφορές κλιμάκων καλούνται διαφορές κέντρου-περίγυρου (*center-surround differences*), καθώς από την μικρότερη κλίμακα αφαιρείται η μεγαλύτερη, και από κάθε σημείο αφαιρούνται αυτά τα οποία βρίσκονται γύρω του (Σχήμα 2.4). Η λήψη διαφορών πραγματοποιείται επανα-δειγματοληπτώντας όλες τις εικόνες ώστε να έχουν ίσες διαστάσεις σε pixel, και λαμβάνοντας διαφορές σημείο-προς-σημείο, μεταξύ κλιμάκων που απέχουν κατά $\{2,\ 3\}$. Από την μικρότερη κλίμακα αφαιρείται η μεγαλύτερη, ώστε από κάθε σημείο να αφαιρεθούν τα γειτονικά του. Οι τιμές διαφοράς που δίνουν αρνητικό πρόσημο τίθενται στο μηδέν. Η διαδικασία λήψης διαφορών μιμείται ιδιότητες της τοπικής φλοιώδους απαγόρευσης (*local cortical inhibition*) [54]. Στο Σχήμα 2.6 φαίνονται κάποια παραδείγματα έπειτα από την εφαρμογή των διαφορών σε χάρτες χαρακτηριστικών.

Το τελευταίο στάδιο της επεξεργασίας είναι η κανονικοποίηση των χαρτών. Αυτό το στάδιο έχει στόχο να ενισχύσει χάρτες που έχουν λίγες και υψηλές κορυφές, και να καταπιέσει αυτούς που είναι ομοιόμορφοι. Η κανονικοποίηση πραγματοποιείται πολλαπλασιάζοντας κάθε χάρτη με το τετράγωνο της διαφοράς του ολικού μεγίστου από την μέση τιμή των υπόλοιπων τοπικών μεγίστων του χάρτη. Δηλαδή, εάν ο χάρτης C_i, έχει ολικό μέγιστο M_i, και η μέση τιμή των υπόλοιπων τοπικών μεγίστων είναι $\bar{m}_i$, τότε:

$$C_i^{norm} = (M_i - \bar{m}_i)^2 \cdot C_i \tag{2.3}$$

όπου C_i^{norm}, ο πίνακας μετά την κανονικοποίηση. Οι χάρτες με λίγα και υψηλά μέγιστα



Σχήμα 2.6: Ανάλυση σε αρχείο ήχου με το κελάηδισμα ενός πουλιού. Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω φαίνονται: ο λογάριθμος του φασματογραφήματος, φιλτράρισμα με φίλτρο χρονικής αντίθεσης για τις κλίμακες 0-4 και οι διαφορές 0-2, 1-3, 2-4, και τέλος ο χάρτης σημαντικότητας.

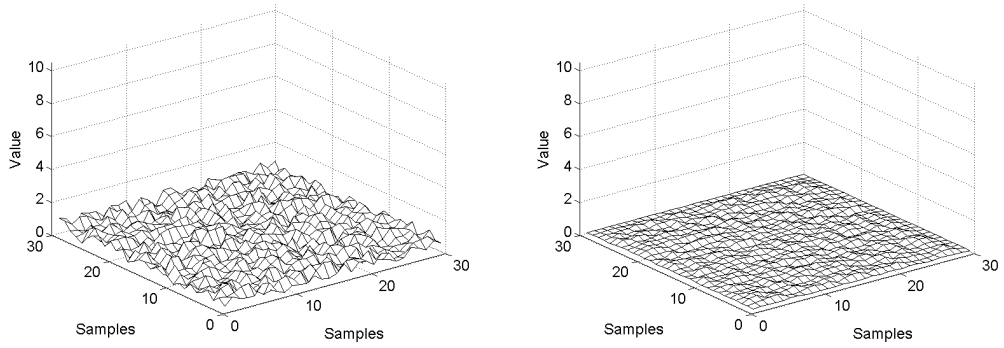
ενισχύονται, ενώ οι χάρτες με πολλές σχεδόν ισοϋψείς κορυφές καταπιέζονται. Στο Σχήμα 2.7 φαίνεται ένας χάρτης που έχει παραχθεί από μια ομοιόμορφη κατανομή, και δεξιά το αποτέλεσμα της κανονικοποίησης του. Στο Σχήμα 2.8 φαίνεται ο ίδιος χάρτης όπου τώρα ενισχύσαμε μια κορυφή ώστε να έχει διπλάσιο ύψος. Βλέπουμε πως ο χάρτης ενισχύεται από το στάδιο της κανονικοποίησης όταν κάποια κορυφή προεξέχει, διαφορετικά καταπιέζεται.

Η κανονικοποίηση γίνεται είτε ολικά (globally) σε ολόκληρο τον χάρτη, είτε τοπικά (locally) σε χρονικά υποσύνολα του. Στην περίπτωση οπτικών ερεθισμάτων συνηθίζεται να γίνεται ολικά, καθώς ερχόμαστε σε επαφή με ολόκληρο το ερέθισμα ταυτόχρονα. Αντίθετα, τα ακουστικά ερεθίσματα παρουσιάζονται σειριακά στον χρόνο και υπάρχει αλληλεπίδραση κυρίως μεταξύ γεγονότων που δεν απέχουν πολύ χρονικά. Επομένως, όταν η ακουστική σκηνή προς ανάλυση έχει διάρκεια μερικών δευτερολέπτων, η κανονικοποίηση γίνεται συνήθως τοπικά σε χρονικά παράθυρα μήκους περίπου 500 ms. Τα παράθυρα είναι επικαλυπτόμενα, με $100 - 150$ ms επικάλυψη μεταξύ διαδοχικών παραθύρων, ώστε να είναι ομαλή η μετάβαση από το ένα στο άλλο και ο χάρτης να μην έχει μεγάλες ασυνέχειες.

Μετά το στάδιο της κανονικοποίησης λαμβάνεται η μέση τιμή των χαρτών σε κάθε χαρακτηριστικό και εξάγεται ο χάρτης σημαντικότητας για το χαρακτηριστικό αυτό. Εάν κάποιο χαρακτηριστικό ήταν έντονο στο αρχικό φασματογράφημα, ο αντίστοιχος χάρτης σημαντικότητας θα έχει κάποιες υψηλές κορυφές, διαφορετικά θα είναι ομοιόμορφος. Ο τελικός χάρτης σημαντικότητας προκύπτει από τον γραμμικό συνδυασμό των χαρτών σημαντικότητας για κάθε χαρακτηριστικό, σε αναλογία με την ιδέα ολοκλήρωσης χαρακτηριστικών σε εικόνες [61].

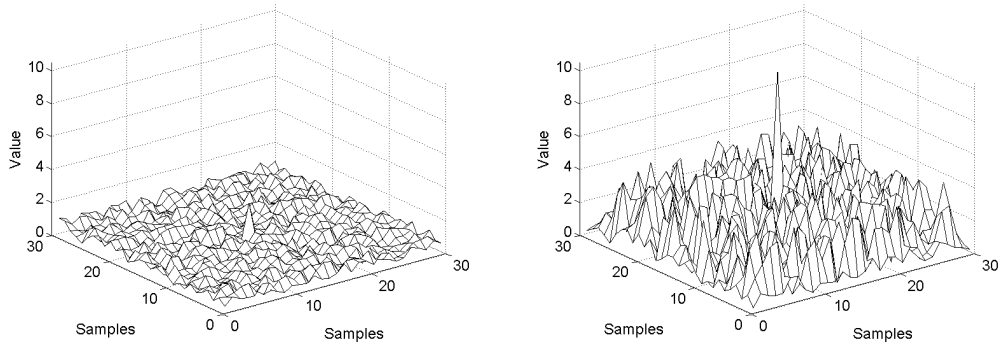
2.2.2 Αξιολόγηση μοντέλου των Kayser et al

Το μοντέλο έχει επιβεβαιωθεί ότι αναπαράγει κάποια γνωστά φαινόμενα της ανθρώπινης ακουστικής αντίληψης όπως έχουν μελετηθεί στην ψυχοακουστική βιβλιογραφία [12]. Αυτά περιλαμβάνουν τις ακόλουθες κατηγορίες ήχων σε θορυβώδες περιβάλλον: οι



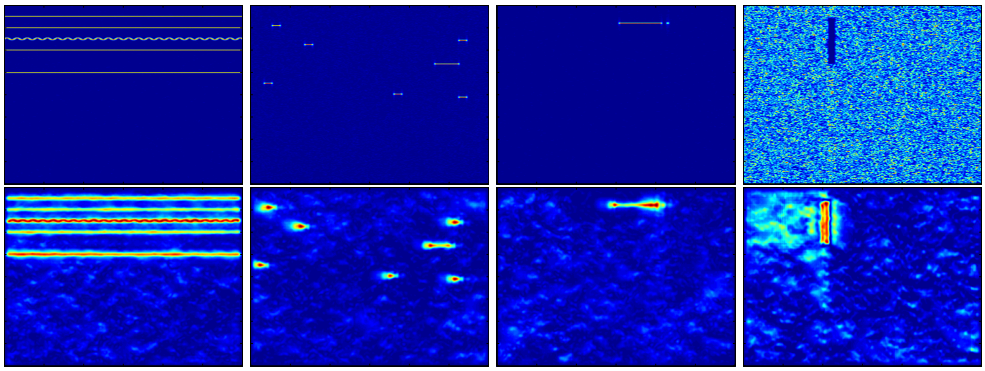
(α΄) (β΄)

Σχήμα 2.7: 2.7α΄ χάρτης που παράχθηκε από ομοιόμορφη κατανομή στο $[0,\ 1]$, 2.7β΄ ο χάρτης μετά την κανονικοποίηση.



(α΄) (β΄)

Σχήμα 2.8: 2.8α΄ χάρτης από ομοιόμορφη κατανομή με ενισχυμένη μία κορυφή του, 2.8β΄ ο χάρτης μετά την κανονικοποίηση.



(α΄) (β΄) (γ΄) (δ΄)

Σχήμα 2.9: Στην πρώτη γραμμή τα φασματογραφήματα και στην δεύτερη οι αντίστοιχοι χάρτες σημαντικότητας των: 2.9α΄ διαμορφωμένων και σταθερών τόνων, 2.9β΄ μακριών και κοντών τόνων, 2.9γ΄ κοντά τοποθετημένων τόνων, 2.9δ΄ συχνοτικού κενού.

μεγάλης διάρκειας τόνοι είναι πιο σημαντικοί από τους τόνους μικρής διάρκειας, τόσο οι μεγάλης όσο και οι μικρής διάρκειας τόνοι ξεχωρίζουν σε θορυβώδες περιβάλλον, οι διαμορφωμένοι τόνοι ξεχωρίζουν ευκολότερα από τους σταθερούς, εάν από θόρυβο μεγάλου συχνοτικού εύρους (*broadband noise*) αφαιρεθεί κάποια χρονική στιγμή ένα εύρος συχνοτήτων, τότε αυτό γίνεται εύκολα αντιληπτό. Τέλος, σε μια ακολουθία δύο κοντά τοποθετημένων στον χρόνο τόνων, ο δεύτερος τόνος είναι λιγότερο εύκολα αντιληπτός από τον πρώτο.

Δημιουργήσαμε ηχητικές σκηνές που αναπαράγουν τις ανωτέρω κατηγορίες ήχων και ελέγξαμε την απόδοση του μοντέλου σε αυτές. Στο Σχήμα 2.9 φαίνονται τα φασματογραφήματα και οι χάρτες σημαντικότητας τέτοιων ηχητικών σκηνών. Βλέπουμε πως με εξαίρεση την περίπτωση συνύπαρξης μακρών και κοντών τόνων, όπου αποτυγχάνει να δώσει μεγαλύτερη σημαντικότητα στον μακρύ τόνο, οι προβλέψεις του μοντέλου συμφωνούν με τις ψυχοακουστικές προβλέψεις.

Ανθρώπινη βαθμολόγηση της σημαντικότητας

Ένας τρόπος ελέγχου της δυνατότητας του μοντέλου να θέτει στάθμες σημαντικότητας είναι μέσω της σύγκρισης δύο σκηνών. Οι συγγραφείς ζήτησαν από άτομα να συγκρίνουν ζεύγη ηχητικών σκηνών, και να επιλέξουν μία ως περισσότερο σημαντική, ή να τις θεωρήσουν εξίσου σημαντικές. Η ίδια ερώτηση τέθηκε και στο μοντέλο, όπου ως ποιο σημαντική σκηνή επιλέγεται εκείνη που δίνει υψηλότερες κορυφές στον χάρτη σημαντικότητας. Οι ηχητικές σκηνές είχαν διάρκεια τριών δευτερολέπτων και σε περιβάλλον θορύβου περιείχαν ήχους της φύσης, ήχους μηχανών και ήχους ανθρώπινης ομιλίας. Τα αποτέλεσμα του πειράματος έδειξαν ότι ανθρώπινη επιλογή ήταν υψηλά συσχετισμένη με την επιλογή του μοντέλου, δίνοντας μέση συσχέτιση (*correlation*) 0.47 ± 0.1(μέση τιμή

± τυπική απόκλιση) για κάθε άτομο. Επίσης, όταν τα άτομα θεωρούσαν τις σκηνές εξίσου σημαντικές, η διαφορά σημαντικότητας που προέβλεπε το μοντέλο ήταν κοντά στο μηδέν.

Πείραμα ανίχνευσης με ανθρώπους

Ένας άλλος τρόπος ελέγχου εάν η πρόβλεψη του μοντέλου συμφωνεί με την ανθρώπινη αντίληψη, είναι η τοποθέτηση ηχητικών σκηνών των οποίων την σημαντικότητα πρέπει να ελεγχθεί, σε θορυβώδες περιβάλλον του οποίου την ένταση μεταβάλλουμε. Σε πρώτο στάδιο, ηχητικές σκηνές ταξινομήθηκαν από το μοντέλο ως περισσότερο ή λιγότερο σημαντικές. Στην συνέχεια τοποθετήθηκαν σε αρχείο ήχου μεγαλύτερης διάρκειας που υπήρχε θόρυβος, και παρουσιάστηκαν σε άτομα μέσω ακουστικών. Η ηχητική σκηνή εμφανιζόταν μόνο σε ένα από τα δύο ακουστικά και ζητήθηκε από τα άτομα να απαντήσουν εάν ανίχνευσαν την ηχητική σκηνή ή όχι. Για τις σκηνές που το μοντέλο ταξινόμησε ως περισσότερο σημαντικές, το ποσοστό ανίχνευσης τους από τα άτομα ήταν μεγαλύτερο σε σύγκριση με τις σκηνές που ταξινομήθηκαν ως λιγότερο σημαντικές.

Πείραμα ανίχνευσης με μαϊμούδες

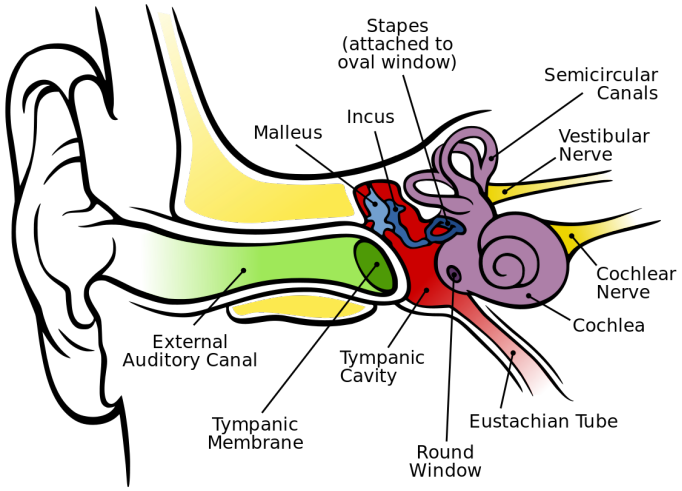
Οι ηχητικές σκηνές που παρουσιάστηκαν στους ανθρώπους παρουσιάστηκαν σε μαϊμούδες. Δύο μεγάφωνα τοποθετήθηκαν πίσω από το ζώο σε αντίθετες κατευθύνσεις μεταξύ τους. Στα μεγάφωνα υπήρχε συνεχώς θόρυβος και κάποια χρονική στιγμή σε ένα από τα δύο παρουσιάζονταν μία ηχητική σκηνή. Οι ηχητικές σκηνές παρουσιάζονταν όταν το ζώο δεν είχε τάση προσανατολισμού προς κάποιο μεγάφωνο. Παρατηρητές έλεγχαν εάν την στιγμή εμφάνισης της ηχητικής σκηνής το ζώο έστριβε προς την κατεύθυνση του μεγαφώνου από το οποίο παρουσιάστηκε η σκηνή, ώστε να θεωρηθεί ότι του τράβηξε την προσοχή. Για τις λιγότερο σημαντικές σκηνές οι μαϊμούδες έδειξαν σχεδόν τυχαία μέση προτίμηση στην πλευρά εμφάνισης της σκηνής. Αντίθετα, όταν η σκηνή θεωρούνταν περισσότερο σημαντική, έδειξαν σαφή προτίμηση προς την κατεύθυνση εμφάνισής της.

2.3 Μοντέλο Ακουστικού Φάσματος

Σε αυτήν την παράγραφο περιγράφουμε το μοντέλο που αναπτύχθηκε από τους Shamma et al [66, 69], το οποίο μιμείται την επεξεργασία ηχητικών σημάτων στο περιφερειακό και μέσο ακουστικό σύστημα του ανθρώπου, και δίνει ως έξοδο μια απεικόνιση που υπάρχουν ενδείξεις ότι δέχεται το κεντρικό ακουστικό σύστημα για την περαιτέρω επεξεργασία και ανάλυση της ηχητικής σκηνής. Η έξοδος του μοντέλου θα αποτελέσει την βάση στην οποία θα στηριχθεί ο αλγόριθμος ανίχνευσης προεξεχόντων γεγονότων σε ηχητικές σκηνές που αναπτύξαμε και παρουσιάζεται στο Κεφάλαιο 3.

2.3.1 Περιγραφή μοντέλου

Το ανθρώπινο ακουστικό σύστημα έχει την αξιοσημείωτη δυνατότητα ανάλυσης περίπλοκων ακουστικών σκηνών. Η μελέτη του αποτελεί πηγή έμπνευσης για την βελτίωση



Σχήμα 2.10: Τομή του περιφερειακού ακουστικού συστήματος του ανθρώπου.

συστημάτων που λαμβάνουν και επεξεργάζονται ηχητική πληροφορία. Κάποιοι βασικοί μηχανισμοί λειτουργίας του έχουν γίνει γνωστοί έπειτα από έρευνα πολλών ετών, ωστόσο υπάρχουν αρκετά σημεία ακόμα να εξερευνηθούν τόσο στο περιφερειακό σύστημα όσο και σε πιο κεντρικές λειτουργίες στον ανθρώπινο εγκέφαλο. Μία τομή του περιφερειακού ακουστικού συστήματος φαίνεται στο Σχήμα 2.10. Θα παρουσιάσουμε μια απλοποιημένη εκδοχή αυτών των λειτουργιών του, η οποία έχει παρατηρηθεί ότι προσεγγίζει αρκετά καλά τον τρόπο λειτουργίας του όταν διεγείρεται από ευρυζωνικούς ήχους με μέση προς υψηλή ένταση, όπως για παράδειγμα η ανθρώπινη φωνή.

Τα στάδια επεξεργασίας του ηχητικού σήματος στο περιφερειακό σύστημα μπορούν να χωρισθούν σε τρία μέρη: το στάδιο ανάλυσης, το στάδιο μεταγωγής (*transduction*), και το στάδιο αναγωγής (*reduction*). Τα ηχητικά κύματα διεγείρουν το τύμπανο του αυτιού (*eardrum*) θέτοντας το σε ταλάντωση, η οποία μεταδίδεται μέσω τριών μικρών οστών (*malleus, incus, stapes* στο Σχήμα 2.10), στον κοχλία και στα όργανα που υπάρχουν εντός αυτού. Ιδιαίτερης σημασίας είναι η κίνηση της βασικής μεμβράνης (*basilar membrane*) η οποία εκτείνεται κατά μήκος του κοχλία. Το πλάτος και η ακαμψία (*stiffness*) της βασικής μεμβράνης μεταβάλλονται κατά μήκος της, με αποτέλεσμα να δείχνει συχνοτική επιλεκτικότητα. Η ταλάντωση που προκαλείται από έναν απλό τόνο δημιουργεί ένα οδεύον κύμα που μεταδίδεται από την βάση του κοχλία στην κορυφή του, προκαλώντας μια μέγιστη μετατόπιση στην βασική μεμβράνη σε κάποιο σημείο της και στην συνέχεια φθίνει με μεγάλο ρυθμό. Το σημείο που εμφανίζεται το μέγιστο εξαρτάται από την συχνότητα του τόνου, με τις χαμηλές συχνότητες να ταξιδεύουν πιο μακρυά, προς την κορυφή του κοχλία. Η μετατόπιση, επομένως, κάθε σημείου της μεμβράνης είναι μια συνάρτηση της συχνότητας του τόνου που την διεγείρει. Έτσι, η βασική μεμβράνη μπορεί να ειδωθεί σαν μια συστοιχία από ζωνοπερατά φίλτρα, με τις κεντρικές τους συχνότητες να φθίνουν καθώς κινούμαστε προς τη κορυφή του κοχλία. Για συχνότητες άνω των 800 Hz, έχει παρατηρη-

θεί πως οι συναρτήσεις μεταφοράς αποτελούν μετατοπίσεις η μία της άλλης ομοιόμορφα τοποθετημένες σε λογαριθμικό άξονα συχνοτήτων. Εάν $x(t)$ είναι το σήμα που εισέρχεται στο ακουστικό σύστημα, τότε:

$$y_1(t;s) = h(t;s) *_t x(t), \tag{2.4}$$

όπου $h(t;s)$ η κρουστική απόκριση του φίλτρου στην θέση s κατά μήκος του κοχλία, και y_1 η μετατόπιση της μεμβράνης στην θέση αυτή, με $s = 0$ για την βάση, και $s > 0$ καθώς κινούμαστε προς την κορυφή του. Το y_1 είναι η έξοδος του σταδίου της ανάλυσης.

Στο επόμενο στάδιο (στάδιο transduction) οι μηχανικές ταλαντώσεις της βασικής μεμβράνης μετατρέπονται σε ηλεκτρικό δυναμικό, μέσω διατεταγμένων κατά μήκος του κοχλία ακουστικών νευρών. Οι μετατοπίσεις της μεμβράνης προκαλούν, σε κάθε σημείο της, την ροή ιονισμένου υγρού το οποίο στέφει λεπτά νήματα (*cilia*) που πρόσκεινται στα εσωτερικά τριχοφόρα κύτταρα (*inner hair cells*). Η στροφή των νημάτων ρυθμίζει την ροή του υγρού μέσω μη γραμμικών καναλιών προς τα τριχοφόρα κύτταρα, το οποίο δημιουργεί διαφορές δυναμικού εγκαρσίως της μεμβράνης τους. Τέλος, οι διαφορές δυναμικού μεταφέρονται στο κεντρικό ακουστικό σύστημα μέσω των ακουστικών νευρών που συνδέονται με τα τριχοφόρα κύτταρα. Αυτά τα τρία στάδια μπορούν να μοντελοποιηθούν αρκετά καλά μέσω μιας διαδικασίας τριών βημάτων: χρήση της ταχύτητας ταλάντωσης της βασικής μεμβράνης για τον ρυθμό ροής φορτίου, μία στιγμιαία μη-γραμμικότητα για τα μη γραμμικά κανάλια, και ένα βαθυπερατό φίλτρο με σχετικά μικρή σταθερά χρόνου($<$ 0.3 ms). Πιο συνοπτικά, οι ταλαντώσεις της βασικής μεμβράνης, $y_1(t;s)$, μετατρέπονται σε δυναμικό, $y_2(t;s)$, εντός των τριχοφόρων κυττάρων, ως εξής:

$$y_2(t;s) = g(\partial_t y_1(t;s)) *_t w(t), \tag{2.5}$$

όπου g σιγμοειδής μη-γραμμικότητα, και w η κρουστική απόκριση του βαθυπερατού φίλτρου. Η συνάρτηση g είναι της μορφής:

$$g(u) = \frac{1}{1 + e^{-\gamma u}} - \frac{1}{2}, \tag{2.6}$$

όπου γ είναι το κέρδος.

Μετά την μεταφορά του ηλεκτρικού δυναμικού στο κεντρικό ακουστικό σύστημα, λαμβάνουν χώρα διαδικασίες για την εξαγωγή διάφορων χαρακτηριστικών του ηχητικού σήματος που το προκάλεσε, όπως το τέμπο, το pitch, και η θέση της πηγής στον χώρο. Ένα ιδιαίτερα σημαντικό χαρακτηριστικό το οποίο συμβάλλει στην αναγνώριση των διαφόρων ήχων, είναι το ακουστικό φάσμα βραχέως χρόνου. Το μοντέλο εξάγει μια προσέγγιση αυτού του ακουστικού φάσματος, ενώ αγνοεί τα υπόλοιπα χαρακτηριστικά. Για τον υπολογισμό του φάσματος χρησιμοποιείται ένα πλευρικό απαγορευτικό δίκτυο (*lateral inhibitory network-LIN*), το οποίο ανιχνεύει ασυνέχειες κατά μήκος του χωρικού άξονα των ακουστικών νευρών, και έπειτα τα ολοκληρώνει σε διάστημα μερικών χιλιοστών του δευτερολέπτου. Αυτές οι ασυνέχειες οφείλονται σε διαφορετικές συχνότητες, φάσεις, ή πλάτη που πιθανώς έχουν οι κυματομορφές στα διάφορα κανάλια. Ο τρόπος λειτουργίας του LIN μπορεί να περιγραφεί με τρία στάδια:

1. Μία παράγωγος κατά μήκος του χωρικού άξονα του κοχλία για την ενίσχυση των ασυνεχειών, που ακολουθείται από μία τοπική ομαλοποίηση λόγω του πεπερασμένου χωρικού εύρους αλληλεπίδρασης μεταξύ των ασυνεχειών:

$$\begin{aligned} y_3(t;s) &= \partial_s(y_2(t;s)) *_s v(s) \\ &= (g'(\partial_t y_1(t;s)) \cdot \partial_s \partial_t y_1(t;s)) *_t w(t) *_s v(s). \end{aligned} \tag{2.7}$$

2. Ένας ανορθωτής μισού κύματος, για την μοντελοποίηση της μη γραμμικότητας κατωφλίου που εμφανίζεται στα LIN δίκτυα, το οποίο μπορεί εκφραστεί ως:

$$y_4(t;s) = max(y_3(t;s), 0). \tag{2.8}$$

3. Ένας, ολοκληρωτής με μεγάλη χρονική σταθερά. Αυτό πραγματοποιείται διότι οι κεντρικοί ακουστικοί νευρώνες αδυνατούν να παρακολουθήσουν τις ταχείς χρονικές μεταβολές (υψηλότερες μερικών εκατοντάδων hertz). Η έξοδος του LIN είναι:

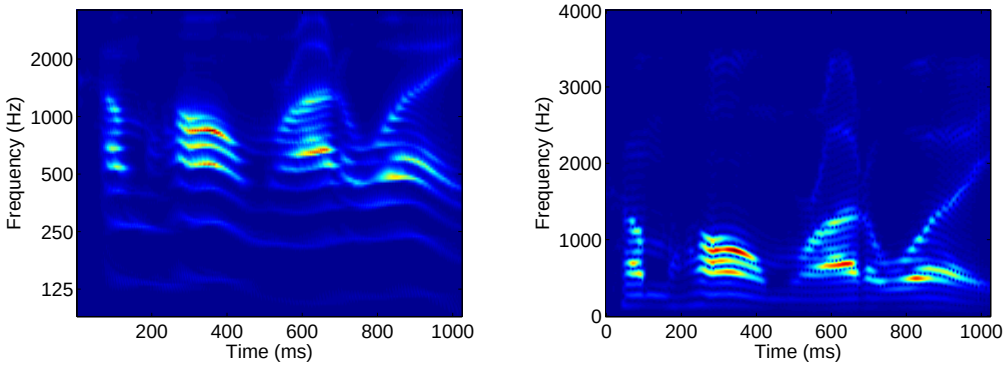
$$y_5(t;s) = y_4(t;s) *_t \Pi_T(t), \tag{2.9}$$

όπου Π_T, παράθυρο διάρκειας $T \approx 10 - 20\ ms$.

Η έξοδος του LIN είναι και η έξοδος του μοντέλου, δηλαδή το ακουστικό φάσμα βραχέως χρόνου. Τα στάδια του μοντέλου είναι απλοποιημένες εκδοχές των διαδικασιών που συμβαίνουν στο ακουστικό σύστημα, που για ορισμένες εφαρμογές ίσως είναι κρίσιμες. Ωστόσο, πειραματικές δοκιμές έχουνε δείξει ότι το μοντέλο διατηρεί όλη την φασματική πληροφορία του ακουστικού σήματος. Στην συνέχεια δείχνουμε ορισμένα παραδείγματα ακουστικών φασμάτων, όπως υπολογίστηκαν από το μοντέλο, καθώς και το φάσμα που υπολογίζεται με μετασχηματισμό Fourier βραχέως χρόνου(*STFT*), προς χάριν σύγκρισης των δύο απεικονίσεων.

2.3.2 Παραδείγματα απεικονίσεων

Σε αυτή την ενότητα παρουσιάζουμε απεικονίσεις του ηχητικού φάσματος σημάτων, όπως υπολογίστηκαν με την διαδικασία που περιγράφηκε στην προηγούμενη παράγραφο, και τα συγκρίνουμε με το φασματογράφημα όπως υπολογίζεται με χρήση μετασχηματισμού Fourier βραχέως χρόνου (STFT). Οι ηχητικές σκηνές περιλαμβάνουν ανθρώπινη ομιλία, μουσική, απλούς τόνους και φυσικούς ήχους. Για τον υπολογισμό του φάσματος χρησιμοποιήθηκαν 128 φίλτρα σε εύρος 5.4 οκτάβων. Η συνάρτηση g θεωρείται ότι λαμβάνει τιμές γύρω από την γραμμική της περιοχή και προσεγγίζεται με μία γραμμική συνάρτηση, ενώ το βαθυπερατό φιλτράρισμα με το φίλτρο w παραλείπεται (εξίσωση 2.5). Η παραγώγιση κατά μήκος του άξονα του κοχλία προσεγγίζεται με διαφορές μεταξύ των διαδοχικών καναλιών (από το κανάλι μικρότερης συχνότητας αφαιρείται η μεγαλύτερη), και τέλος η χρονική ολοκλήρωση γίνεται μέσω δύο σταδίων: ενός βαθυπερατού φιλτραρίσματος, και υποδειγματοληψίας με παράγοντα αντίστροφο του μήκους του παραθύρου ολοκλήρωσης (*leaky integration*).



(α΄) (β΄)

Σχήμα 2.11: Ακουστικό φάσμα της φράσης “come home right away”, από άνδρα ομιλητή με χρήση μοντέλου ακουστικού φάσματος (αριστερά), με STFT (δεξιά).

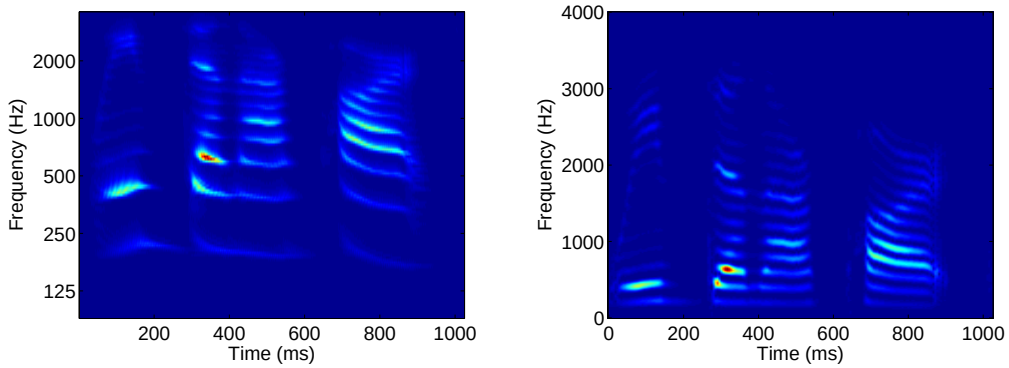
Στα Σχήματα 2.11, 2.12, βλέπουμε φάσματα από ανθρώπινη ομιλία. Στο Σχήμα 2.13 το νιαούρισμα μιας γάτας σε θορυβώδες περιβάλλον. Παρατηρούμε την αξιοσημείωτη ομοιότητα μεταξύ της απεικόνισης του ακουστικού μοντέλου και του STFT. Αυτή οφείλεται στην απλοποίηση αρκετών σταδίων της διαδικασίας, που ανάγουν τον μετασχηματισμό σε φιλτράρισμα του σήματος με απότομα φίλτρα (λόγω των διαφορών κατά μήκος του κοχλία), που είναι ανάλογο με την δειγματοληψία που πραγματοποιεί ο DFT στο σήμα, μία ανόρθωση μισού κύματος και μια ολοκλήρωση βραχέως χρόνου. Η χρήση μη-γραμμικής συνάρτησης g θα παραμόρφωνε το συχνοτικό περιεχόμενο του σήματος, αποκόπτοντας κυρίως τις υψηλές συχνότητες, και διαφοροποιούσε σε μεγαλύτερο βαθμό τις δύο απεικονίσεις. Σε αυτήν την εργασία γίνεται χρήση του απλοποιημένου μοντέλου.

2.4 Μέθοδοι Αξιολόγησης της Σημαντικότητας

Μη γνωρίζοντας τον τρόπο που ο ανθρώπινος εγκέφαλος επεξεργάζεται την εισερχόμενη πληροφορία, η υπολογιστική μοντελοποίηση της ανθρώπινης προσοχής βασίζεται σε πειραματικές ενδείξεις. Από την λήψη απόκρισης από τους χρήστες σε ερεθίσματα, εκτιμάται η σημαντικότητα των ερεθισμάτων στην οποία στηρίζεται η ανάπτυξη και αξιολόγηση των υπολογιστικών μοντέλων. Στις επόμενες παραγράφους παρουσιάζονται κάποιες από τις μεθόδους αξιολόγησης που χρησιμοποιούνται, καθώς και η μέθοδος που ακολουθείται σε αυτή την εργασία.

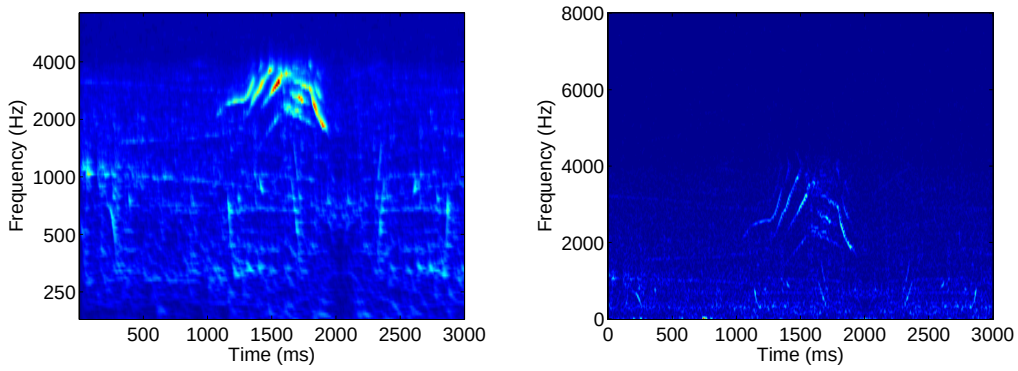
2.4.1 Αξιολόγηση της Οπτικής Σημαντικότητας

Ως μέθοδος αξιολόγησης των αλγορίθμων ανίχνευσης σημαντικών σημείων σε εικόνες και βίντεο, χρησιμοποιείται ευρέως τα τελευταία χρόνια ο *παρακολουθητής βλέμματος* (*eye tracker*) [17]. Ο παρακολουθητής βλέμματος έχοντας την δυνατότητα παρακολούθησης της κίνησης του ανθρώπινου οφθαλμού και εκτίμησης του σημείου της εικόνας στο



(α΄) (β΄)

Σχήμα 2.12: Ακουστικό φάσμα της φράσης “we’ve done apart”, από γυναίκα ομιλήτρια με χρήση μοντέλου ακουστικού φάσματος (αριστερά), με STFT (δεξιά).



(α΄) (β΄)

Σχήμα 2.13: Ακουστικό φάσμα του νιαουρίσματος μιας γάτας σε θορυβώδες περιβάλλον, με χρήση μοντέλου ακουστικού φάσματος (αριστερά), με STFT (δεξιά).

οποίο εστίασε, αποτελεί έναν άμεσο τρόπο σύγκρισης της εξόδου των αλγορίθμων με μετρήσεις που έχουν ληφθεί από ανθρώπους. Έχει το πλεονέκτημα ότι συλλαμβάνει στιγμιαία την κίνηση του ανθρώπινου οφθαλμού μη-απαιτώντας εκ των υστέρων αναφορά των χρηστών του τι είδαν στην εικόνα ώστε να σκεφτούν την απάντηση τους, κάνοντας δυνατό τον έλεγχο εάν ένα αντικείμενο τράβηξε αυθόρμητα την προσοχή τους (*pop-out*) ή έπειτα από σάρωση της εικόνας και παρατήρησης άλλων αντικειμένων πριν από αυτό. Είναι, επομένως, δυνατός ο προσδιορισμός των προεξεχόντων περιοχών μιας εικόνας, και η απόδοση επιπέδων σημαντικότητας στα αντικείμενα που την συνθέτουν με βάση την τροχιά που έχει διαγράψει το βλέμμα των παρατηρητών πάνω σε αυτή.

Ωστόσο, το βλέμμα των χρηστών είναι πιθανό να κατευθυνθεί σε μία περιοχή λόγω υψηλότερου επιπέδου διεργασιών, όπως ο στόχος του παρατηρητή, και όχι των χαρακτηριστικών χαμηλού επιπέδου της εικόνας.

2.4.2 Αξιολόγηση της Ακουστικής Σημαντικότητας

Η έλλειψη ενός εύκολου τρόπου μέτρησης της κατάστασης ενός φυσικού μέσου με την οποία σχετίζεται άμεσα η ακουστική προσοχή, όπως είναι η κίνηση του οφθαλμού, δυσχεραίνει την αξιολόγηση και ανάπτυξη των μεθόδων μοντελοποίησης της. Οι υπάρχουσες μέθοδοι προσπαθούν με έμμεσο τρόπο να μετρήσουν εάν ένα ηχητικό γεγονός τράβηξε αυθόρμητα την ακουστική προσοχή, και να θέσουν στάθμες σημαντικότητας.

Μία μέθοδος αξιολόγησης που έχει χρησιμοποιηθεί από τους Kayser et al [31], και Duangudom [15] είναι μέσω σύγκρισης της σημαντικότητας δύο ηχητικών σκηνών. Ζητείται από χρήστες να ακούσουν ταυτόχρονα δύο ηχητικές σκηνές διάρκειας 1 έως 3 δευτερολέπτων, και να επιλέξουν μία ως περισσότερο προεξέχουσα από την άλλη, ή να τις θεωρήσουν εξίσου προεξέχουσες. Διαπιστώθηκε υψηλός βαθμός συσχέτισης της απόκρισης των χρηστών με τις στάθμες σημαντικότητας που έθετε το μοντέλο.

Οι Kayser et al για την μέτρηση της σημαντικότητας των ηχητικών σκηνών πραγματοποίησαν ένα πείραμα με χρήση μαϊμούδων [31]. Παρουσίασαν σε μαϊμούδες ηχητικές σκηνές ταυτόχρονα από δύο μεγάφωνα που βρισκόταν αντιδιαμετρικά του ζώου. Από το ένα μεγάφωνο παράγονταν η ηχητική σκηνή ενώ από το άλλο θόρυβος. Ανάλογα με την σημαντικότητα της ηχητικής σκηνής ανέμεναν το ζώο να στραφεί προς το μεγάφωνο από το οποίο παρουσιάστηκε αυτή. Παρατηρήθηκε συσχέτιση του προσανατολισμού του ζώου με την έξοδο του μοντέλου των ερευνητών. Όσο πιο προεξέχων προέβλεπε το μοντέλο ότι είναι ο ήχος, έδειχναν μεγαλύτερη τάση προσανατολισμού προς την κατεύθυνση που εμφανίστηκε.

Μία άλλη μέθοδος για την μέτρηση της σημαντικότητας είναι η πραγματοποίηση δύο εργασιών ταυτόχρονα (*dual task*). Τα dual tasks έχουν χρησιμοποιηθεί εκτενώς για την διερεύνηση γνωσιακών χαρακτηριστικών του ανθρώπου γενικότερα, αλλά και λειτουργιών του ακουστικού συστήματος πιο συγκεκριμένα [10]. Σε αυτού του είδους τα πειράματα ζητείται από άτομα να πραγματοποιήσουν ταυτόχρονα δύο εργασίες οι οποίες μοιράζονται υπολογιστικούς πόρους του ανθρώπου, και εξετάζεται η απόδοση σε κάθε μία όταν πραγματοποιούνται ταυτόχρονα σε σύγκριση με την διεξαγωγή κάθε μίας χωριστά.

Η Duangudom [15] προσπάθησε να μετρήσει την σημαντικότητα των ηχητικών σκη-

νών μέσω dual task πειράματος. Ζήτησε από έναν αριθμό ατόμων να ακούσει δύο ηχητικές σκηνές ταυτόχρονα. Η μία αποτελούνταν από μη-επικαλυπτόμενους απλούς τόνους συχνότητας 100 Hz και 200 Hz πολύ μικρής διάρκειας ο καθένας, με μικρή παύση μεταξύ τους, και τα άτομα έπρεπε να μετρήσουν το πλήθος των τόνων συχνότητας 200 Hz που εμφανίστηκε. Αυτή ήταν η σκηνή στην οποία έπρεπε να συγκεντρωθεί η προσοχή των ατόμων. Η άλλη σκηνή αποτελούνταν από σύμπλεγμα 4 τόνων, και ένας από αυτούς διαμορφωνόταν κατά πλάτος για μια χρονική διάρκεια σε ορισμένες μόνο από τις σκηνές. Για κάθε σκηνή που τους παρουσιάζονταν τα άτομα έπρεπε να απαντήσουν εάν εμφανίστηκε ο διαμορφωμένος τόνος ή όχι σε αυτή. Η υπόθεση είναι ότι ακόμη και εάν η προσοχή είναι στραμμένη αλλού, ένας προεξέχων ήχος θα γίνει αντιληπτός. Τα αποτελέσματα έδειξαν συσχέτιση του δείκτη διαμόρφωσης με την σημαντικότητα του ήχου. Όσο πιο υψηλός ήταν ο δείκτης διαμόρφωσης, τόσο πιο εύκολα εντοπίσιμος ήταν ο διαμορφωμένος τόνος.

Τέλος, μία ακόμη μέθοδος αξιολόγησης των αλγορίθμων είναι μέσω της σημείωσης των σημαντικών γεγονότων μίας βάσης ηχητικών δεδομένων. Οι Kim et al [33] ζήτησαν από έναν αριθμό ατόμων να ακούσει ηχητικά αρχεία από μία σειρά συνεδριάσεων σε κλειστές αίθουσες, και με την υπόθεση ότι βρίσκονται στον χώρο συνεδρίασης να σημειώσουν με κατάλληλο λογισμικό οποιαδήποτε χρονική διάρκεια έστρεψαν την προσοχή τους ακούσια ή μη σε κάποια ηχητική πηγή. Τα αρχεία αυτά περιείχαν διάφορες κατηγορίες ήχων, όπως ανθρώπινη ομιλία, γέλιο, χτύπημα της πόρτας, και είχαν διάρκεια δώδεκα ώρες. Παρατηρήθηκε ότι κάποιοι ήχοι, όπως ανθρώπινα βήματα ή χτύπημα της πόρτας, τραβούσαν πάντα την προσοχή των περισσότερων χρηστών, ενώ κάποιοι άλλοι άλλοτε τραβούσαν την προσοχή όλων και άλλοτε όχι. Ωστόσο υπήρχε σημασιολογική επικάλυψη (ανήκαν σε ίδια κατηγορία) μεταξύ των ήχων που τραβούσαν πάντα την προσοχή των περισσότερων χρηστών και αυτών που άλλοτε τραβούσαν την προσοχή όλων και άλλοτε όχι. Αυτό καταδεικνύει την σημασία υψηλότερου επιπέδου διεργασιών που διαμορφώνουν (*modulate*) την ανθρώπινη προσοχή, και συνιστά ότι οι προεξέχοντες ήχοι πιθανώς να μην τραβούν πάντοτε την προσοχή όλων των χρηστών.

Σε αυτή την εργασία δεν αναπτύσσεται κάποια νέα μέθοδος αξιολόγησης, αλλά υιοθετείται η μέθοδος της σημείωσης σημαντικών γεγονότων σε μία βάση δεδομένων, η οποία περιγράφεται στην ενότητα 2.4.3.

2.4.3 Η Βάση δεδομένων COGNIMUSE

Η βάση που χρησιμοποιήθηκε αποτελείται από αποσπάσματα έξι ταινιών που παρήχθησαν στο Hollywood, και είναι οι ακόλουθες: “Chicago” (CHI), “Crash” (CRA), “The Departed” (DEP), “Finding Nemo” (FNE), “Gladiator” (GLA), και “Lord of the Rings - the Return of the King” (LOR) [1]. Οι ταινίες αυτές επιλέχτηκαν διότι περιείχαν ηχητικές σκηνές με διάφορες κατηγορίες ήχων, όπως μουσική, ανθρώπινη ομιλία, φυσικούς και συνθετικούς ήχους, σε διάφορα επίπεδα έντασης και ήταν δυνατός ο έλεγχος του αλγόριθμου στην ανίχνευση μεταβολών σε διαφορετικά χαρακτηριστικά του ήχου. Η συνολική διάρκεια των αποσπασμάτων είναι 185.1 λεπτά. Ζητήθηκε από τρία άτομα να ακούσουν τα αποσπάσματα, χωρίς να βλέπουν την εικόνα, και να σημειώσουν με χρήση του προγράμματος επισημείωσης ANVIL [34], ποια σημεία τους “τράβηξαν την προσοχή”. Η

Πίνακας 2.1: Ποσοστό της διάρκειας της ταινίας που θεωρήθηκε σημαντικό από ακριβώς ένα, δύο και τρία άτομα στις στήλες 3 έως 5 αντίστοιχα, και τουλάχιστον ένα και δύο στις στήλες 6 και 7.

Movie	Duration (minutes)	One (%)	Two (%)	Three (%)	≥ One (%)	≥ Two (%)
CHI	30.14	17.16	21.48	36.28	74.92	57.76
CRA	26.62	18.51	22.40	33.86	74.77	56.26
DEP	30.47	22.35	11.98	18.00	52.33	29.98
FNE	30.29	23.14	20.05	33.47	76.66	53.52
GLA	30.05	19.21	16.53	43.44	79.18	59.97
LOR	37.56	17.46	18.27	40.34	76.07	58.61
Total	185.13	19.58	18.36	34.45	72.39	52.81

ακρίβεια της επισημείωσης είναι σε επίπεδο καρέ (frame) εικόνας, και τα συγκεκριμένα αποσπάσματα έχουν συχνότητα 25 καρέ ανά δευτερόλεπτο. Η επισημείωση δεν γινόταν σε πραγματικό χρόνο, καθώς τα άτομα έπρεπε να εντοπίσουν σε ποιο καρέ ήταν η αρχή ενός γεγονότος που τους τράβηξε την προσοχή. Κάποια στατιστικά στοιχεία των αποσπασμάτων και της επισημείωσης των χρηστών φαίνονται στον Πίνακα 2.1.

Τα ερεθίσματα των ταινιών είναι αρκετά περίπλοκα και η προσοχή των χρηστών επηρεάζεται έντονα από διαδικασίες υψηλότερου επιπέδου, που σχετίζονται με την σημασιολογία των σκηνών. Σε αρκετές περιπτώσεις είναι σκοπίμως δομημένα ώστε να κατευθύνουν την προσοχή των θεατών σε συγκεκριμένα σημεία, και να γεννούν συγκεκριμένα συναισθήματα. Συνεπώς, εκτός από τα σημεία που τράβηξαν αυθόρμητα την προσοχή των χρηστών, έχουν σημειωθεί και άλλα που οφείλονται στην διαμόρφωση της προσοχής από σημασιολογική πληροφορία. Επίσης, κάποια σημεία τα οποία είναι προεξέχοντα αλλά όχι πολύ έντονα, δεν σημειώνονται όπως φαίνεται και από τα αποτελέσματα των Kim et al [33].

Επειδή η ανθρώπινη προσοχή επηρεάζεται από διάφορους παράγοντες και παρατηρείται ασυμφωνία μεταξύ της σημείωσης των χρηστών, ως σημαντικά γεγονότα θα θεωρούνται εκείνα που σημειώθηκαν από τουλάχιστον δύο χρήστες. Αυτά τα γεγονότα έχουν αυξημένη πιθανότητα να είναι σημειωμένα επειδή τράβηξαν ακούσια την προσοχή των ατόμων, και όχι λόγω σημασιολογικής πληροφορίας της σκηνής. Όπως φαίνεται στον Πίνακα 2.1, υπάρχει συμμετρία στη διάρκεια των σημαντικών και μη-σημαντικών γεγονότων, με τα σημαντικά να καλύπτουν χρονικά το 52.8% της διάρκειας των ταινιών.

2.4.4 Μέτρα Αξιολόγησης της επίδοσης

Έστω ένα πρόβλημα δυαδικής ταξινόμησης, όπου τα δεδομένα διαμερίζονται σε μία θετική (*positive*) κλάση και μία αρνητική (*negative*). Έστω, επίσης, ένας δυαδικός ταξινο-

Πίνακας 2.2: Πίνακας σύγχυσης για την αξιολόγηση ενός ταξινομητή.

		True Class	
		positive	negative
Predicted Class	positive	tp	fp
	negative	fn	tn

μητής για το πρόβλημα. Με βάση την πραγματική κλάση των δεδομένων και την έξοδο του ταξινομητή, κάθε ένα λαμβάνει έναν από τους ακόλουθους τέσσερις χαρακτηρισμούς: εάν το δεδομένο ανήκει στην θετική κλάση και ταξινομηθεί ως θετικό τότε ονομάζεται *true positive* (tp), ενώ εάν ταξινομηθεί ως αρνητικό ονομάζεται *false negative* (fn). Εάν ανήκει στην αρνητική κλάση και ταξινομηθεί ως αρνητικό καλείται *true negative* (tn), ενώ εάν ταξινομηθεί ως θετικό καλείται *false positive* (fp).

Πιο συνοπτικά, ο Πίνακας 2.2 (*πίνακας σύγχησης - confusion matrix*) δείχνει τα τέσσερα δυνατά αποτελέσματα έπειτα από την ταξινόμηση. Τα στοιχεία της διαγωνίου του πίνακα αποτελούν τις ορθές ταξινομήσεις (χαρακτηρισμός true), ενώ τα στοιχεία εκτός διαγωνίου τις εσφαλμένες ταξινομήσεις (χαρακτηρισμός false). Με χρήση αυτών των χαρακτηρισμών ορίζονται μέτρα αξιολόγησης της επίδοσης που θα χρησιμοποιηθούν σε αυτήν την εργασία και περιγράφονται στη συνέχεια.

Ένα μέτρο της επίδοσης ενός ταξινομητή που χρησιμοποιείται ευρέως είναι η *ακρίβεια ταξινόμησης* (*accuracy*), το οποίο ορίζεται ως εξής:

$$acc = \frac{tp + tn}{tp + fp + tn + fn} \tag{2.10}$$

Η ακρίβεια δείχνει τι ποσοστό των δεδομένων ταξινομήθηκε ορθά. Μπορεί επίσης να ειδωθεί ως η πιθανότητα ένα νέο δεδομένο να ταξινομηθεί ορθά. Επιθυμητό σε ένα πρόβλημα ταξινόμησης είναι το accuracy να είναι υψηλό.

Ένα άλλο μέτρο που χρησιμοποιείται είναι το *precision*:

$$precision = \frac{tp}{tp + fp} \tag{2.11}$$

το οποίο δείχνει τι ποσοστό των σημείων που ταξινομήθηκαν ως θετικά είναι πράγματι θετικά. 'Η ισοδύναμα την πιθανότητα ένα δεδομένο που έχει ταξινομηθεί ως θετικό, να έχει προέλθει από την θετική κλάση.

Ένα επίσης πολύ σημαντικό μέτρο είναι το *recall* ή *true positive rate*:

$$recall = \frac{tp}{tp + fn} \tag{2.12}$$

το οποίο δείχνει τι ποσοστό των θετικών δεδομένων ανιχνεύθηκαν (πιθανότητα ανίχνευσης ενός θετικού δεδομένου). Στα προβλήματα ταξινόμησης παρατηρείται συνήθως συμπληρωματικότητα των precision και recall, με αύξηση του ενός να οδηγεί σε πτώση του άλλου.

Σε αντιστοιχία με το recall ορίζεται και το *false positive rate* ενός ταξινομητή:

$$fpr = \frac{fp}{fp + tn} \tag{2.13}$$

το οποίο δείχνει τι ποσοστό των αρνητικών δεδομένων δεν ανιχνεύθηκαν. Ισοδύναμα, την πιθανότητα ένα δεδομένο που ταξινομήθηκε ως θετικό να έχει προέλθει από την αρνητική κλάση.

Τέλος με βάση τα ανωτέρω μέτρα ορίζεται το *f-score* ενός ταξινομητή ως εξής:

$$f_{sc} = 2\frac{precision \cdot recall}{precision + recall} \tag{2.14}$$

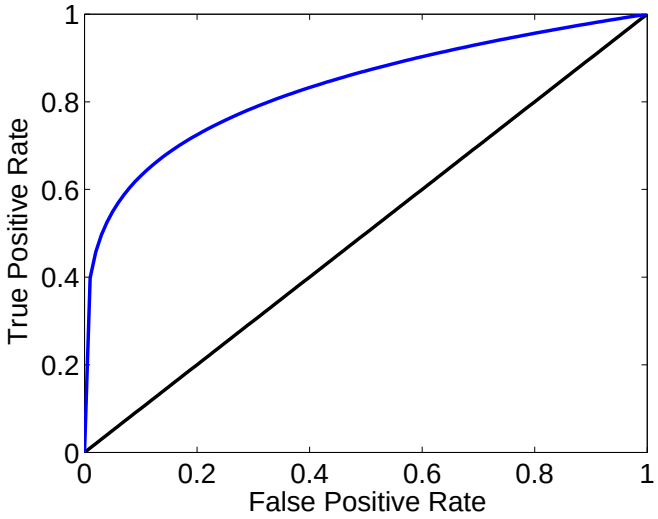
Στην παρούσα εργασία θα γίνει χρήση των ανωτέρω μέτρων για την αξιολόγηση των μεθόδων ανίχνευσης σημαντικών γεγονότων. Ως θετική κλάση θα θεωρείται το σύνολο των σημαντικών σημείων ενώ ως αρνητική το σύνολο των μη-σημαντικών. Επειδή είναι επιθυμητή η ανίχνευση σημαντικών γεγονότων, δίνεται βάση στην επίτευξη υψηλότερου recall παρά precision, ωστόσο και αυτό είναι καλό να κυμαίνεται σε υψηλά επίπεδα. Το accuracy επίσης είναι επιθυμητό να είναι υψηλό.

Μεταβάλλοντας κάποιες από τις παραμέτρους ενός ταξινομητή λαμβάνονται διαφορετικά ζεύγη (fpr, tpr) στο επίπεδο, τα οποία σχηματίζουν μία καμπύλη. Η καμπύλη αυτή συχνά καλείται *Receiver Operating Characteristic - ROC* [19]. Επειδή τα tpr, fpr λαμβάνουν τιμές στο διάστημα $[0, 1]$, η καμπύλη βρίσκεται εντός του μοναδιαίου τετραγώνου. Ένα παράδειγμα καμπύλης φαίνεται στο Σχήμα 2.14. Τα σημεία $(0, 0)$ και $(1, 1)$ της καμπύλης αντιστοιχούν στην ταξινόμηση όλων των σημείων ως αρνητικών (negatives) και θετικών (positives), αντίστοιχα. Το σημείο $(0, 1)$ αντιστοιχεί στην ιδανική ταξινόμηση, όπου η κλάση όλων των στοιχείων έχει προβλεφθεί σωστά. Τα σημεία της διαγωνίου αντιστοιχούν σε τυχαία ταξινόμηση (*random guess*) καθώς οι πιθανότητες ένα δείγμα να έχει προέλθει από την ορθή ή λάθος κλάση είναι ίσες. Σημεία άνω της διαγωνίου αντιστοιχούν σε επίδοση καλύτερη της τυχαίας ταξινόμησης ενώ σημεία κάτω της διαγωνίου χειρότερης αυτής. Ωστόσο, εάν κάποιος ταξινομητής παράγει συστηματικά ζεύγη (tpr, fpr) κάτω της διαγωνίου μπορεί να θεωρηθεί αποδοτικός καθώς με τη συμπληρωματική ταξινομήση λαμβάνονται σημεία συμμετρικά του κέντρου $(0.5, 0.5)$ τα οποία βρίσκονται άνω της διαγωνίου. Επομένως είναι επιθυμητό τα σημεία να απέχουν μεγάλη απόσταση από τη διαγώνιο και να βρίσκονται κοντά στα σημεία $(0, 1)$ και $(1, 0)$.

Ένα μέτρο που χαρακτηρίζει την καμπύλη ROC είναι το εμβαδόν του χωρίου (*Area Under Curve - AUC*) που εμπεριέχεται μεταξύ αυτής και της διαγωνίου του μοναδιαίου τετραγώνου, και είναι ένα μέτρο της ευρωστίας του ταξινομητή. Επιθυμητή είναι η επίτευξη υψηλού AUC κατά απόλυτη τιμή. Η μέγιστη τιμή που μπορεί να λάβει το AUC είναι 0.5, και η ελάχιστη μηδέν. Για την αξιολόγηση των υπολογιστικών μοντέλων που θα εξετασθούν θα γίνει υπολογισμός ROC καμπυλών και των AUC που προκύπτουν από αυτές.

2.4.5 Πρόβλεψη ενός χρήστη από τους υπόλοιπους

Σε προβλήματα ταξινόμησης όπου χρησιμοποιείται μία βάση δεδομένων για εξαγωγή χαρακτηριστικών και ελέγχου των υπολογιστικών μοντέλων σύμφωνα με ταξινόμηση που



Σχήμα 2.14: Παράδειγμα ROC καμπύλης (με μπλε χρώμα).

έχουν πραγματοποιήσει χρήστες, συχνά χρησιμοποιούνται τα δεδομένα των χρηστών για να γίνει μία εκτίμηση του πόσο δύσκολη είναι η ταξινόμηση στη συγκεκριμένη βάση και των άνω φραγμάτων στην απόδοση των αλγορίθμων.

Τα δεδομένα των χρηστών χωρίζονται σε δεδομένα εκπαίδευσης και επαλήθευσης, και γίνεται έλεγχος σε ποιο βαθμό τα δεδομένα επαλήθευσης μπορούν να προβλέψουν τα δεδομένα εκπαίδευσης. Δηλαδή, τα δεδομένα επαλήθευσης χειρίζονται ως να ήταν η έξοδος ενός υπολογιστικού μοντέλου του οποίου η απόδοση ελέγχεται. Για την περίπτωση ταξινόμησης σε κλάσεις σημαντικότητας στην βάση που κάνουμε έλεγχο του μοντέλου, θα θεωρήσουμε την επισημείωση των δύο χρηστών ως δεδομένα αλήθειας και την επισημείωση του τρίτου χρήστη ως δεδομένα επαλήθευσης. Θα ελέγξουμε με τι ακρίβεια ο τρίτος χρήστης μπορεί να προβλέψει τους άλλους δύο, και για τους τρεις δυνατούς συνδυασμούς.

Στον Πίνακα 2.3 φαίνονται μέσα ποσοστά ταξινόμησης ως προς τους χρήστες για κάθε ταινία της βάσης, ενώ στον Πίνακα 2.4 φαίνεται το f-score κάθε χρήστη. Λαμβάνεται μέση ακρίβεια ταξινόμησης 75.1%, f-score 72.72%, και recall 84.6%, με μικρή διακύμανση μεταξύ των χρηστών. Σύμφωνα με αυτές τις μετρήσεις η πρόβλεψη των υπόλοιπων χρηστών από έναν έχει "μέτρια" δυσκολία η οποία χαρακτηρίζει την βάση δεδομένων. Ανάλογη απόδοση αναμένεται να έχουν και οι αλγόριθμοι ταξινόμησης σημαντικών γεγονότων.

Πίνακας 2.3: Μέσο ποσοστό αναγνώρισης για τους χρήστες σε κάθε ταινία της βάσης.

Movie	Acc	Prec	Recall
CHI	72.80	64.54	83.62
CRA	71.43	62.00	82.01
DEP	80.57	54.31	82.26
FNE	72.24	61.46	83.44
GLA	77.07	71.75	88.94
LOR	75.91	69.42	87.00
Total	75.11	64.15	84.69

Πίνακας 2.4: F-score για κάθε χρήστη σε κάθε ταινία της βάσης, και συνολικό F-score.

Movie	User 1	User 2	User 3	Total
CHI	72.93	73.57	71.72	72.74
CRA	69.10	73.97	68.18	70.42
DEP	65.82	67.38	61.88	65.03
FNE	70.76	73.14	68.31	70.74
GLA	77.26	78.75	81.46	79.16
LOR	77.93	78.27	74.91	77.04
Total	72.57	74.33	71.26	72.72

Κεφάλαιο 3

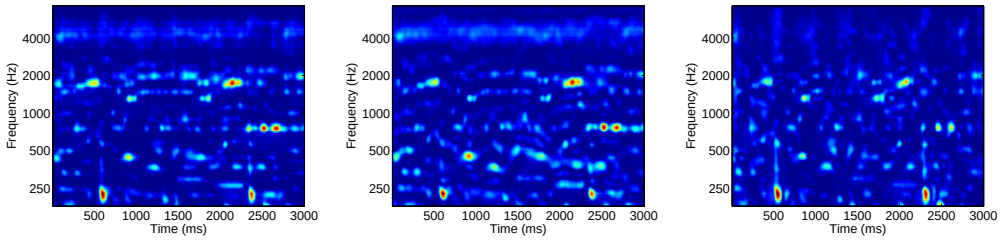
Εφαρμογή Μοντέλου των Kayser et al

Εισαγωγή

Στις επόμενες ενότητες εφαρμόζεται το μοντέλο των Kayser et al [31] για την ταξινόμηση ηχητικών σκηνών. Από τον χάρτη σημαντικότητας υπολογίζεται καμπύλη σημαντικότητας την οποία κατωφλιώνοντας πραγματοποιείται ταξινόμηση των ηχητικών σκηνών. Λαμβάνονται αποτελέσματα που έχουν συσχέτιση με την ανθρώπινη θεώρηση της σημαντικότητας. Επίσης, εισάγεται η έννοια του gist μιας σκηνής και δημιουργούνται διανύσματα από τους χάρτες, τα οποία ταξινομούνται με αλγορίθμους μηχανικής μάθησης. Τέλος δοκιμάζεται η μείωση της διάστασης των gist διανυσμάτων και συγκρίνεται η απόδοση τους με τα πλήρη διανύσματα. Με χρήση των gist διανυσμάτων επιτυγχάνονται υψηλότερα ποσοστά ταξινόμησης σε σύγκριση με την κατωφλίωση των καμπυλών.

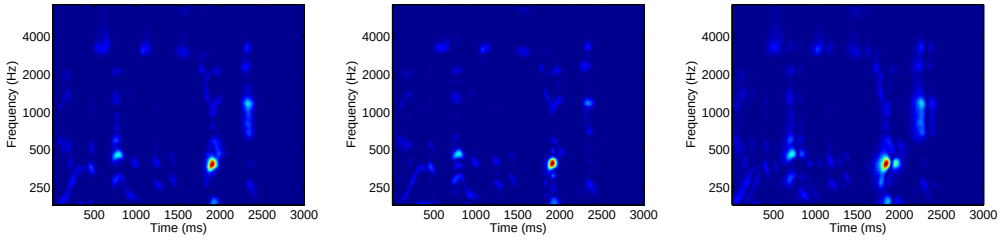
3.1 Εξαγωγή Χαρτών

Χρησιμοποιήθηκε το μοντέλο των Kayser et al για την ταξινόμηση ηχητικών σκηνών στην βάση COGNIMUSE. Ως είσοδος στο μοντέλο δίνεται το ακουστικό φάσμα όπως υπολογίζεται με το μοντέλο του Shamma [69] και περιγράφηκε στην ενότητα 2, αντί για το κλασσικό φασματογράφημα στο οποίο στηρίχθηκε το μοντέλο από τους συγγραφείς. Το κύριο πλεονέκτημα χρήσης του ακουστικού φάσματος είναι η λογαριθμική κατανομή συχνοτήτων με συνέπεια να υπάρχει μεγαλύτερη ανάλυση στις χαμηλές συχνότητες όπου συγκεντρώνεται η ενέργεια πολλών φυσικών σημάτων, όπως η φωνή. Επιπλέον το μοντέλο είναι εμπνευσμένο από τον τρόπο λειτουργίας του ανθρώπινου αυτιού. Στις επόμενες υπο-ενότητες παρουσιάζονται χάρτες χαρακτηριστικών όπως υπολογίσθηκαν από το μοντέλο, καθώς και πως συνδυάζονται για την παραγωγή του χάρτη σημαντικότητας.



(α΄) (β΄) (γ΄)

Σχήμα 3.1: Απόσπασμα μουσικής από την ταινία Departed. Από αριστερά προς τα δεξιά φαίνονται οι χάρτες: ενέργειας, συχνοτικής αντίθεσης, και χρονικής αντίθεσης.



(α΄) (β΄) (γ΄)

Σχήμα 3.2: Απόσπασμα ομιλίας από την ταινία Lord of the Rings με την φράση "you're in the service of the steward now". Από αριστερά προς τα δεξιά φαίνονται οι χάρτες: ενέργειας, συχνοτικής αντίθεσης, και χρονικής αντίθεσης.

3.1.1 Χάρτες χαρακτηριστικών

Στο μοντέλο των Kayser et al, που περιγράφηκε στην ενότητα 2, εξάγονται τρεις χάρτες χαρακτηριστικών που ο κάθε ένας έχει στόχο να τονίσει μία διαφορετική ιδιότητα του ηχητικού περιβάλλοντος. Γίνεται υπολογισμός των χαρτών ενέργειας, χρονικής αντίθεσης, και συχνοτικής αντίθεσης.

Ο χάρτης ενέργειας υπολογίζεται από το μοντέλο φιλτράροντας το ακουστικό φάσμα με το φίλτρο ενέργειας. Έχει τονισμένα τα σημεία που το σήμα έχει υψηλή ενέργεια, και στόχος του είναι να ανιχνεύσει ηχητικά γεγονότα με μεγάλη ένταση. Στα Σχήματα 3.1α΄, 3.2α΄ φαίνονται οι χάρτες ενέργειας δύο σκηνών της βάσης COGNIMUSE που περιέχουν μουσική και ομιλία, αντίστοιχα. Στην περίπτωση της μουσικής υψηλές τιμές παρατηρούνται στον χάρτη σε όλο το εύρος συχνοτήτων όπου και κατανέμεται συνήθως η ενέργεια μουσικών ερεθισμάτων. Αντίθετα, στην ομιλία υψηλές τιμές παρατηρούνται κυρίως στις χαμηλές συχνότητες, στα σημεία εμφάνισης των λέξεων και κυρίως των έμφωνων ήχων.

Με χρήση του φίλτρου συχνοτικής αντίθεσης υπολογίζεται ο χάρτης συχνοτικής αντίθεσης. Ο χάρτης έχει τονισμένα τα σημεία όπου κατά μήκος του συχνοτικού άξονα παρατηρείται κάποια μεταβολή, όπως για παράδειγμα κατά την εμφάνιση κάποιου σταθερού

τόνου. Στα Σχήματα 3.1β΄, 3.2β΄ φαίνονται χάρτες συχνοτικής αντίθεσης από σκηνές μουσικής και ομιλίας, αντίστοιχα. Ιδιαίτερα στην περίπτωση του χάρτη της μουσικής, είναι εμφανής η τάση διατήρησης οριζόντιων ακμών.

Τέλος, ο χάρτης χρονικής αντίθεσης προκύπτει από το φίλτρο χρονικής αντίθεσης. Έχει τονισμένα τα σημεία όπου κατά μήκος του χρονικού άξονα παρατηρείται μεταβολή του συχνοτικού περιεχομένου του σήματος. Με χρήση αυτού του χάρτη ανιχνεύονται τα σημεία εμφάνισης (onset) και παύσης (offset) των ηχητικών γεγονότων, τα οποία είναι υποψήφια να έλκουν την ανθρώπινη προσοχή καθώς παρατηρείται μεταβολή στις ιδιότητες του ηχητικού περιβάλλοντος με την εμφάνιση και παύση ηχητικών πηγών. Στα Σχήματα 3.1γ΄, 3.2γ΄ φαίνονται χάρτες χρονικής αντίθεσης από μουσική και ομιλία. Οι χάρτες έχουν πράγματι τονισμένες κάθετες στον χρονικό άξονα ακμές που αντιστοιχούν στο onset και offset των ήχων.

3.1.2 Χάρτης σημαντικότητας

Ο τελικός χάρτης σημαντικότητας από τον οποίο θα ταξινομούνται οι σκηνές ως σημαντικές ή μη, προκύπτει από τον συνδυασμό των επιμέρους χαρτών χαρακτηριστικών. Οι Kayser et al παρήγαγαν τον τελικό χάρτη αθροίζοντας τους χάρτες χαρακτηριστικών. Σε αυτή την εργασία εξετάζεται ο γραμμικός και μη-γραμμικός συνδυασμός των χαρτών. Στην πρώτη σειρά του Σχήματος 3.3, φαίνονται χάρτες σημαντικότητας για κάθε περίπτωση συνδυασμού.

Γραμμικός Συνδυασμός Χαρτών

Κατά τον γραμμικό συνδυασμό χαρτών, ο τελικός χάρτης σημαντικότητας S, προκύπτει ως εξής:

$$S = w_1 S_1 + w_2 S_2 + w_3 S_3, \tag{3.1}$$

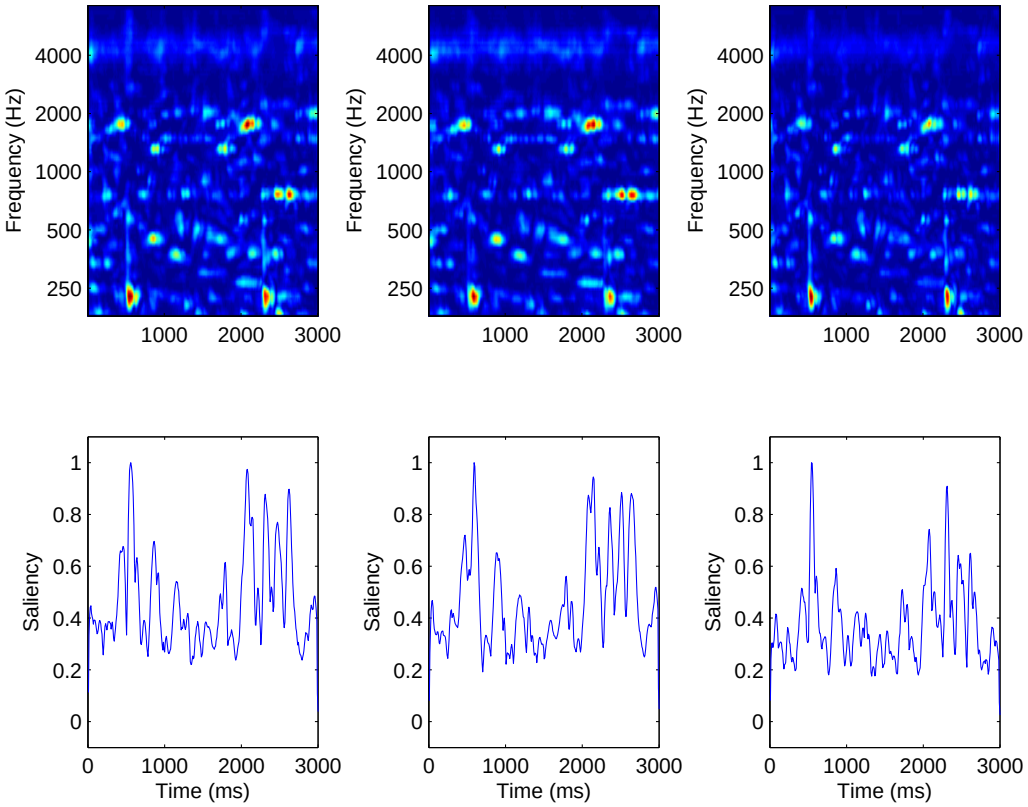
όπου S_i και w_i, $i = 1, 2, 3$, οι χάρτες χαρακτηριστικών και τα βάρη στάθμισης αντίστοιχα. Τα βάρη στάθμισης μπορούν να επιλεγούν ώστε να δοθεί μεγαλύτερη έμφαση σε κάποιο χαρακτηριστικό. Επιλέχθηκε τα βάρη να είναι ίσα με την μονάδα, $w_i = 1, i = 1, 2, 3$, ώστε τα αποτελέσματα να είναι άμεσα συγκρίσιμα με τους Kayser et al.

Συνδυασμός με βάρη αντίστροφα της διασποράς

Στην περίπτωση συνδυασμού των χαρτών με βάρη αντίστροφα της διασποράς, ο χάρτης σημαντικότητας δίνεται από την εξής σχέση:

$$S = \sum_{i=1}^{3} \frac{1}{Var(S_i)} S_i, \tag{3.2}$$

όπου $Var(S_i)$ η διακύμανση του χάρτη i, υπολογιζόμενη ως ο χάρτης να ήταν ένα διάνυσμα με συνιστώσες όλα τα στοιχεία του. Χάρτες με μεγάλη διακύμανση σταθμίζονται με μικρά βάρη ενώ χάρτες με μικρή διακύμανση με μεγάλα βάρη.



Σχήμα 3.3: Στην πρώτη σειρά φαίνονται χάρτες σημαντικότητας όπως υπολογίστηκαν από τους χάρτες του Σχήματος 3.1 με γραμμικό συνδυασμό, βάρη αντίστροφα της τυπικής απόκλισης, και μη γραμμικά, από αριστερά προς τα δεξιά. Στην δεύτερη σειρά οι αντίστοιχες καμπύλες σημαντικότητας με εφαρμογή της μέγιστης τιμής κατά μήκος του συχνοτικού άξονα.

Μη-Γραμμικός Συνδυασμός Χαρτών

Κατά τον μη-γραμμικό συνδυασμό των χαρτών για την δημιουργία του χάρτη σημαντικότητας, η τιμή σε κάθε σημείο του χάρτη υπολογίζεται από την εφαρμογή μιας μη-γραμμικής συνάρτησης με είσοδο τα αντίστοιχα σημεία των χαρτών χαρακτηριστικών. Επειδή για την ύπαρξη σημαντικού γεγονότος, αρκεί αυτό να εμφανισθεί σε τουλάχιστον μία διάσταση, δηλαδή έναν χάρτη, επιλέχτηκε η συνάρτηση της μέγιστης τιμής για την παραγωγή του τελικού χάρτη σημαντικότητας. Το (i, j) στοιχείο του τελικού χάρτη δίνεται από την σχέση:

$$S(i,j) = \max\{S_1(i,j), S_2(i,j), S_3(i,j)\}. \tag{3.3}$$

3.2 Καμπύλη Σημαντικότητας

Από τον χάρτη σημαντικότητας μπορεί να εξαχθεί καμπύλη σημαντικότητας (*saliency curve*) της οποίας η τιμή να δείχνει την σημαντικότητα του ηχητικού σήματος κάθε χρονική στιγμή. Υψηλότερη τιμή της καμπύλης κάποια χρονική στιγμή αντιστοιχεί σε μεγαλύτερη πιθανότητα η χρονική στιγμή να έχει σημειωθεί ως σημαντική. Η καμπύλη μπορεί να υπολογισθεί από τον χάρτη εξαλείφοντας την διάσταση που δεν αντιστοιχεί στον χρόνο. Η εξάλειψη του συχνοτικού άξονα πρέπει να γίνει ώστε τις χρονικές στιγμές που εμφανίζονται υψηλές τιμές κατά μήκος του, να αντιστοιχιστούν σε υψηλές τιμές στην καμπύλη. Επιλέχθηκε η συνάρτηση της μέγιστης τιμής για την απεικόνιση κάθε λωρίδας κάθετης του χρονικού άξονα σε ένα σημείο. Επίσης, με αυτόν τον τρόπο διασφαλίζεται ότι ήχοι με πολύ μικρό εύρος συχνοτήτων που έλκουν την ανθρώπινη προσοχή (όπως ένα ημίτονο), θα επηρεάσουν την καμπύλη σημαντικότητας.

Στην δεύτερη σειρά του Σχήματος 3.3 φαίνονται οι καμπύλες σημαντικότητας των χαρτών που βρίσκονται στην πρώτη σειρά. Με χρήση της καμπύλης σημαντικότητας μπορεί να γίνει ταξινόμηση των ήχων σε σημαντικούς και μη.

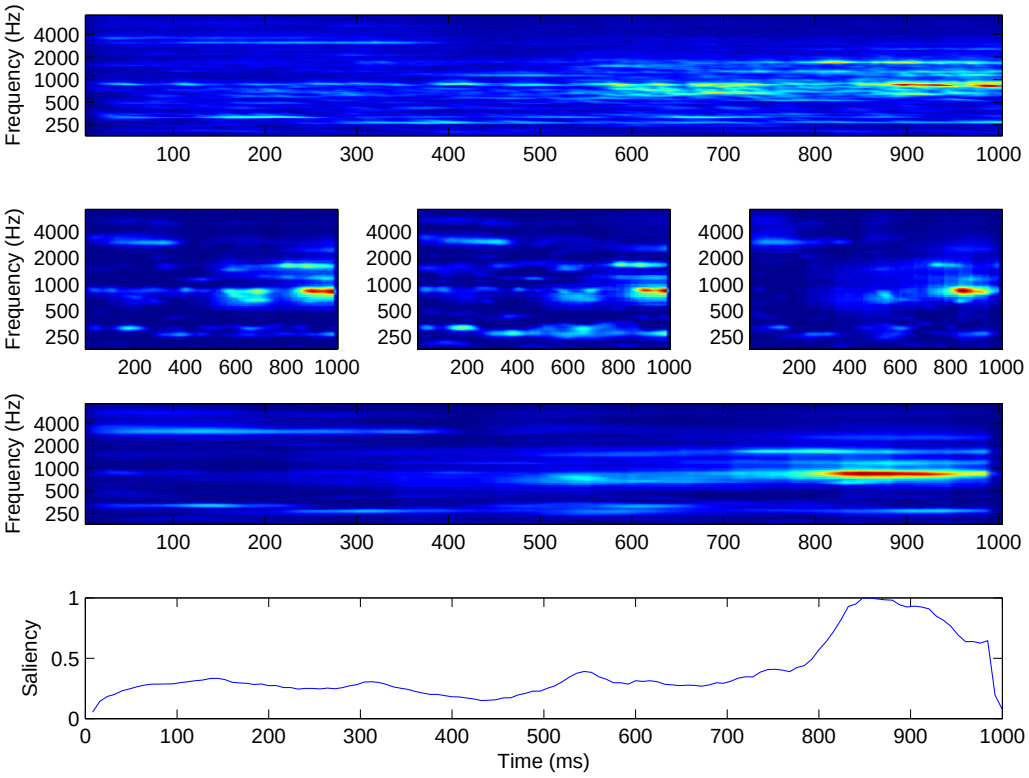
3.3 Μέθοδος Κατωφλίωσης και Ταξινόμηση

Ένας αρκετά απλός αλλά αποδοτικός τρόπος λήψης απόφασης για την σημαντικότητα κάθε σκηνής είναι μέσω κατωφλίωσης της καμπύλης σημαντικότητας. Στην μέθοδο κατωφλίωσης, επιλέγεται μια τιμή-κατώφλι και μόνο σημεία της καμπύλης που υπερβαίνουν αυτή την τιμή ταξινομούνται ως σημαντικά. Τα υπόλοιπα ταξινομούνται ως μη-σημαντικά.

Για τον υπολογισμό της καμπύλης σημαντικότητας το χρονικό σήμα κάθε ταινίας χωρίζεται σε μη-επικαλυπτόμενα παράθυρα διάρκειας ενός δευτερόλεπτου και σε κάθε παράθυρο υπολογίζεται ο χάρτης σημαντικότητας. Από τον χάρτη σημαντικότητας υπολογίζεται η καμπύλη σημαντικότητας με τον τρόπο που περιγράφηκε προηγουμένως. Η καμπύλη σημαντικότητας για ολόκληρη την ταινία προκύπτει από την συνένωση των επιμέρους καμπυλών. Τέλος απαιτείται η επιλογή ενός κατωφλίου για την ταξινόμηση των σκηνών. Η έξοδος κάθε σταδίου της διαδικασίας φαίνεται στο Σχήμα 3.4 για μία σκηνή από την ταινία Gladiator.

Αρχικά δοκιμάστηκε να γίνει ταξινόμηση των σκηνών με χρήση των χαρτών χαρακτηριστικών και όχι του χάρτη σημαντικότητας (δηλαδή θέτοντας ένα μόνο από τα βάρη ίσο με ένα, και τα υπόλοιπα μηδέν στο γραμμικό συνδυασμό) για να εξετασθεί η δυνατότητα κάθε ενός να ανιχνεύει σημαντικά γεγονότα που οφείλονται στο αντίστοιχο χαρακτηριστικό. Έπειτα ελέγχθηκε ο χάρτης σημαντικότητας όπως παράγεται από τους συνδυασμούς των επιμέρους χαρτών που περιγράφηκαν σε προηγούμενη ενότητα. Τα αποτελέσματα ταξινόμησης φαίνονται συγκεντρωτικά στον Πίνακα 3.1. Για κάθε τρόπο συνδυασμού φαίνεται ο μέσος όρος επιτυχίας για όλες τις ταινίες της βάσης σύμφωνα με τα μέτρα που χρησιμοποιούνται.

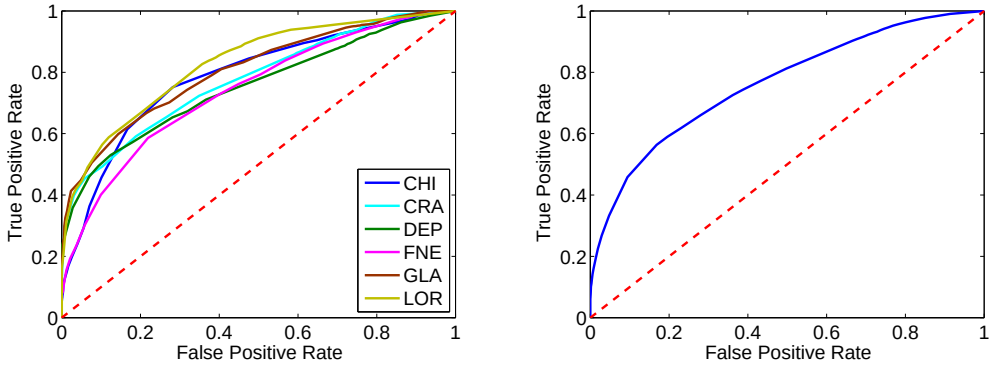
Τα αποτελέσματα ταξινόμησης είναι εφάμιλλα για όλους τους συνδυασμούς χαρτών, τόσο χρησιμοποιώντας μόνο κάθε χάρτη χαρακτηριστικών όσο και συνδυάζοντας τους.



Σχήμα 3.4: Στάδια εξαγωγής καμπύλης σημαντικότητας από τους χάρτες. Από πάνω προς τα κάτω και αριστερά προς τα δεξιά σε κάθε γραμμή: φασματογράφημα, χάρτες ενέργειας, συχνοτικής και χρονικής αντίθεσης, χάρτης σημαντικότητας, καμπύλη σημαντικότητας.

Πίνακας 3.1: Αποτελέσματα ταξινόμησης στην βάση COGNIMUSE με χρήση μοντέλου Kayser και κατωφλίωσης καμπύλης σημαντικότητας.

Map	Accuracy	Precision	Recall
EO	68.07	71.40	71.96
ESI	68.23	71.31	72.38
EPI	64.90	68.50	72.18
LIN	66.28	70.94	69.33
VAR	64.08	67.13	70.19
MAX	65.94	70.87	68.97



(α΄) (β΄)

Σχήμα 3.5: ROC καμπύλες για κάθε ταινία ξεχωριστά και ο μέσος όρος τους με κατωφλίωση της καμπύλης σημαντικότητας όπως υπολογίστηκε από τον χάρτη σημαντικότητας.

Οι χάρτες χρονικής και συχνοτικής αντίθεσης δεν συνεισέφεραν στην απόδοση ταξινόμησης όταν συνδυάστηκαν με τον χάρτη ενέργειας. Αντίθετα, σύμφωνα με τα αποτελέσματα των Kayser et al [31] ο χάρτης ενέργειας δεν αρκεί για την επιτυχή ταξινόμηση των σκηνών. Όταν οι συγγραφείς χρησιμοποίησαν μόνο τον χάρτη ενέργειας παρατήρησαν πολύ μικρότερη συσχέτιση με την ανθρώπινη θεώρηση της σημαντικότητας σε σύγκριση με τον χάρτη γραμμικού συνδυασμού. Αίτια στα οποία πιθανώς οφείλεται αυτή η διαφορά είναι το είδος των ήχων στους οποίους εξετάζεται το μοντέλο, οι οποίοι είναι πολύ πιο σύνθετοι από αυτούς που χρησιμοποίησαν οι Kayser et al με συνέπεια την υψηλή συσχέτιση μεταξύ των χαρτών, καθώς και η επισημειώση της βάσης από τους χρήστες η οποία είναι επηρεασμένη και από παράγοντες που δεν σχετίζονται με τον ήχο (όπως σημασιολογία).

Η υψηλή συσχέτιση των χαρτών χαρακτηριστικών είναι εμφανής στα Σχήματα 3.1 και 3.2 για τις συγκεκριμένες σκηνές. Υψηλή συσχέτιση παρατηρήθηκε για κάθε ηχητική σκηνή της βάσης δεδομένων. Οι Kalinli & Narayanan [29] με υπολογισμό αμοιβαίας πληροφορίας μεταξύ των χαρτών κατέληξαν σε παρόμοια συμπεράσματα όσον αφορά την συσχέτιση μεταξύ των χαρτών.

Για την περίπτωση γραμμικού συνδυασμού των χαρτών έγινε εξαγωγή της καμπύλης ROC για κάθε ταινία, η οποία φαίνεται στο Σχήμα 3.5. Στο ίδιο σχήμα φαίνεται και η καμπύλη του μέσου όρου. Το εμβαδόν του χωρίου μεταξύ κάθε καμπύλης και της διαγωνίου φαίνεται στον Πίνακα 3.2, όπου επίσης φαίνονται τα μέτρα accuracy, precision, και recall για κάθε ταινία της βάσης για μία τιμή κατωφλίου. Για το ίδιο κατώφλι, ο αλγόριθμος ταξινομεί κατά μέσο όρο το 48.11% κάθε ταινίας ως σημαντικό, ενώ οι χρήστες το 52.81% και 34.45% εάν απαιτείται η συμφωνία τουλάχιστον δύο ή τριών χρηστών αντίστοιχα, για την θεώρηση ενός γεγονότος ως σημαντικού. Το πλήθος σημείων που ταξινομείται σημαντικό από τον αλγόριθμο είναι κοντά στο πλήθος που ταξινομούν οι χρήστες ως σημαντικό.

Πίνακας 3.2: Απόδοση του χάρτη σημαντικότητας με μέθοδο κατωφλίωσης στην βάση COGNIMUSE.

Movie	Acc	Prec	Rec	AUC
CHI	65.15	63.51	93.24	0.2850
CRA	68.88	73.80	69.29	0.2748
DEP	63.10	43.19	73.17	0.2527
FNE	66.40	67.14	72.89	0.2410
GLA	71.76	77.16	75.17	0.3050
LOR	63.46	95.46	39.54	0.3276
Average	66.28	70.94	69.33	0.2636

3.4 Η έννοια του Gist μιας σκηνής

Από μελέτες σχετικά με την αντίληψη οπτικών σκηνών [48], έχει παρατηρηθεί πως ο άνθρωπος έχει την ικανότητα να αναγνωρίζει ορισμένα βασικά στοιχεία που συνθέτουν μια σκηνή όταν έλθει σε επαφή με αυτήν για διάστημα μικρότερο των 100 ms, οσοδήποτε περίπλοκη και εάν είναι αυτή. Τα στοιχεία αυτά συμπεριλαμβάνουν την κατηγορία μιας σκηνής (π.χ. μία οδός ή ένα δάσος), την χωρική διάταξη της σκηνής (π.χ. μια οδός με κάθετα τετράγωνα), και πιθανώς την ταυτότητα ορισμένων αντικειμένων που εμφανίζονται σε αυτή. Τα χαρακτηριστικά αυτά αναφέρονται συχνά με τον όρο *gist* μιας σκηνής, και είναι αυτά που ο ανθρώπινος εγκέφαλος χρησιμοποιεί για την αναπαράσταση των σκηνών. Το gist μιας σκηνής περιλαμβάνει τόσο χαρακτηριστικά χαμηλού επιπέδου για την αντιληπτική αναπαράσταση των σκηνών, όσο και υψηλότερου επιπέδου για την αναπαράσταση και κατανόηση του νοηματικού περιεχομένου μιας σκηνής.

Λόγω της ενσωμάτωσης στον όρο gist διαφόρων σταδίων επεξεργασίας, συνήθως αναφερόμαστε σε *αντιληπτικό gist* (*perceptual gist*), και *νοηματικό gist* (*conceptual gist*). Το νοηματικό gist αναφέρεται στα χαρακτηριστικά που εξάγονται από την σκηνή για την αναπαράσταση σημασιολογικής πληροφορίας. Σε μία μελέτη [51], παρουσίασαν σε ανθρώπους εικόνες για διάστημα 100 ms. Όταν υπήρχε παύση μερικών δευτερολέπτων μεταξύ της παρουσίασης δύο διαδοχικών εικόνων, ήταν εύκολο για τα άτομα να τις θυμούνται. Αντίθετα όταν οι εικόνες παρουσιάζονταν συνεχόμενα η μία μετά την άλλη (μορφή βίντεο), με ρυθμό 125 ms για κάθε εικόνα, η απόδοση έπεφτε στο επίπεδο της τύχης δείχνοντας ότι πιθανώς η διαδικασία της αποθήκευσης μίας εικόνας στην μνήμη διακόπτεται για την επεξεργασία της επόμενης. Άλλες παρόμοιες μελέτες έδειξαν ότι 100 ms είναι αρκετά για την εξαγωγή του νοηματικού gist και κατανόηση της σκηνής, ωστόσο απαιτούνται μερικές ακόμα εκατοντάδες ms για την αναπαράσταση του στη μνήμη και την ύπαρξη δυνατότητας συμπερασμού του περιεχομένου της σκηνής.

Με τον όρο αντιληπτικό gist αναφερόμαστε στα χαρακτηριστικά χαμηλού επιπέδου που εξάγονται από μια σκηνή τη στιγμή της αντίληψής της, για την αναπαράσταση της στον ανθρώπινο εγκέφαλο. Τα χαρακτηριστικά αυτά εξαρτώνται και από άλλους παρά-

(α΄) (β΄)

Σχήμα 3.6: Παράδειγμα όπου η επεξεργασία των εικόνων γίνεται από γενική προς την λεπτομερή πληροφορία.

γοντες (πέραν των ιδιοτήτων της σκηνής), όπως από τον στόχο του παρατηρητή, και τον χρόνο θέασης της σκηνής. Για παράδειγμα, σε μια μελέτη [56] με στόχο την σύγκριση της πληροφορίας που μεταφέρουν οι ακμές (egdes) και οι ομοιόμορφες μάζες (blobs) μιας εικόνας, συνδύασαν την χαμηλή χωρική συχνότητα μιας εικόνας με την υψηλή χωρική συχνότητα μιας άλλης εικόνας (Σχήμα 3.6), και το αποτέλεσμα το παρουσίασαν σε ανθρώπους. Όταν η παρουσίαση κάθε εικόνας διαρκούσε 30 ms, τα άτομα απαντούσαν ότι είδαν την εικόνα από την οποία λήφθηκε η χαμηλή συχνότητα. Αντίθετα, όταν η παρουσίαση της εικόνας διαρκούσε 150 ms, τα άτομα επέλεγαν την εικόνα οπό την οποία προήλθε η υψηλή συχνότητα. Υπήρχε, επομένως μια ροή επεξεργασίας από την αδρή προς την πιο λεπτομερή πληροφορία, και αντίστοιχη αναπαράσταση στον εγκέφαλο.

Σε παρόμοια μελέτη για την εξέταση του ρόλου του χρώματος στην αντίληψη μιας σκηνής, παρουσίασαν σε άτομα εικόνες για 150 ms με είτε το πραγματικό τους χρώμα είτε τροποποιώντας το (Σχήμα 3.7), και τους ζήτησαν να τις ταξινομήσουν (π.χ. οδός, κουζίνα, δάσος) όσο πιο γρήγορα μπορούσαν. Όταν το χρώμα είναι διαγνωστικό χαρακτηριστικό για την κατηγορία της σκηνής (όπως σε σκηνές με φυσικά τοπία), η προσθήκη πραγματικού χρώματος σε γκρίζα εικόνα επιταχύνει την αναγνώριση της, ενώ η προσθήκη μη-πραγματικού την επιβραδύνει. Η ταχύτητα αναγνώρισης δεν επηρεάζεται όταν το χρώμα δεν είναι διαγνωστικό χαρακτηριστικό (όπως σκηνές με περιβάλλοντα φτιαγμένα από ανθρώπους).

Με βάση τα ανωτέρω, συμπεραίνουμε ότι πολύ σύντομα ($\approx$100 ms) πραγματοποιείται η εξαγωγή κάποιων αδρών χαρακτηριστικών από τις σκηνές, ικανών να οδηγήσουν στην αναγνώριση τους, αν και η ακριβής ταυτότητα των αντικειμένων που παρουσιάζονται σε αυτές ίσως δεν έχει αποκαλυφθεί. Παρόμοιες συμπεράσματα έχουν εξαχθεί και για ηχητι-

(α΄) (β΄)

Σχήμα 3.7: Αριστερά: παράδειγμα περιβάλλοντος (φτιαγμένο από ανθρώπους) που η ταχύτητα αναγνώρισης του δεν επηρεάζεται με τροποποίηση του χρώματος. Δεξιά: φυσικό τοπίο για το οποίο υπάρχει επιρροή του χρώματος στην ταχύτητα αναγνώρισης του.

κές σκηνές, τα οποία θα εκμεταλλευτούμε για την δόμηση διανύσματος χαρακτηριστικών για τον χαρακτηρισμό τους.

3.5 Το Gist στις ηχητικές σκηνές

Σε αναλογία με την όραση, υπάρχουν ενδείξεις ότι το gist εμφανίζεται και στην ανθρώπινη ακοή [21]. Πριν κατευθυνθεί η προσοχή μας σε μία ηχητική σκηνή για την λεπτομερή επεξεργασία του ηχητικού σήματος, η ηχητική σκηνή επεξεργάζεται ως ένα σύνολο και όχι κάθε πηγή που την απαρτίζει. Για παράδειγμα κατά την είσοδο σε ένα δωμάτιο γεμάτο με άτομα που μιλάνε, αρχικά ακούγεται ένα βουητό και όχι κάποια συγκεκριμένη συζήτηση. Πρώτα ακούμε μια ορχήστρα στο σύνολο της και έπειτα τα όργανα που την απαρτίζουν, εάν επικεντρώσουμε σε κάθε ένα από αυτά. Σε πειράματα με εναλλασσόμενους τόνους, τα άτομα δήλωναν ότι αρχικά αντιλαμβάνονταν ένα ηχητικό ρεύμα (*stream*), αλλά έπειτα από λίγα δευτερόλεπτα συγκέντρωσης στην σκηνή άκουγαν δύο ρεύματα [13, 9]. Πειράματα έχουν δείξει ότι το ηχόχρωμα (*timbre*) προσδιορίζεται πριν το pitch [53, 52], με συνέπεια πρώτα να γίνεται αντιληπτή η μορφή του συχνοτικού φάσματος (κατανομή συχνοτήτων), και έπειτα πιο συγκεκριμένα οι αρμονικές που παράγουν την αίσθηση του pitch. Η ίδια μελέτη έδειξε ότι μικρής διάρκειας ακουστικά ερεθίσματα μπορούν να ταξινομηθούν γρήγορα με βάση το ηχόχρωμα τους, όπως για παράδειγμα τα φωνήεντα ή οι ήχοι από μουσικά όργανα με διάρκεια μικρότερη των 10 ms.

3.6 Εξαγωγή Gist διανύσματος

Υποθέτοντας, με βάση τα ανωτέρω, ότι μερικές δεκάδες msec είναι αρκετά για την αντίληψη ορισμένων χαρακτηριστικών του σήματος, όπως η κατηγορία της σκηνής, το φύλο του ομιλητή, θα εξάγουμε διάνυσμα χαρακτηριστικών από τον χάρτη σημαντικότητας χρησιμοποιώντας τοπική πληροφορία του χάρτη για τον υπολογισμό κάθε συνιστώ-

σας του διανύσματος. Πιο συγκεκριμένα, διαμερίζουμε τον χάρτη σε ορθογώνιο πλέγμα $m \times n$ (συχνότητα $\times$ χρόνος). Κάθε κελί του πλέγματος είναι ο χάρτης σημαντικότητας του σήματος στην περιοχή συχνοτήτων και χρονική διάρκεια που καθορίζει το πλέγμα, και συλλαμβάνει τις μεταβολές των χαρακτηριστικών στην περιοχή. Ο χάρτης σημαντικότητας έχει τονίσει σημεία υποψήφια να έλξουν την ανθρώπινη προσοχή. Υπολογίζοντας την τιμή μιας συνάρτησης f, σε κάθε κελί του χάρτη και θεωρώντας το διάνυσμα με συνιστώσες αυτές τις τιμές, αναμένουμε από χάρτες σημαντικών σκηνών να προκύψουν διανύσματα με κάποιες από τις συνιστώσες τους να ξεχωρίζουν από τις υπόλοιπες, με κατάλληλη επιλογή της f. Αντίθετα, στις μη-σημαντικές σκηνές τα διανύσματα θα είναι ομοιόμορφα.

Οι συναρτήσεις f που δοκιμάστηκαν είναι η μέση και η μέγιστη τιμή του χάρτη σε κάθε κελί:

$$f(i) = \frac{1}{|(j,k) \in B_i|} \sum_{(j,k) \in B_i} S(j,k), \quad i = 1, 2, \ldots, m \cdot n \tag{3.4}$$

$$f(i) = max_{(j,k) \in B_i} S(j,k), \quad i = 1, 2, \ldots, m \cdot n \tag{3.5}$$

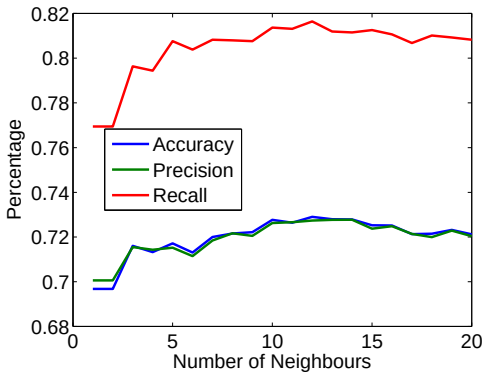
όπου B_i το κελί i, και S ο χάρτης σημαντικότητας. Τα πλέγματα που χρησιμοποιήθηκαν ήταν 4×20, και 1×125 (πλήθος στηλών του χάρτη).

Αναπαριστώντας κάθε σκηνή με ένα διάνυσμα, ταξινομείται ολόκληρη η σκηνή ως σημαντική ή όχι. Επειδή η επισημείωση των χρηστών έχει γίνει σε επίπεδο καρέ εικόνας, και κάθε σκηνή περιέχει περισσότερα του ενός καρέ, πρέπει να αποδοθεί σημαντικότητα στις σκηνές με βάση τη σημαντικότητα των καρέ που περιέχουν. Θεωρούμε ότι μία σκηνή είναι σημαντική εάν τουλάχιστον 300 ms (8 καρέ εικόνας) της σκηνής έχουν σημειωθεί σημαντικά. Για σκηνές διάρκειας 1 sec, τα 300 ms αντιστοιχούν στο 30% της διάρκειας της σκηνής (8 από τα 25 καρέ). Αυτή η επιλογή έγινε διότι παρατηρήθηκε στην βάση δεδομένων πως η ελάχιστη διάρκεια των (σημαντικών) ηχητικών γεγονότων είναι περίπου 300 ms, και θεωρούμε πως μία σκηνή είναι σημαντική εάν τουλάχιστον μία φορά έλξει την ανθρώπινη προσοχή, δηλαδή περιέχει τουλάχιστον ένα σημαντικό γεγονός.

Για την ταξινόμηση των σκηνών χρησιμοποιήθηκε ο αλγόριθμος kNN (*k-Nearest Neighbours*) με χρήση Ευκλείδιας απόστασης. Στο αποτέλεσμα της ταξινόμησης πραγματοποιείται φιλτράρισμα μέσου (median) πέντε σημείων για εξομάλυνση. Στο Σχήμα 3.8 φαίνεται τα μέσα accuracy, precision, και recall με χρήση της συνάρτησης του μεγίστου και πλέγματος 4×20 για την εξαγωγή του gist διανύσματος. Στον Πίνακα 3.3 φαίνονται αναλυτικά τα ποσοστά επιτυχίας για κάθε ταινία για την τιμή του k που επιτεύχθηκε το μέγιστο μέσο recall για κάθε πλέγμα. Τα αποτελέσματα ταξινόμησης με χρήση της συνάρτησης μέσης τιμής ήταν εφάμιλλα με αυτά της μέγιστης.

3.6.1 Εφαρμογή PCA στα gist διανύσματα

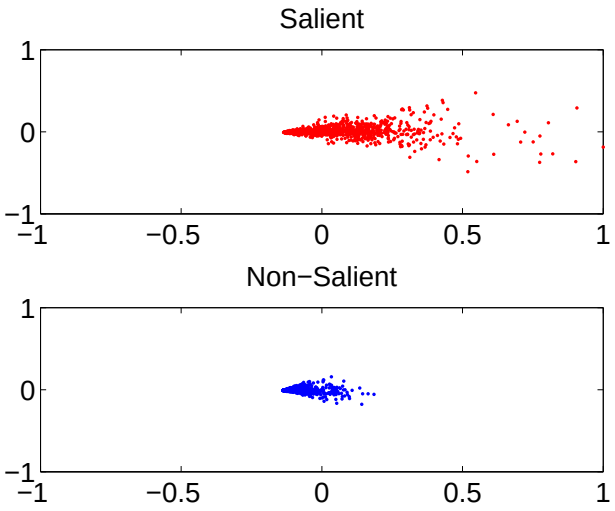
Στα gist διανύσματα εφαρμόστηκε ανάλυση σε πρωτεύουσες συνιστώσες (PCA) διατηρώντας τις πρώτες δύο κύριες συνιστώσες για οπτικοποίηση του χώρου που κατανέμονται. Στο Σχήμα 3.9 φαίνονται τα διανύσματα για πλέγμα 4×20 και τη συνάρτηση



Σχήμα 3.8: Μέσο accuracy, precision και recall για τις ταινίες της βάσης με χρήση gist διανύσματος και kNN αλγορίθμου, ως συνάρτηση του αριθμού των κοντινότερων k γειτόνων.

Πίνακας 3.3: Ποσοστά επιτυχίας με χρήση gist διανύσματος και kNN αλγορίθμου με $k = 12$ και $k = 17$ για 4×20 και 1×125 grid αντίστοιχα.

	4×20 Grid			1×125 Grid		
Movie	Acc	Prec	Rec	Acc	Prec	Rec
CHI	71.29	70.77	88.63	68.03	66.94	92.05
CRA	79.71	82.99	81.73	77.52	78.53	84.22
DEP	70.13	53.48	71.43	63.29	46.45	75.08
FNE	66.26	64.10	91.52	66.59	63.72	94.83
GLA	73.53	72.97	90.22	72.25	70.83	93.03
LOR	76.48	89.45	69.44	76.17	86.54	71.93
Average	72.90	72.74	81.64	70.72	69.31	84.66



Σχήμα 3.9: Σημεία σε 2 διαστάσεις έπειτα από την εφαρμογή PCA σε gist διανύσματα της ταινίας Gladiator.

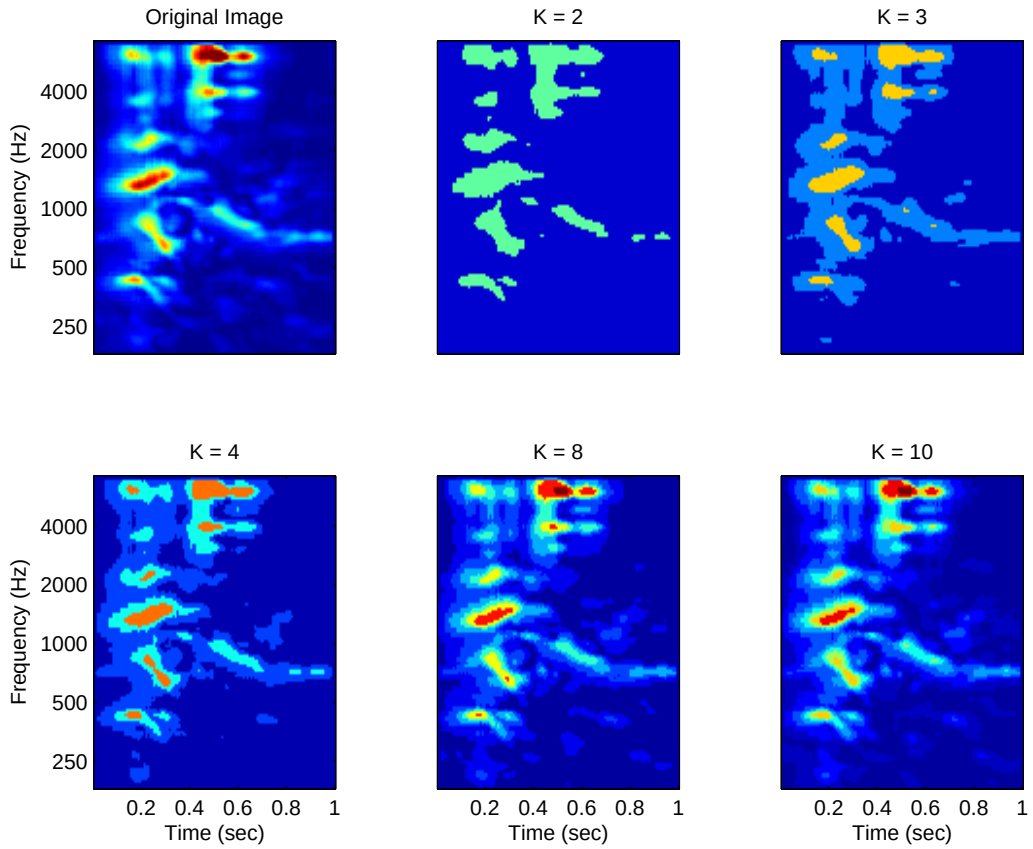
μεγίστου, από την ταινία Gladiator. Τα σημεία που αντιστοιχούν στα μη-σημαντικά γεγονότα (με βάση τη σημείωση των χρηστών) είναι υποσύνολο των σημείων των σημαντικών γεγονότων. Παρόμοιες εικόνες λήφθηκαν και για τις υπόλοιπες ταινίες της βάσης. Πραγματοποιήθηκε ταξινόμηση των δισδιάστατων διανυσμάτων με χρήση του αλγορίθμου kNN με 15 γείτονες, και τα μέσα accuracy, precision και recall που επιτεύχθηκαν είναι 64.72, 73.17 και 69.85% αντίστοιχα.

3.6.2 Κατάτμηση χάρτη με K-Means

Οι χάρτες σημαντικότητας έχουν σε αρκετές περιπτώσεις ομοιόμορφες περιοχές με μεγάλο εμβαδόν από τις οποίες εξάγονται συνιστώσες με ίδιες τιμές για το gist διάνυσμα, προκαλώντας αύξηση της διάστασης των διανυσμάτων, ενώ η πληροφορία θα μπορούσε πιθανώς να διατηρηθεί σε λιγότερες συνιστώσες. Για παράδειγμα, το υπόβαθρο των χαρτών έχει σχεδόν μηδενική τιμή και μεγάλο εμβαδόν στους περισσότερους χάρτες με συνέπεια πολλές από τις συνιστώσες να έχουν πολύ μικρή τιμή. Επιπλέον, για την ταξινόμηση ενός χάρτη ως σημαντικού αρκεί να υπάρχει μια περιοχή με υψηλές τιμές.

Για την μείωση της διάστασης των διανυσμάτων, δοκιμάστηκε η κατάτμηση του χάρτη σε περιοχές με βάση τις τιμές τους και υπολογισμό της εξόδου της συνάρτησης f με είσοδο αυτές τις περιοχές. Οι περιοχές αυτές έχουν αυθαίρετο σχήμα, το οποίο εξαρτάται από την κατανομή των τιμών στον χάρτη, και είναι μη-συνεκτικές. Στόχος είναι σημεία του χάρτη που έχουν "κοντινές" τιμές να ανήκουν στην ίδια περιοχή, ώστε να γίνουν πιο διακριτές οι συνιστώσες του διανύσματος.

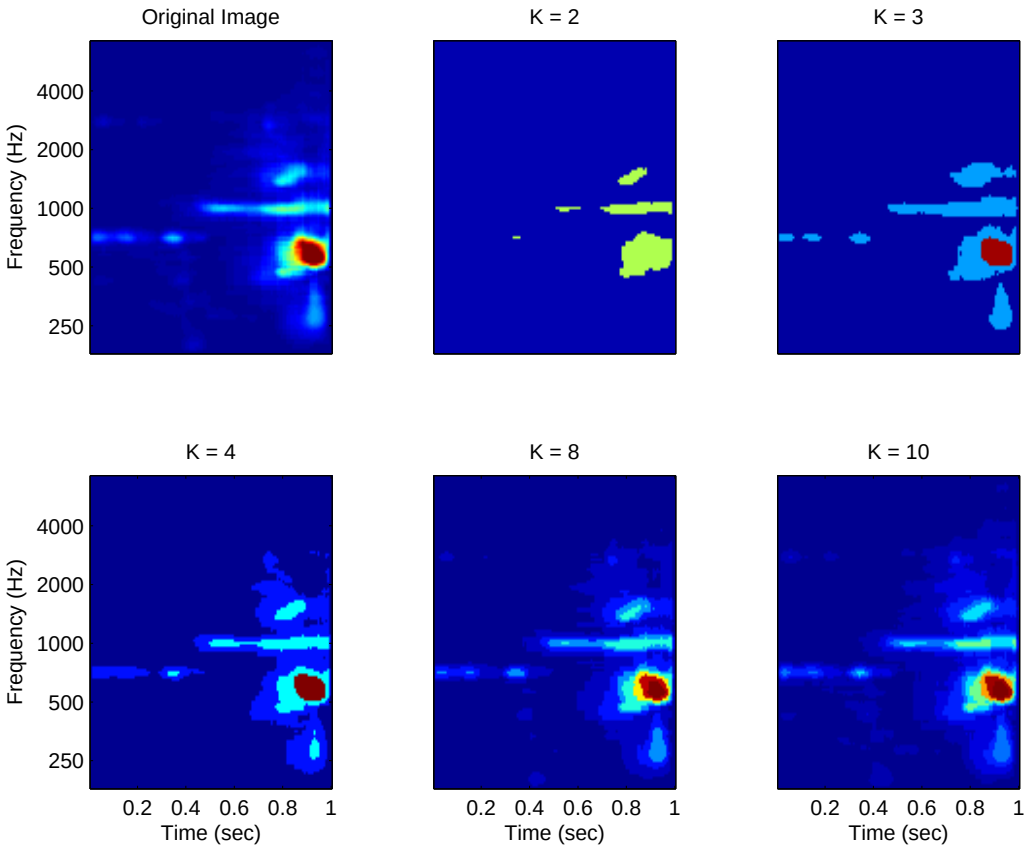
Για την κατάτμηση του χάρτη χρησιμοποιήθηκε ο αλγόριθμος K-Means με Ευκλείδια απόσταση. Κάθε ομάδα (*cluster*) που σχηματίζεται από τον αλγόριθμο ορίζει και μια



Σχήμα 3.10: Κατάτμηση χάρτη σημαντικότητας με χρήση K-Means αλγορίθμου. Πάνω από κάθε χάρτη φαίνεται το πλήθος των clusters του αλγορίθμου, ενώ κάθε περιοχή αναπαρίσταται με το κέντρο του cluster στο οποίο ανήκει.

περιοχή της κατατμημένης εικόνας που αποτελείται από τα σημεία που ανήκουν σε αυτό, ενώ το κέντρο του ισούται με την μέση τιμή των pixels της εικόνας σε εκείνη την περιοχή. Στα Σχήματα 3.10, 3.11 φαίνεται το αποτέλεσμα της κατάτμησης χαρτών που περιέχουν σημαντικά γεγονότα (με βάση τη σημείωση των χρηστών), για διάφορες τιμές της παραμέτρου K. Κάθε περιοχή αναπαρίσταται με το κέντρο του cluster στο οποίο ανήκει.

Το διάνυσμα χαρακτηριστικών για κάθε σκηνή, έχει συνιστώσες που δίνονται από την έξοδο της συνάρτησης f στα στοιχεία κάθε cluster διατεταγμένες σε αύξουσα σειρά. Χρησιμοποιώντας την συνάρτηση της μέγιστης τιμής σε κάθε περιοχή έγινε έλεγχος της ικανότητας του διανύσματος χαρακτηριστικών να διαχωρίζει σημαντικές από μη-σημαντικές σκηνές. Δοκιμάστηκε πλήθος cluster (μήκος διανύσματος) από δύο έως δώδεκα. Τα ποσοστά επιτυχίας υπολείπονταν αρκετά αυτών που επιτυγχάνονται με το διάνυσμα που εξάγεται με χρήση πλέγματος, και μεταβάλλονταν ελάχιστα ως προς το πλήθος των clusters. Στον Πίνακα 3.4 φαίνονται τα μέσα accuracy, precision και recall που επιτεύχθηκαν με διανύσματα που στηρίζονται στην έννοια του gist. Η ακολουθία των τιμών του χάρτη, που



Σχήμα 3.11: Κατάτμηση χάρτη σημαντικότητας με χρήση K-Means αλγορίθμου. Πάνω από κάθε χάρτη φαίνεται το πλήθος των clusters του αλγορίθμου.

Πίνακας 3.4: Απόδοση στην βάση COGNIMUSE του gist διανύσματος. FULL: διάνυσμα με χρήση τετραγωνικού πλέγματος, PCA-i: διατήρηση i πρωτευουσών συνιστωσών από το FULL, SEGM-k: διάνυσμα με κατάτμηση του χάρτη με k-Means.

	Dims	Accuracy(%)	Precision(%)	Recall(%)
FULL	80	72.9	72.7	81.6
PCA-2	2	64.7	73.2	69.9
SEGM-9	9	65.3	72.6	67.7

συλλαμβάνεται από το gist διάνυσμα με χρήση πλέγματος, φαίνεται πως είναι σημαντική για την κατηγοριοποίηση των σκηνών ή λόγω της μεγάλης διάστασης των διανυσμάτων η ταξινόμηση διευκολύνεται χωρίς τα διανύσματα να περιέχουν κάποια πληροφορία για την ταξινόμηση των σκηνών.

Κεφάλαιο 4

Χρονικός Χάρτης Σημαντικότητας

Εισαγωγή

Εμπνευσμένοι από το μοντέλο των Kayser el al [31] εφαρμόζουμε ένα μοντέλο όπου αντί για διδιάστατους χάρτες στηρίζεται ολοκληρωτικά σε μονοδιάστατα χαρακτηριστικά. Οι διεργασίες σε κάθε στάδιο είναι παρόμοιες με αυτές των Kayser et al, αλλά προσαρμοσμένες σε μία διάσταση. Με είσοδο το ηχητικό σήμα, η έξοδος του μοντέλου είναι μία καμπύλη σημαντικότητας. Το μοντέλο ονομάσθηκε χρονικός χάρτης σημαντικότητας διότι από τον χρονο-συχνοτικό χάρτη των Kayser et al, διατηρείται μόνο η διάσταση του χρόνου και εξετάζεται πως τα χαρακτηριστικά μεταβάλλονται κατά μήκος αυτής.

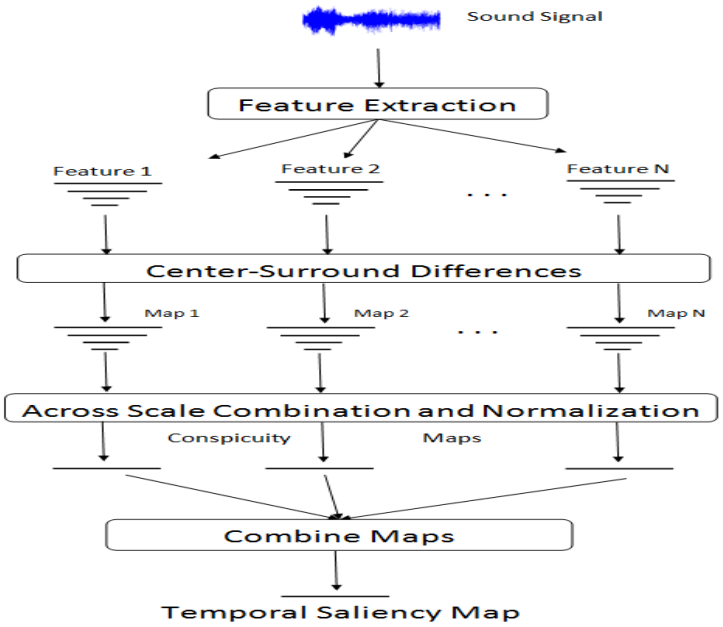
4.1 Στάδια του Μοντέλου

Το διάγραμμα ροής του μοντέλου φαίνεται στο Σχήμα 4.1. Τα στάδια του μοντέλου είναι η πολυ-κλιμακωτή εξαγωγή χαρακτηριστικών, διαφορές κέντρου-περίγυρου μεταξύ κλιμάκων, κανονικοποίηση καμπυλών, και συνδυασμός τους για υπολογισμό της τελικής καμπύλης σημαντικότητας. Περιγραφή τους γίνεται στις επόμενες ενότητες.

Στόχος του μοντέλου είναι η ανίχνευση σημείων στα χαρακτηριστικά που εξάγονται, τα οποία διαφέρουν από τα γειτονικά τους. Δηλαδή, περιοχές που παρατηρούνται μεγάλες μεταβολές στα χαρακτηριστικά. Ως γειτονικά ενός σημείου, θεωρούνται τα σημεία τα οποία απέχουν έως μια απόσταση από αυτό η οποία μπορεί να δοθεί ως είσοδος στον αλγόριθμο.

4.1.1 Εξαγωγή Χαρακτηριστικών

Στο πρώτο στάδιο του μοντέλου πραγματοποιείται η εξαγωγή κάποιων μονοδιάστατων χαρακτηριστικών από το ηχητικό σήμα. Τα χαρακτηριστικά αυτά θα είναι οι υποψήφιες διαστάσεις στις οποίες θα εξετασθούν μεταβολές, και κάποια από αυτά έχουν επιλεγεί διότι υπάρχουν ενδείξεις ότι επηρεάζουν την κατεύθυνση της ανθρώπινης προσοχής, ενώ άλλα για πειραματικούς σκοπούς. Τα χαρακτηριστικά που χρησιμοποιούνται είναι η



Σχήμα 4.1: Διάγραμμα ροής του μοντέλου εξαγωγής χρονικού χάρτη σημαντικότητας.

ενέργεια βραχέως χρόνου, το loudness, το roughness, και η fractal διάσταση. Ο τρόπος υπολογισμού τους, περιγράφεται αναλυτικά σε επόμενη ενότητα.

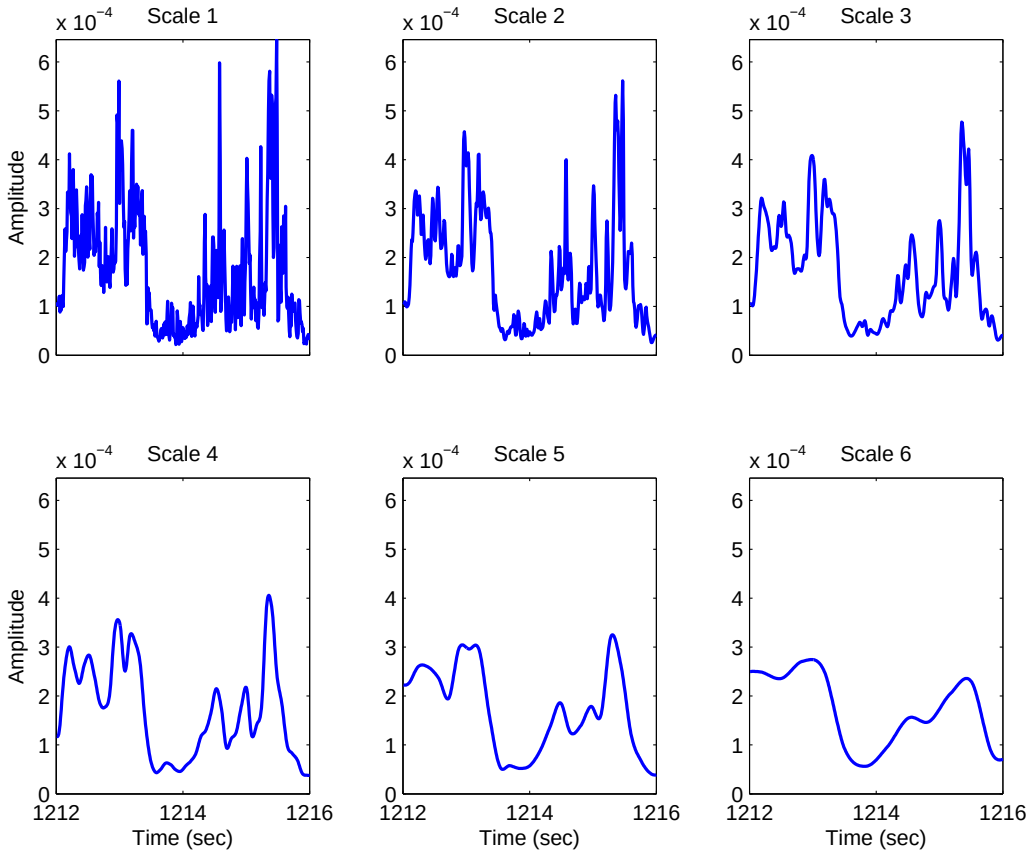
4.1.2 Πολυκλιμακωτή Ανάλυση

Η πολυκλιμακωτή ανάλυση επιτρέπει να εξετασθούν αλληλεπιδράσεις μεταξύ μη γειτονικών σημείων του σήματος και να θεωρηθεί μια γενικότερη μορφή του (τρόπος που μεταβάλλεται). Είναι ως να εξετάζονται όχι μόνο τοπικά οι διακυμάνσεις του αλλά σε ευρύτερο χρονικό πλαίσιο. Για την παραγωγή κάθε κλίμακας, αρχικά φιλτράρεται η προηγούμενη κλίμακα με Gaussian πυρήνα πέντε σημείων, $g = [1,\ 5,\ 10,\ 5,\ 1]/22$, για εξομάλυνση. Στην συνέχεια υπο-δειγματοληπτείται με παράγοντα 2 (η προηγούμενη κλίμακα) ώστε σημεία τα οποία απέχουν απόσταση 2 να γίνουν γειτονικά. Ως αρχική κλίμακα (κλίμακα 1) θεωρείται το αρχικό σήμα.

Γίνεται χρήση $N = 7$ κλιμάκων, το οποίο συνεπάγεται δειγματοληψία του αρχικού σήματος με παράγοντα από $1/2^0$ έως $1/2^{N-1}$. Με χρήση N κλιμάκων καθίσταται δυνατή η εξέταση μεταβολών στο σήμα μεταξύ σημείων που απέχουν έως 2^{N-1} δείγματα. Στο Σχήμα 4.2 φαίνονται 6 κλίμακες ενός εκ των χαρακτηριστικών που εξάγονται.

4.1.3 Διαφορές Κέντρου-Περίγυρου

Αυτό το στάδιο του μοντέλου στοχεύει στην διατήρηση των σημείων που το σήμα μεταβάλλεται ή λαμβάνει μέγιστη τιμή και την αποκοπή ομοιόμορφων περιοχών και κοι-



Σχήμα 4.2: Πολλαπλές κλίμακες του χαρακτηριστικού της ενέργειας από απόσπασμα της ταινίας Gladiator.

λάδων. Για τον σκοπό αυτό, λαμβάνονται διαφορές μεταξύ των κλιμάκων για κάθε χαρακτηριστικό που υπολογίσθηκαν στο προηγούμενο στάδιο. Για λήψη των διαφορών πραγματοποιείται σε όλες τις κλίμακες παρεμβολή ώστε να έχουν ίδιο μήκος με την αρχική κλίμακα (κλίμακα 1), και λαμβάνονται διαφορές σημείο προς σημείο. Διαφορές λαμβάνονται μεταξύ των κλιμάκων $c = \{2, 3, 4\}$ (κλίμακες κέντρου), και $s = c + d,\ d = \{2, 3\}$ (κλίμακες περίγυρου), όπου από τις κλίμακες κέντρου αφαιρούνται οι κλίμακες περίγυρου. Οι αρνητικές τιμές απεικονίζονται στο μηδέν. Λόγω της παρεμβολής που πραγματοποιείται, προσεγγίζεται η λήψη διαφορών μεταξύ σημείων του σήματος που απέχουν $2^d - 1$ δείγματα.

Η υπο-δειγματοληψία που πραγματοποιήθηκε κατά την παραγωγή των κλιμάκων, έχει ως επακόλουθο οι μεγαλύτερες κλίμακες να έχουν μικρότερη μέγιστη και μεγαλύτερη ελάχιστη τιμή από τις μικρότερες κλίμακες. Επομένως κατά την λήψη των διαφορών, οι κορυφές (που είναι υποψήφιες για έλξη της προσοχής) διατηρούνται, ενώ οι κοιλάδες τίθενται στο μηδέν. Ως συνέπεια αυτού, η ελάχιστη τιμή της εξόδου γίνεται ίση με μηδέν και πραγματοποιείται μια μετατόπιση των τιμών του σήματος τοπικά προς αυτή. Επίσης οι ομοιόμορφες περιοχές λαμβάνουν μηδενική τιμή. Στο Σχήμα 4.3 φαίνονται οι διαφορές για τις κλίμακες του Σχήματος 4.2.

4.1.4 Κανονικοποίηση

Το στάδιο της κανονικοποίησης είναι αυτό που θα συγκεντρώσει την ενέργεια του σήματος στα υποψήφια σημεία σημαντικότητας. Στόχος της κανονικοποίησης είναι να τονίσει σημεία στα οποία παρατηρούνται μεταβολές και τοπικά μέγιστα, και να καταπιέσει τις ομοιόμορφες περιοχές. Επιλέχθηκε η μη-γραμμική κανονικοποίηση των Kaya και Elhilali [30] προς επίτευξη αυτού του σκοπού. Ανάλογος τρόπος κανονικοποίησης έχει δοκιμαστεί και από τους Itti και Koch [26] σε εικόνες, με ικανοποιητικά ποσοστά αναγνώρισης. Σε κάθε μία από τις καμπύλες που προκύπτουν έπειτα από το στάδιο των διαφορών κέντρου-περίγυρου, πραγματοποιείται η εξής κανονικοποίηση:

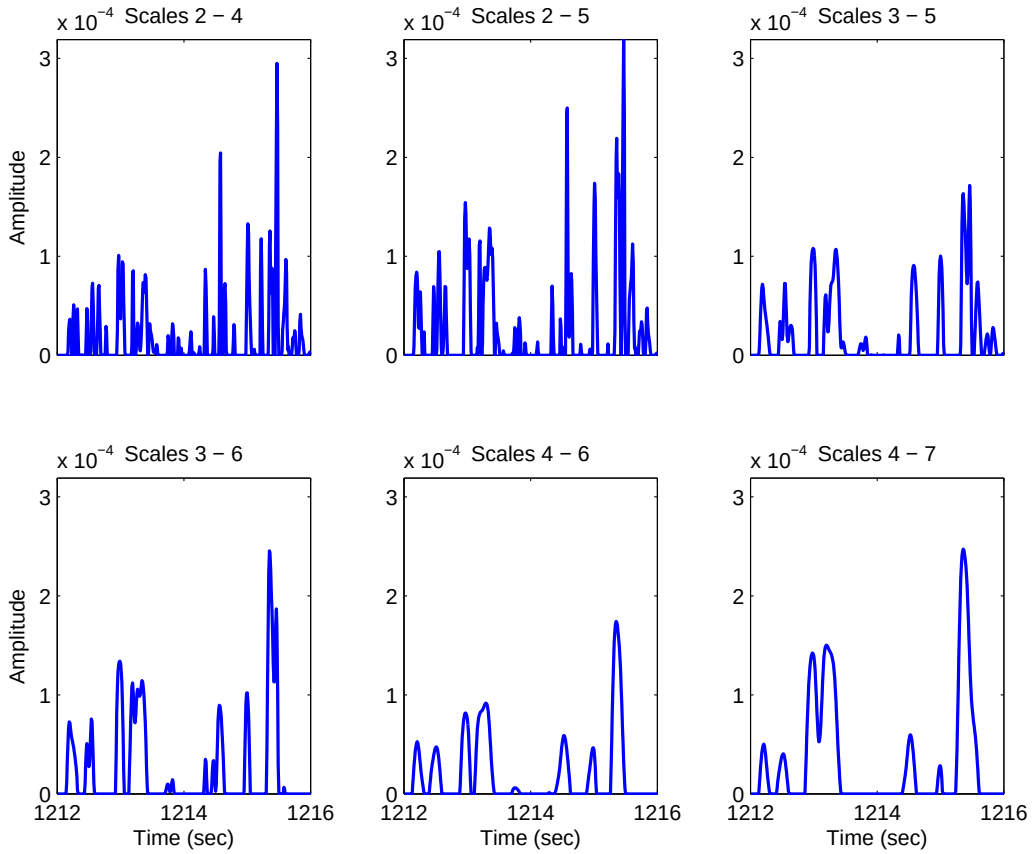
$$M \leftarrow |M + M * DoG|_{\geq 0} \tag{4.1}$$

όπου M η καμπύλη σε κάθε επανάληψη, DoG μονοδιάστατη διαφορά από Gaussians (*Difference of Gaussians*). Με τον συμβολισμό $|\cdot|_{\geq 0}$ δηλώνεται ότι οι αρνητικές τιμές της καμπύλης απεικονίζονται στο μηδέν.

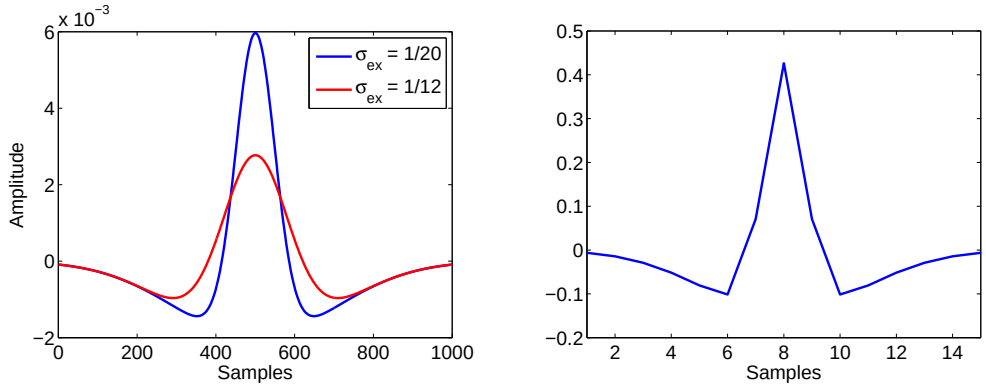
Το φίλτρο μονοδιάστατης διαφοράς από Gaussians έχει την μορφή των καμπυλών του Σχήματος 4.4α΄. Η γενική εξίσωση που περιγράφει τις καμπύλες είναι η ακόλουθη:

$$DoG(x) = \frac{1}{\sqrt{2\pi}\sigma_{ex}} e^{-x^2/(2\sigma_{ex}^2)} - \frac{1}{\sqrt{2\pi}\sigma_{inh}} e^{-x^2/(2\sigma_{inh}^2)} \tag{4.2}$$

όπου $\sigma_{ex},\ \sigma_{inh}$ οι τυπικές αποκλίσεις των Γκαουσσιανών, με $\sigma_{ex} < \sigma_{inh}$. Οι δύο Gaussians έχουν την ίδια μέση τιμή και συνεπώς το φίλτρο είναι συμμετρικό ως προς τον κατακόρυφο άξονα. Η περιοχή στην οποία λαμβάνει θετικές τιμές καλείται περιοχή διέγερσης (*excitation phase*), ενώ η περιοχή των αρνητικών τιμών καλείται περιοχή αποκοπής



Σχήμα 4.3: Διαφορές κλιμάκων του χαρακτηριστικού της ενέργειας αποσπάσματος της ταινίας Gladiator.



(α΄) (β΄)

Σχήμα 4.4: Αριστερά: μονοδιάστατα DoG φίλτρα με διαφορετικά σ_{ex}, και ίδια σ_{inh} $(= 0.2)$. Δεξιά: το φίλτρο που χρησιμοποιείται σε αυτή την εργασία με $\sigma_{ex} = .05,\ \sigma_{inh} = 0.2$.

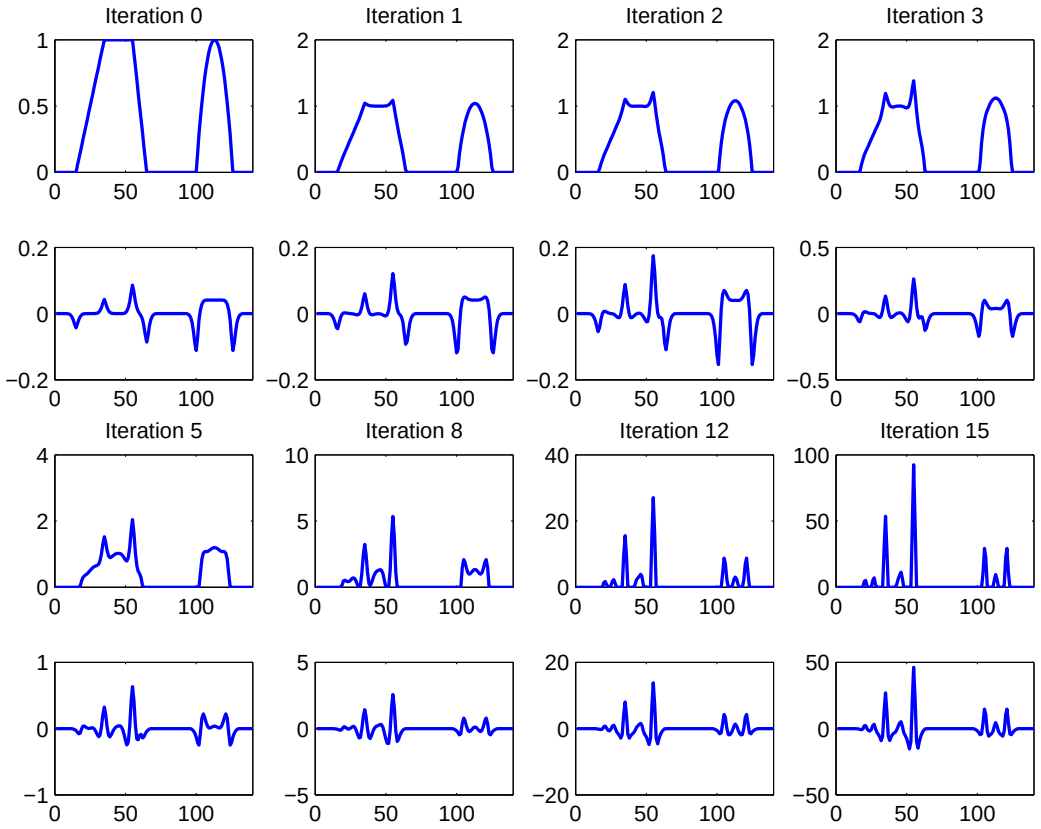
(*inhibition phase*). Σε αυτή την εργασία χρησιμοποιείται ένα φίλτρο 15 σημείων με τυπικές αποκλίσεις $\sigma_{ex} = 0.05,\ \sigma_{inh} = 0.2$ (Σχήμα 4.4β΄). Το μήκος του αντιστοιχεί σε διάρκεια 150 ms, με την περίοδο δειγματοληψίας των χαρακτηριστικών να είναι 10 ms.

Όταν το DoG φίλτρο εφαρμοσθεί σε ομοιόμορφη περιοχή ενός θετικού σήματος (περιοχή που το σήμα είναι σχεδόν σταθερό), θα δώσει μηδενική έξοδο διότι το εμβαδόν μεταξύ της καμπύλης του και του άξονα τετμημένων είναι μηδέν. Συνεπώς, η άθροιση του στο σήμα δεν θα το μεταβάλει.

Εάν το σήμα αρχίζει να αυξάνει, η έξοδος θα γίνει αρχικά αρνητική λόγω της αλληλεπίδρασης της μίας ζώνης αποκοπής με υψηλότερες τιμές. Μεγαλύτερη κλίση του σήματος έχει ως συνέπεια την λήψη μικρότερων τιμών στην έξοδο (μεγαλύτερων κατά μέτρο), με την ελάχιστη τιμή να επιτυγχάνεται όταν ολόκληρη η ζώνη αποκοπής και μόνο αυτή επικαλύπτεται με την αύξουσα περιοχή του σήματος. Η άθροιση του φιλτραρισμένου χάρτη στον αρχικό θα μειώσει τις τιμές του στην περιοχή, καθώς και στο τέλος της περιοχής που το σήμα ήταν σταθερό. Η επικάλυψη της ζώνης διέγερσης με την αύξουσα περιοχή θα αυξήσει την τιμή της εξόδου (πιθανώς να γίνει θετική) ενώ η δεύτερη ζώνη αποκοπής θα την μειώσει και πάλι φέρνοντας την σε τιμές κοντά στο μηδέν.

Η μετάβαση από αύξουσα περιοχή σε σταθερή ή φθίνουσα, οδηγεί σε αύξηση αρχικά της εξόδου κάνοντας την θετική, και έπειτα μείωση της μέχρι τον μηδενισμό. Ως συνέπεια, οι τιμές του χάρτη στο τέλος της περιοχής που ήταν αύξων θα αυξηθούν και οι κορυφές του θα ενισχυθούν. Η μετάβαση από φθίνουσα περιοχή σε σταθερή ή αύξουσα (περιοχή κοιλάδας) δίνει αρνητικές τιμές στην έξοδο και κατά συνέπεια μείωση του σήματος στο πέρας της φθίνουσας περιοχής. Επομένως τόσο οι αύξουσες όσο και οι φθίνουσες περιοχές του σήματος συρρικνώνονται (λόγω της απεικόνισης αρνητικών τιμών στο μηδέν) και το σήμα λαμβάνει μεγαλύτερες τιμές στο πέρας και αρχή τους, αντίστοιχα.

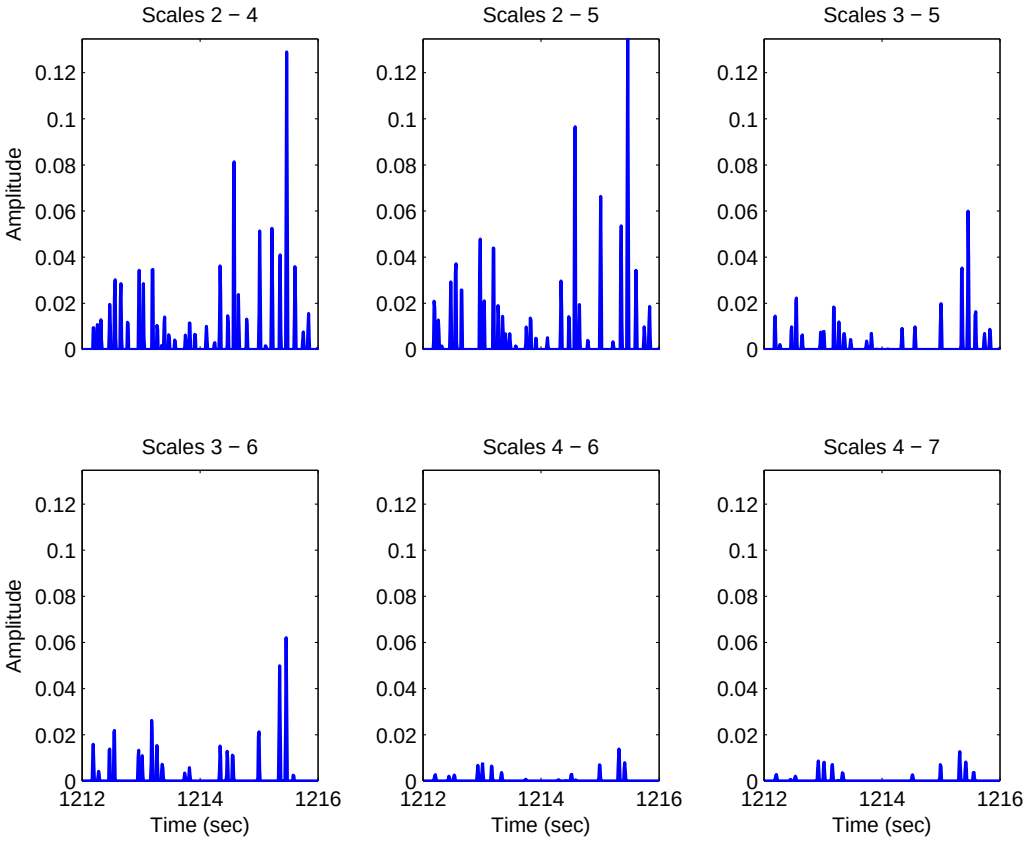
Στο Σχήμα 4.5 φαίνεται ένα παράδειγμα εφαρμογής της κανονικοποίησης σε μία συνάρτηση που αυξάνεται και μειώνεται κατά τμήματα γραμμικά και τετραγωνικά. Τα ση-



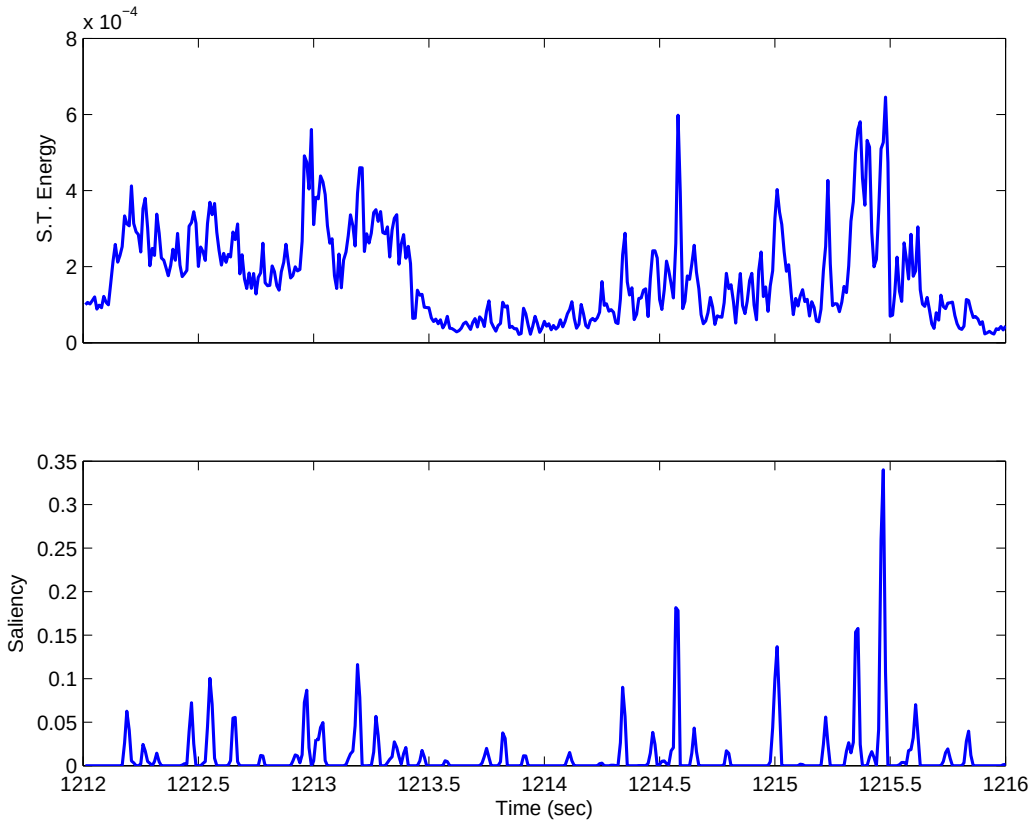
Σχήμα 4.5: Εφαρμογή επαναληπτικής κανονικοποίησης με χρήση DoG φίλτρου σε γραμμικές και τετραγωνικές μεταβολές του σήματος. Στις περιττές γραμμές φαίνεται η καμπύλη στην i επανάληψη, ενώ στις άρτιες το φιλτράρισμα της καμπύλης με DoG φίλτρο του Σχήματος 4.4β΄.

μεία που ενισχύονται στον χάρτη είναι κυρίως αυτά στα οποία υπάρχουν μεταβολές, συμπεριλαμβανομένων των τοπικών μεγίστων. Μεγαλύτερη κλίση της συνάρτησης οδηγεί σε μεγαλύτερες τιμές στην έξοδο. Η εφαρμογή της κανονικοποίησης στους χάρτες έχει ως αποτέλεσμα την δημιουργία μιας αραιής αναπαράστασης (*sparse representation*) με κορυφές τοποθετημένες στα σημεία που υπάρχουν μεταβολές και μέγιστα. Στο Σχήμα 4.6 φαίνεται η εφαρμογή της κανονικοποίησης στην ενέργεια του σήματος από απόσπασμα της ταινίας Gladiator. Έχει παρατηρηθεί ότι 12 έως 15 επαναλήψεις της διαδικασίας είναι αρκετές για να υπάρξει σύγκλιση. Στο εξής θα θεωρείται ότι εκτελούνται 15 επαναλήψεις σε αυτό το στάδιο, εκτός εάν αναφέρεται διαφορετικά.

Μετά την κανονικοποίηση των καμπυλών πραγματοποιείται ο συνδυασμός τους για κάθε χαρακτηριστικό. Έχει επιλεγεί να συνδυάζονται με άθροιση δίνοντας ίσο βάρος σε κάθε κανονικοποιημένη διαφορά από κλίμακες. Η έξοδος αυτού του σταδίου είναι η καμπύλη σημαντικότητας του κάθε χαρακτηριστικού. Θεωρώντας την επίδραση που έχει κάθε χαρακτηριστικό στην ανθρώπινη προσοχή μπορεί να χρησιμοποιηθεί για την ταξι-



Σχήμα 4.6: Εφαρμογή κανονικοποίησης με χρήση DoG φίλτρου στις διαφορές κέντρου-περίγυρου της ενέργειας αποσπάσματος της ταινίας Gladiator.



Σχήμα 4.7: Κυματομορφή του χαρακτηριστικού της ενέργειας και καμπύλη σημαντικότητας της, αποσπάσματος της ταινίας Gladiator.

νόμηση των σκηνών σε σημαντικές και μη, με βάση μόνο αυτό. Υψηλή τιμή στην καμπύλη κάποιου χαρακτηριστικού δείχνει την ύπαρξη μεταβολών ή μεγίστων, ικανών να κάνουν την σκηνή προεξέχουσα. Στο Σχήμα 4.7 φαίνεται η καμπύλη σημαντικότητας και η καμπύλη του χαρακτηριστικού της ενέργειας του αποσπάσματος από την ταινία Gladiator. Από την κυματομορφή του χαρακτηριστικού έχουν τονισθεί τα σημεία στα οποία υπάρχουν μεταβολές.

4.1.5 Συνδυασμός Χαρτών

Το τελικό στάδιο του μοντέλου είναι ο συνδυασμός των χαρτών (καμπυλών) για την δημιουργία του τελικού χάρτη σημαντικότητας. Ο συνδυασμός των χαρτών γίνεται γραμμικά:

$$S = \sum_i w_i S_i \tag{4.3}$$

με S τον τελικό χάρτη σημαντικότητας, S_i η καμπύλη σημαντικότητας για το i χαρακτηριστικό σταθμισμένη με βάρος w_i. Έγινε χρήση δύο διαφορετικών συνόλων τιμών για τα βάρη. Στο πρώτο σύνολο, όλα τα βάρη θεωρούνται ίσα (και ίσα με την μονάδα), ενώ στο δεύτερο το βάρος κάθε χαρακτηριστικού είναι ανάλογο της συσχέτισης του με ανθρώπινες θεωρήσεις της σημαντικότητας. Ο υπολογισμός των βαρών για την δεύτερη περίπτωση, περιγράφεται αναλυτικά στην ενότητα 5.2. Τα δύο σύνολα από βάρη είχαν παρόμοια επίδοση και οι καμπύλες που παράγονται υψηλή συσχέτιση μεταξύ τους.

4.2 Χαρακτηριστικά

Στις επόμενες υποενότητες γίνεται περιγραφή των χαρακτηριστικών που χρησιμοποιούνται στο μοντέλο που περιγράφηκε στην ενότητα 4.1.

4.2.1 Ενέργεια Βραχέως Χρόνου

Η ενέργεια ενός διακριτού σήματος x ορίζεται συνήθως ως εξής:

$$E = \sum_{m=-\infty}^{+\infty} x^2(m) \tag{4.4}$$

Τα σήματα έχουν συνήθως αρκετές τοπικές διακυμάνσεις στην ενέργεια που δεν είναι δυνατή η απεικόνιση τους εάν αυτή υπολογιστεί σε όλο το χρονικό εύρος του σήματος. Είναι χρησιμότερο να υπολογισθεί η ενέργεια τοπικά σε μικρά χρονικά παράθυρα του σήματος. Η ενέργεια βραχέως χρόνου του σήματος την στιγμή n ορίζεται ως:

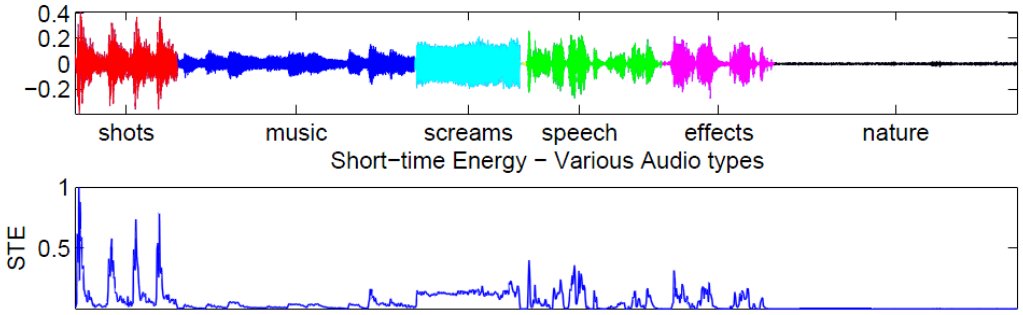
$$E_n = \sum_{m=-\infty}^{+\infty} (x(m)w(n-m))^2 \tag{4.5}$$

όπου w, είναι παράθυρο διάρκειας μερικών ms.

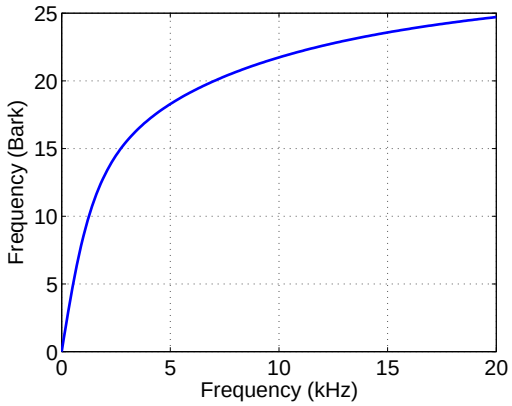
Η ενέργεια βραχέως χρόνου είναι ένα μέγεθος εύκολο να υπολογισθεί και μπορεί να χρησιμοποιηθεί για την ταξινόμηση των ήχων σε κατηγορίες. Στο Σχήμα 4.8 φαίνεται η ενέργεια βραχέως χρόνου για ήχους από διαφορετικές κατηγορίες. Μεταξύ κάποιων κατηγοριών είναι εμφανής η διαφορά. Όπως θα φανεί σε επόμενα κεφάλαια το χαρακτηριστικό της ενέργειας θα είναι αρκετά αποδοτικό στη διάκριση των κλάσεων σημαντικότητας.

4.2.2 Loudness

Το loudness ενός σήματος σχετίζεται με την αίσθηση της έντασης του σήματος. Έχει παρατηρηθεί ότι ήχοι με την ίδια ένταση μετρούμενη σε decibel δεν γίνονται αντιληπτοί ως να έχουν την ίδια. Το ανθρώπινο ακουστικό σύστημα δείχνει συχνοτική εξάρτηση στην αντιλαμβανόμενη ένταση των ήχων. Ως συνέπεια, απαιτείται ο μετασχηματισμός της έντασης του σήματος σε ένα μέγεθος του οποίου η τιμή να είναι αύξουσα με την αντιλαμβανόμενη ένταση. Το loudness του σήματος έχει αυτή την ιδιότητα και θα αποτελέσει ένα από τα χαρακτηριστικά που θα χρησιμοποιηθούν.



Σχήμα 4.8: Κυματομορφή του ηχητικού σήματος που αποτελείται από διάφορες κατηγορίες ήχων (πρώτη σειρά), και η ενέργεια βραχέως χρόνου του σήματος (δεύτερη σειρά).



Σχήμα 4.9: Κλίμακα Bark ως συνάρτηση της γραμμικής κλίμακας συχνοτήτων.

Διάφορα μοντέλα έχουν προταθεί για τον υπολογισμό του loudness [46, 68, 45]. Σε αυτή την εργασία υιοθετείται το μοντέλο του Yang [68] το οποίο είναι ένα μοντέλο προσέγγισης του loudness από πειραματικές μετρήσεις. Το πρώτο στάδιο του μοντέλου είναι ο χωρισμός του σήματος σε χρονικά πλαίσια μήκους 30 ms με 20 ms επικάλυψη μεταξύ διαδοχικών πλαισίων. Κάθε πλαίσιο σταθμίζεται με παράθυρο Hanning το οποίο περιγράφεται από την ακόλουθη εξίσωση:

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), \quad n = 0, 1, \ldots, N-1 \tag{4.6}$$

Ακολουθεί ο υπολογισμός του φάσματος ισχύος σε κάθε παράθυρο, το οποίο ισούται με το πλάτος του μετασχηματισμού Fourier υψωμένο στο τετράγωνο. Το επόμενο στάδιο είναι ο χωρισμός του φάσματος σε μπάντες συχνοτήτων και ο υπολογισμός της ενέργειας σε κάθε μπάντα. Ο χωρισμός σε μπάντες γίνεται με βάση την κλίμακα Bark.

Η κλίμακα Bark βασίζεται στη συχνοτική εξάρτηση της αντιλαμβανόμενης έντασης.

Πίνακας 4.1: Συχνοτικά όρια (σε Hz) κάθε μπάντας της κλίμακας Bark.

z(Bark)	Center(Hz)	Upper(Hz)	z(Bark)	Center(Hz)	Upper(Hz)
1	50	100	13	1850	2000
2	150	200	14	2150	2320
3	250	300	15	2500	2700
4	350	400	16	2900	3150
5	450	510	17	3400	3700
6	570	630	18	4000	4400
7	700	770	19	4800	5300
8	840	920	20	5800	6400
9	1000	1080	21	7000	7700
10	1170	1270	22	8500	9500
11	1370	1480	23	10500	12000
12	1600	1720	24	13500	15500

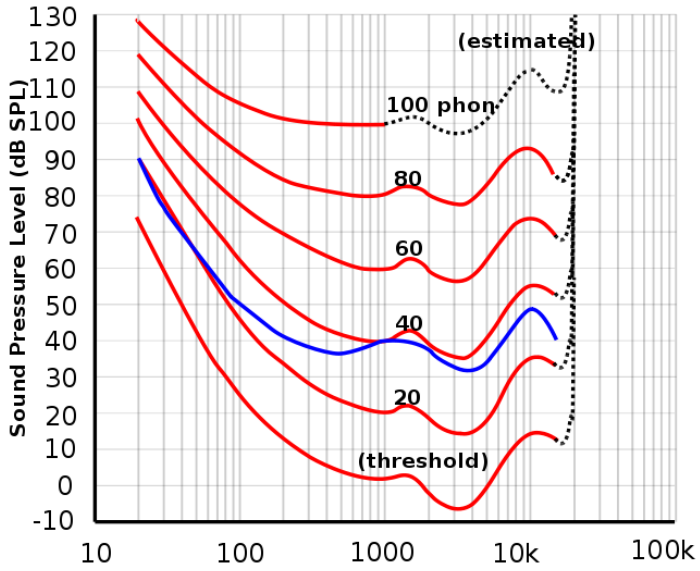
Γίνεται εκτίμηση του κρίσιμου εύρους συχνοτήτων (*critical bandwidth*) για το οποίο δεν μεταβάλλεται η αντιλαμβανόμενη ένταση με την μεταβολή της συχνότητας μέσα στα όρια του. Περαιτέρω αύξηση του εύρους συχνοτήτων οδηγεί σε αύξηση της αίσθησης του loudness παρόλο που η ένταση του σήματος διατηρείται σταθερή. Στις χαμηλές συχνότητες (έως 500 Hz) το κρίσιμο εύρος παραμένει σταθερό και περίπου ίσο με 100 Hz, ενώ στην συνέχεια αυξάνεται εκθετικά. Στο Σχήμα 4.9 φαίνεται η καμπύλη που συνδέει την γραμμική κλίμακα με την κλίμακα Bark η οποία προκύπτει από την ακόλουθη εξίσωση:

$$Z_{bark}(f) = 13 \arctan(0.00076 f) + 3.5 \arctan\left((f/7500)^2\right) \tag{4.7}$$

όπου f η συχνότητα σε Hz. Στον Πίνακα 4.1 φαίνονται τα συχνοτικά όρια για τον χωρισμό σε Bark μπάντες. Σε κάθε μπάντα πραγματοποιείται άθροιση των πλατών του φάσματος ισχύος που ανήκουν σε αυτή, και έπειτα λαμβάνεται ο λογάριθμος του, ώστε η ενέργεια να μετράται σε dB. Πιο συνοπτικά, ο μετασχηματισμός είναι ο εξής:

$$spl(z) = 10 \log_{10} \left(\sum_{k \in z \text{ Band}} |X(k)|^2 \right) \tag{4.8}$$

όπου z ο αριθμός της μπάντας, και X ο μετασχηματισμός Fourier του σήματος. Λόγω αλληλεπίδρασης μεταξύ των μπάντων, παρατηρούνται φαινόμενα συγκάλυψης (masking). Για να ληφθεί υπόψιν αυτή η αλληλεπίδραση γίνεται στάθμιση της ενέργειας κάθε μπάντας με την ακόλουθη συνάρτηση, γνωστή ως συνάρτηση διάχυσης (*spreading function*) [55]:



Σχήμα 4.10: Ισοϋψείς καμπύλες loudness όπως υπολογίστηκαν από το διεθνές στάνταρ ISO [23]. Με μπλε χρώμα φαίνεται προγενέστερη εκτίμηση της καμπύλης των 40 phons. Πηγή Wikipedia.

$$S(i,j) = 15.81 + 7.5(i - j + 0.474) - 17.5\sqrt{1 + (i - j + 0.474)^2} \tag{4.9}$$

$$SPL' = S \cdot SPL \tag{4.10}$$

με $i, j = 1, 2, \ldots$, |i-j| < 25, i η Bark μπάντα του σήματος που συγκαλύπτεται (*masked band*), j η μπάντα του σήματος που συγκαλύπτει (*masking band*), και SPL διάνυσμα με στοιχεία τα spl.

Από μετρήσεις που έχουν γίνει σε ανθρώπους για την αντιλαμβανόμενη ένταση των ήχων, έχουν υπολογισθεί καμπύλες οι οποίες ως συνάρτηση της συχνότητας δίνουν σταθερή αντιλαμβανόμενη ένταση. Τέτοιες καμπύλες φαίνονται στο Σχήμα 4.10. Για την μέτρηση της ίδιας αντιλαμβανόμενης έντασης χρησιμοποιείται η κλίμακα των phons. Ίδιος αριθμός από phons αντιστοιχεί στην ίδια αντιλαμβανόμενη ένταση. Για τον υπολογισμό του loudness σε phons αρχικά επιλέγονται κάποιες στάθμες (καμπύλες) αναφοράς. Έπειτα με χρήση της κεντρικής συχνότητας κάθε Bark μπάντας πραγματοποιείται γραμμική παρεμβολή στις δύο καμπύλες μεταξύ των οποίων βρίσκεται η υπολογισμένη σε dB ένταση (ενέργεια), και έτσι υπολογίζεται η ένταση σε phon (έστω l).

Ένα μειονέκτημα της κλίμακας των phons είναι ότι η αντιστοίχιση με την αντιλαμβανόμενη ένταση δεν είναι γραμμική και διπλασιασμός του αριθμού των phons δεν συνεπάγεται διπλασιασμό της. Για την επίτευξη γραμμικότητας του loudness εισήχθη η sone κλίμακα που συνδέεται με την phon μέσω της ακόλουθης σχέσης [3]:

$$S(z) = \begin{cases} \left(\frac{l(z)}{40}\right)^{2.642} & , \quad \text{εάν } l(z) < 40\, phons, \\ 2^{(l(z)-40)/10} & , \quad \text{αλλιώς} \end{cases} \tag{4.11}$$

όπου l η ένταση σε phons, και S η ένταση σε sones. Η στάθμη αναφοράς είναι 1 sone, που είναι η αντιλαμβανόμενη ένταση που προκαλεί ένας τόνος συχνότητας 1 kHz με ένταση 40 dB, και συνήθως χρησιμοποιείται ως στάθμη αναφοράς. Για διπλασιασμό του loudness (2 sones) απαιτείται η ένταση του τόνου να γίνει 50 dB (αύξηση 10 dB).

Από την σχέση 4.11 υπολογίζεται το loudness που προκαλείται από κάθε μπάντα συχνοτήτων (*specific loudness*). Για τον υπολογισμό του ολικού loudness (*total loudness*), που προκαλείται από το σήμα απαιτείται η άθροιση του loudness στις μπάντες:

$$TL = \sum_{z} S(z) \tag{4.12}$$

4.2.3 Roughness

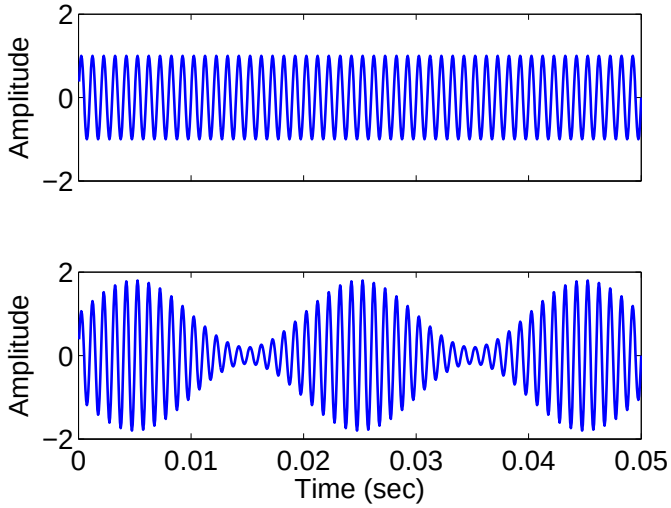
Με τον όρο roughness γίνεται αναφορά στην τραχύτητα ορισμένων ήχων. Εμφανίζεται όταν παρατηρούνται ταλαντώσεις της περιβάλλουσας του σήματος με ρυθμό μικρότερο από ένα κρίσιμο εύρος συχνοτήτων [72], οι οποίες καθορίζονται από την ύπαρξη και ένταση κοντινών συχνοτικών συνιστωσών στο φάσμα του σήματος. Η αίσθηση της τραχύτητας του ήχου συνδέεται με την αδυναμία του ακουστικού συστήματος να διαχωρίσει συχνότητες που απέχουν λιγότερο από το κρίσιμο εύρος [8].

Οι διακυμάνσεις της περιβάλλουσας μπορούν να χωρισθούν σε τρεις επικαλυπτόμενες κατηγορίες με βάση το αντιληπτικό τους αποτέλεσμα [50, 64]: αργές διακυμάνσεις (≤ 15 Hz), οι οποίες προκαλούν διακυμάνσεις στο loudness που είναι γνωστές και ως "beating" φαινόμενο. Έπειτα με την αύξηση του ρυθμού ταλαντώσεων το loudness σταθεροποιείται και αυξάνεται η τραχύτητα του ήχου. Επιπλέον αύξηση του ρυθμού ταλαντώσεων($75 - 150$ Hz) οδηγεί σε μεγιστοποίηση του roughness, και έπειτα σταδιακή μείωση του μέχρι που σχεδόν εξαφανίζεται.

Στην περίπτωση δύο ημιτόνων με συχνότητες $f_1,\ f_2$, η εμφάνιση του roughness εξαρτάται από την διαφορά $|f_1 - f_2|$. Εάν αυτή η διαφορά είναι μικρότερη από ένα κρίσιμο εύρος συχνοτήτων τότε παρατηρείται εμφάνιση του roughness ή διακυμάνσεων στο loudness. Εάν η διαφορά υπερβεί ένα κατώφλι, οι ήχοι ακούγονται ως δύο διαφορετικά ρεύματα με σχεδόν μηδενικό roughness. Πολλοί σύνθετοι ήχοι έχουν κοντινές συχνότητες, ωστόσο δεν αφήνουν όλοι την ίδια αίσθηση τραχύτητας.

Ένας, επίσης, σημαντικός παράγοντας όταν εμφανίζεται το roughness είναι το πλάτος της ταλάντωσης της περιβάλλουσας (διαφορά μεταξύ κορυφών και κοιλάδων). Αναφέρεται και ως βαθμός του roughness, και εξαρτάται από το σχετικό πλάτος των συνιστωσών στο φάσμα του σήματος, με ίσες συνιστώσες να προκαλούν μέγιστο ρυθμό ταλάντωσης και μέγιστο βαθμό τραχύτητας.

Το roughness ενός σήματος που το φάσμα του αποτελείται από δύο τόνους με συχνότητες f_1, f_2 και πλάτη A_1, A_2, με $A_{min} = \min\{A_1, A_2\},\ f_{min} = \min\{f_1, f_2\},\ f_{max} = \max\{f_1, f_2\}$, δίνεται από την σχέση [57]:



Σχήμα 4.11: Παράδειγμα σήματος όπου η εμφάνιση ταλαντώσεων στην περιβάλλουσα οδηγεί σε αύξηση του roughness.

$$R = X^{0.1} \cdot 0.5Y^{3.11} \cdot Z \tag{4.13}$$

όπου

$$X = A_1 \cdot A_2 \tag{4.14}$$

$$Y = 2A_{min}/(A_1 + A_2) \tag{4.15}$$

και

$$Z = e^{-b_1 s(f_{max}-f_{min})} - e^{-b_2 s(f_{max}-f_{min})} \tag{4.16}$$

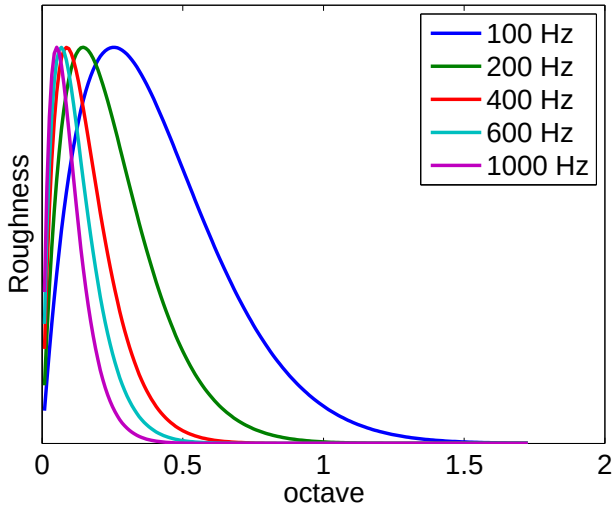
με

$$b_1 = 3.5,\ b_2 = 5.75 \quad \text{και} \quad s = 0.24/(s_1 f_{min} + s_2),\ s_1 = 0.0207,\ s_2 = 18.96$$

Ο όρος X αντιπροσωπεύει την εξάρτηση του roughness από την ένταση των ημιτόνων [64, 57, 60], ο όρος Y από το σχετικό πλάτος των κορυφών, ενώ ο όρος Z την εξάρτηση από τον ρυθμό μεταβολής του πλάτους της κυματομορφής (διαφορά συχνότητας των ημιτόνων), και την μικρότερη των δύο συχνοτήτων. Στο Σχήμα 4.12 φαίνεται πως μεταβάλλεται το roughness ως συνάρτηση της διαφοράς συχνοτήτων των δύο τόνων και της ελάχιστης συχνότητας. Για τον υπολογισμό του roughness ενός σήματος που το φάσμα του έχει περισσότερες από δύο ημιτονικές συνιστώσες (κορυφές), αθροίζεται το roughness μεταξύ όλων των ζευγών κορυφών στο φάσμα.

4.2.4 Fractal Διάσταση Σήματος

Τα fractals έχουν χρησιμοποιηθεί με επιτυχία στην μοντελοποίηση πολλών φαινομένων με υψηλή γεωμετρική πολυπλοκότητα [37, 2], όπως σε εικόνες με φυσικές σκηνές



Σχήμα 4.12: Καμπύλη roughness σήματος με δύο ημιτονικές συνιστώσες ίδιου πλάτους ως συνάρτηση της διαφοράς συχνοτήτων και της ελάχιστης συχνότητας.

(π.χ. βουνά και δάση), την κίνηση σε τηλεπικοινωνιακά δίκτυα, αλλά και σε βιολογικές διεργασίες. Έχουν φανεί επίσης χρήσιμα στην επεξεργασία και μοντελοποίηση χαρακτηριστικών των ακουστικών σημάτων, όπως μεταβολές του pitch σε σήματα μουσικής, στην σύνθεση και αναγνώριση ήχων, καθώς και στην συμπίεση τους. Σε αυτή την εργασία θα γίνει χρήση της fractal διάστασης ενός σήματος.

Η fractal διάσταση ενός σήματος είναι ένα μέτρο του βαθμού τμηματοποίησης του. Για τον υπολογισμό της θα γίνει χρήση της *Minkowski-Bouligand διάστασης*. Στόχος της διάστασης Minkowski-Bouligand είναι η μέτρηση του μήκους μιας πιθανώς μη ομαλής καμπύλης. Για τον σκοπό αυτό ο Minkowski δοκίμασε να καλύψει την καμπύλη με δίσκους ακτίνας ε, να υπολογίσει το εμβαδόν $A_B(\varepsilon)$ του χωρίου που δημιουργείται και τέλος να λάβει το όριο $\varepsilon \to 0$ του εμβαδού προς την διάμετρο του δίσκου. Πιο συγκεκριμένα, έστω S μία συνάρτηση ορισμένη στο $[0, T]$. Το γράφημα της S, ορίζεται ως εξής:

$$F = \{(t, S(t)) \in \mathbb{R}^2 : 0 \leq t \leq T\} \tag{4.17}$$

Εάν B είναι ο μοναδιαίος κύκλος στο επίπεδο με κέντρο την αρχή των αξόνων, και εB η κλιμακωμένη κατά ε έκδοση του, υπολογίζεται το εμβαδόν του χωρίου που δημιουργείται κεντράροντας τον δίσκο εB σε κάθε σημείο της καμπύλης και λαμβάνοντας την ένωση των σημείων:

$$A_B(\varepsilon) = area(F \oplus \varepsilon B) \tag{4.18}$$

όπου

$$F \oplus \varepsilon B = \{z + \varepsilon b \in \mathbb{R}^2 : z \in F,\ b \in B\} \tag{4.19}$$

Όπως έχει δειχθεί από τους Maragos & Sun [42], το B αρκεί να είναι οποιοδήποτε συμπαγές, απλά συνεκτικό και συμμετρικό υποσύνολο του $\mathbb{R}^2$. Η fractal διάσταση του σήματος

προκύπτει ότι είναι ίση με:

$$D = 2 - \lim_{\varepsilon \to 0} \frac{\log(A_B(\varepsilon))}{\log \varepsilon}. \tag{4.20}$$

Υποθέτοντας ότι $\log(A_B(\varepsilon)) \approx (2 - D) \log \varepsilon + const$ καθώς $\varepsilon \to 0$, η fractal διάσταση μπορεί να υπολογισθεί με χρήση ελαχίστων τετραγώνων υπολογίζοντας το $A_B(\varepsilon)$ για διάφορα ε.

Με χρήση της ανωτέρω διαδικασίας είναι δυνατός ο υπολογισμός του εμβαδού $A_B(\varepsilon)$, και της διάστασης Minkowski. Ωστόσο, έχει δειχθεί [42, 38] ότι ο υπολογισμός του εμβαδού $A_B(\varepsilon)$ μπορεί να γίνει χωρίς την χρήση διδιάστατης γεωμετρίας, μόνο με σήματα σε μία διάσταση, ως εξής:

$$A_B(\varepsilon) = \int_0^T S \oplus G_\varepsilon(t) - S \ominus G_\varepsilon(t)\, dt + O(\varepsilon^2), \tag{4.21}$$

όπου

$$S \oplus G_\varepsilon(t) = \sup_x \{S(t) + G_\varepsilon(t - x)\} \tag{4.22}$$

$$S \ominus G_\varepsilon(t) = \inf_x \{S(t) - G_\varepsilon(x - t)\} \tag{4.23}$$

πράξεις dilation και erosion αντίστοιχα, και

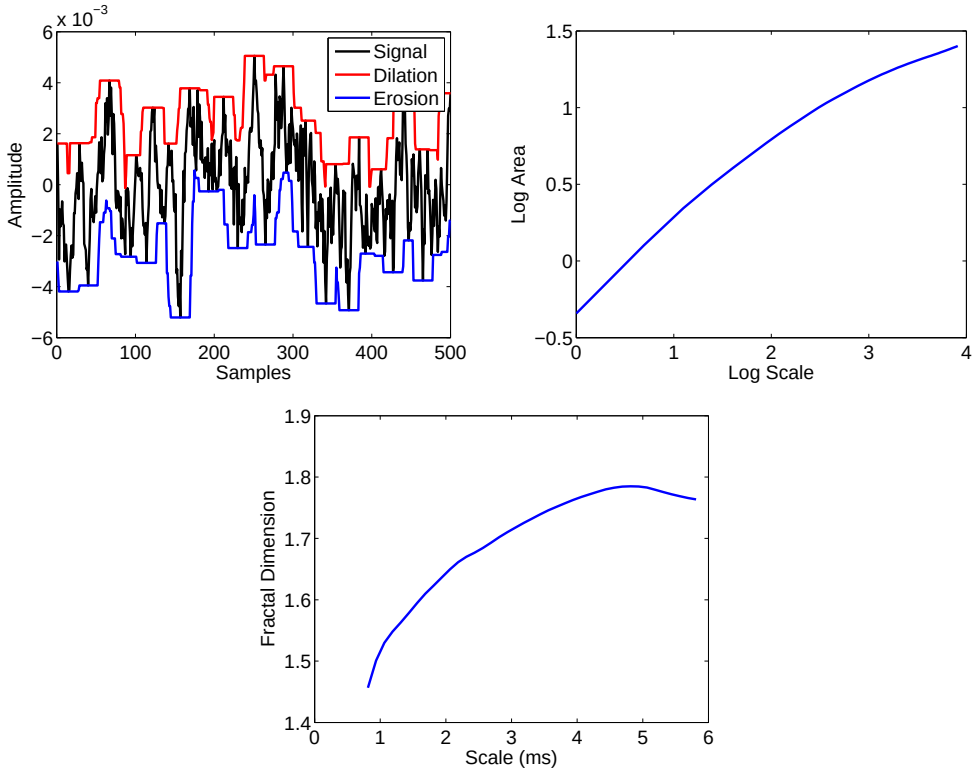
$$G_\varepsilon(t) = \sup\{y \in \mathbb{R} : (t, y) \in \varepsilon B\} \tag{4.24}$$

Ο υπολογισμός του εμβαδού μέσω της σχέσης 4.21 μειώνει την υπολογιστική πολυπλοκότητα καθώς δεν απαιτείται ο χειρισμός δισδιάστατων σχημάτων και όλες οι πράξεις γίνονται σε μία διάσταση. Επιπλέον, προσφέρει έναν πρακτικό και αποδοτικό αλγόριθμο εκτίμησης του στις διάφορες κλίμακες στην περίπτωση ψηφιακών σημάτων.

Στην περίπτωση διακριτών σημάτων, αρκεί στις ανωτέρω εξισώσεις να αντικατασταθούν τα ολοκληρώματα με αθροίσματα στο χρονικό εύρος των σημάτων. Η κλίμακα λαμβάνει διακριτές τιμές $\varepsilon = 1, 2, \ldots, \varepsilon_{max}$. Εάν το σύνολο B επιλεγεί να είναι κυρτό με ακτίνα 1, τότε η συνάρτηση G μπορεί να έχει είτε τριγωνικό σχήμα, με $G[-1] = G[1] = 0,\ G[0] = h$, είτε ορθογώνιο, με $G[-1] = G[1] = G[0] = h$, και μηδέν αλλού. Για τον υπολογισμό του εμβαδού $A_B(\varepsilon)$ στις διάφορες κλίμακες χρησιμοποιείται ο εξής επαναληπτικός αλγόριθμος [38]:

$$\left\{ \begin{array}{ll} S \oplus G[n] = \max_{-1 \le k \le 1}\{S[n + k] + G[k]\} & \\ S \ominus G[n] = \min_{-1 \le k \le 1}\{S[n + k] - G[k]\} & , \quad \varepsilon = 1 \\ & \\ S \oplus G_{\varepsilon+1}[n] = (S \oplus G_\varepsilon[n]) \oplus G & \\ S \ominus G_{\varepsilon+1}[n] = (S \ominus G_\varepsilon[n]) \ominus G & , \quad \varepsilon > 1 \end{array} \right\} \tag{4.25}$$

Τέλος, βρίσκεται η κλίση της καμπύλης $(\log \varepsilon,\ \log(A_B(\varepsilon)))$ με χρήση ελαχίστων τετραγώνων, και υπολογίζεται η fractal διάσταση. Επειδή για φυσικά σήματα (μη-συνθετικά)



Σχήμα 4.13: Από αριστερά προς τα δεξιά: dilation και erosion αποσπάσματος με φωνή, εμβαδόν μεταξύ των δύο καμπυλών ως συνάρτηση της κλίμακας, και fractal διάσταση του σήματος ως συνάρτηση της κλίμακας.

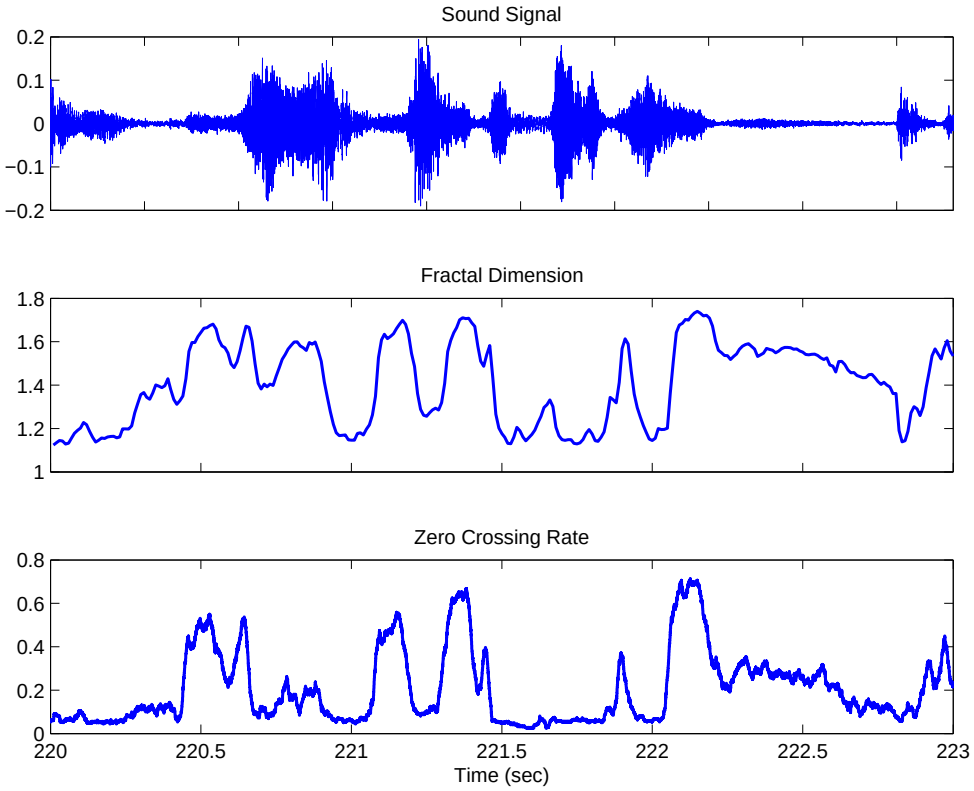
η fractal διάσταση δεν είναι σταθερή με μεταβολή της κλίμακας ε, γίνεται εκτίμησή της σε κινούμενα παράθυρα μήκους w από κλίμακες. Σε αυτήν την εργασία έχει επιλεγεί $w = 10$. Επίσης, επιλέγεται $h = 0$ και η συνάρτηση G γίνεται επίπεδη.

Στο Σχήμα 4.13 φαίνεται η εφαρμογή των dilation και erosion σε μικρό απόσπασμα ομιλίας. Στο ίδιο Σχήμα φαίνονται επίσης ο λογάριθμος του εμβαδού μεταξύ των δύο καμπυλών ως συνάρτηση της κλίμακας, καθώς και η fractal διάσταση.

Στο Σχήμα 4.14 φαίνεται η fractal διάσταση για την κλίμακα $\varepsilon = 1$, και το zero-crossing rate ενός αποσπάσματος ομιλίας από την ταινία Gladiator. Παρατηρείται υψηλή συσχέτιση μεταξύ των δύο μεγεθών όπως έχει τονιστεί και στο [41]. Η υψηλή συσχέτιση παρατηρήθηκε όχι μόνο σε σήματα φωνής, αλλά σε όλες τις κατηγορίες ήχων που εμφανίζονται στην βάση δεδομένων που χρησιμοποιείται.

4.2.5 Συσχέτιση μεταξύ των χαρακτηριστικών

Υπολογίστηκε η γραμμική συσχέτιση μεταξύ των χαρακτηριστικών ώστε να ελεγχθεί εάν κάποιο μπορεί να προβλεφθεί από τα υπόλοιπα. Διαπιστώθηκε ότι η ενέργεια είναι συσχετισμένη με το loudness (όπως αναμενόταν), και λιγότερο με το roughness. Υψηλή



Σχήμα 4.14: Fractal διάσταση και zero-crossing rate αποσπάσματος ομιλίας από την ταινία Gladiator, στην δεύτερη και τρίτη σειρά αντίστοιχα.

Πίνακας 4.2: Συσχέτιση μεταξύ των χαρακτηριστικών που εξήχθησαν.

Feature	Energy	Loudness	Roughness	Fractal Dim.
Energy	1	0.66	0.38	0.05
Loudness	0.66	1	0.64	0.18
Roughenss	0.38	0.64	1	0.41
Fractal Dim.	0.05	0.18	0.41	1

Πίνακας 4.3: Συσχέτιση μεταξύ των χαρακτηριστικών στην έξοδο του μοντέλου.

Feature	Energy	Loudness	Roughness	Fractal Dim.
Energy	1	0.37	0.04	-0.029
Loudness	0.37	1	0.19	0.099
Roughenss	0.04	0.19	1	0.18
Fractal Dim.	-0.029	0.099	0.18	1

συσχέτιση υπάρχει επίσης, μεταξύ loudness και roughness, η οποία οφείλεται κυρίως στην εξάρτηση του roughness από το πλάτος των αρμονικών συνιστωσών. Μικρή συσχέτιση υπήρχε μεταξύ ενέργειας, loudness και fractal διάστασης. Τέλος συσχέτιση υπήρχε μεταξύ roughness και fractal διάστασης. Στον Πίνακα 4.2 φαίνονται τα αποτελέσματα για κάθε ζευγάρι χαρακτηριστικών.

Έγινε υπολογισμός, επίσης, της συσχέτισης μεταξύ των χαρακτηριστικών στην έξοδο του μοντέλου (Πίνακας 4.3). Η ενέργεια και το loudness έχουν και πάλι την υψηλότερη συσχέτιση μεταξύ των ζευγών, ωστόσο μικρότερη σε σύγκριση με τα χαρακτηριστικά πριν δοθούν ως είσοδος στο μοντέλο. Η διάταξη συσχέτισης μεταξύ όλων των ζευγών χαρακτηριστικών δεν άλλαξε από την είσοδο στην έξοδο του μοντέλου, και παρατηρήθηκε μείωση της κατά απόλυτη τιμή.

Τέλος υπολογίσθηκε η συσχέτιση μεταξύ κάθε χαρακτηριστικού στην είσοδο και στην έξοδο του μοντέλου, ώστε να διαπιστωθεί η επίδραση του μοντέλου στα χαρακτηριστικά. Στον Πίνακα 4.4 φαίνονται τα αποτελέσματα.

Πίνακας 4.4: Συσχέτιση εισόδου-εξόδου του μοντέλου για τα χαρακτηριστικά που εξήχθησαν. Με $'$ η έξοδος του μοντέλου.

Feature	Energy	Loudness	Roughness	Fractal Dim.
Energy$'$	0.70	0.41	0.23	0.02
Loudness$'$	0.24	0.44	0.29	0.11
Roughenss$'$	0.02	0.07	0.52	0.15
Fractal Dim.$'$	-0.06	-0.06	0.10	0.50

Κεφάλαιο 5

Εφαρμογή Μοντέλου Χρονικού Χάρτη Σημαντικότητας

Εισαγωγή

Σε αυτό το κεφάλαιο εφαρμόζεται το μοντέλο του χρονικού χάρτη σημαντικότητας στο πρόβλημα της ταξινόμησης ηχητικών σκηνών σε μία βάση δεδομένων, σε σημαντικές και μη. Αρχικά παρουσιάζεται η έξοδος του μοντέλου σε διάφορες κατηγορίες ήχων. Στη συνέχεια, κατωφλιώνοντας την καμπύλη σημαντικότητας λαμβάνεται απόφαση για τη σημαντικότητα των ηχητικών σκηνών, και εξετάζεται η συνεισφορά του κάθε χαρακτηριστικού στην ταξινόμηση. Επίσης, χρησιμοποιώντας τα χαρακτηριστικά που υπολογίσθηκαν στο προηγούμενο κεφάλαιο δημιουργούνται διανύσματα και ταξινομούνται με χρήση των SVM ταξινομητών. Με χρήση των διανυσμάτων επιτυγχάνεται υψηλότερη ακρίβεια ταξινόμησης. Τέλος, από τα διανύσματα χαρακτηριστικών υπολογίζονται ιστογράμματα και πραγματοποιείται ταξινόμηση με χρήση αυτών. Με τα ιστογραφικά διανύσματα παρατηρείται μια μικρή πτώση στην ακρίβεια ταξινόμησης αλλά σημαντική αύξηση στο ποσοστό σημείων που ανιχνεύονται (recall).

5.1 Η έξοδος του Μοντέλου

Σε αυτή την ενότητα παρουσιάζεται η έξοδος του μοντέλου του χρονικού χάρτη σημαντικότητας καθώς και των χαρακτηριστικών που περιγράφηκαν στο προηγούμενο κεφάλαιο για διάφορες κατηγορίες ήχων. Γίνεται χρήση απλών ήχων, αλλά και πιο πολύπλοκων που εμφανίζονται στην φύση.

5.1.1 Καμπύλες σημαντικότητας απλών ήχων

Αρχικά ελέγχεται το μοντέλο σε συνθετικούς ήχους με απλή δομή, ώστε να διαπιστωθεί η δυνατότητα του να τονίζει σημεία σημαντικότητας των χαρακτηριστικών που

χρησιμοποιούνται. Οι ήχοι είναι επιλεγμένοι ώστε να υπάρχει μεταβολή σε κάποιο από τα χαρακτηριστικά.

Γραμμική αύξηση της συχνότητας

Έχει παρατηρηθεί ότι τόνοι των οποίων η συχνότητα αυξάνεται ως συνάρτηση του χρόνου έχουν την τάση να έλκουν την ανθρώπινη προσοχή πιο "εύκολα" από τόνους με σταθερή συχνότητα. Αυτό το φαινόμενο συχνά αντιστοιχίζεται με φαινόμενο της οπτικής προσοχής του ανθρώπου όπου οι ακμές σε μία εικόνα με κλίση διαφορετική των 0 ή 90 μοιρών είναι πιο προεξέχουσες και άμεσα αντιληπτές [35].

Εξετάζεται εδώ η γραμμική αύξηση της συχνότητας (*chirp*) ενός σταθερού αρχικά τόνου. Δείχνεται ότι το μέρος όπου η συχνότητα αυξάνεται είναι πιο προεξέχων από το σταθερό. Το φασματογράφημα ενός τέτοιου τόνου φαίνεται στο Σχήμα 5.1. Αρχικά ο τόνος έχει συχνότητα 500 Hz ενώ στη συνέχεια η συχνότητα του αυξάνεται έως τα 2 kHz εντός ενός δευτερολέπτου. Από την γραμμική αύξηση της συχνότητας αναμένουμε να επηρεαστούν τα loudness και fractal διάστασης (λόγω χειρισμού ψηφιακών σημάτων). Αντίθετα, η ενέργεια και το roughness του τόνου παραμένουν σχεδόν σταθερά, με μικρές μεταβολές στην αρχή (onset) των τόνων, και συνεισφέρουν στην καμπύλη σημαντικότητας μόνο εκεί. Στο ίδιο Σχήμα φαίνεται η καμπύλη σημαντικότητας του κάθε χαρακτηριστικού από την έξοδο του μοντέλου και η τελική καμπύλη σημαντικότητας. Βλέπουμε ότι η τελική καμπύλη λαμβάνει υψηλότερες τιμές στην περιοχή της γραμμικής αύξησης. Αντίστοιχα αποτελέσματα λήφθηκαν και για γραμμική μείωση της συχνότητας.

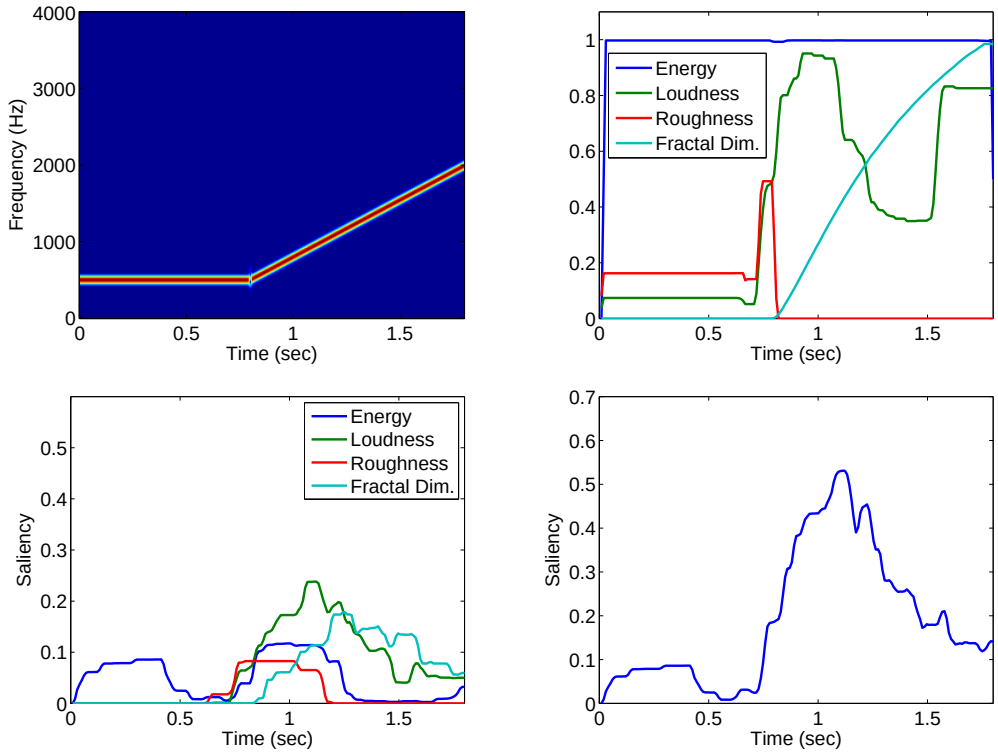
Εμφάνιση θορύβου

Ένα άλλο φαινόμενο το οποίο έλκει την ανθρώπινη προσοχή είναι η διακοπή του ηχητικού σήματος ξαφνικά από θόρυβο (noise burst). Όταν η διάρκεια του θορύβου είναι μερικά ms, δημιουργείται η ψευδαίσθηση της συνέχειας, δηλαδή ότι το ηχητικό σήμα εξακολουθεί να υπάρχει [6].

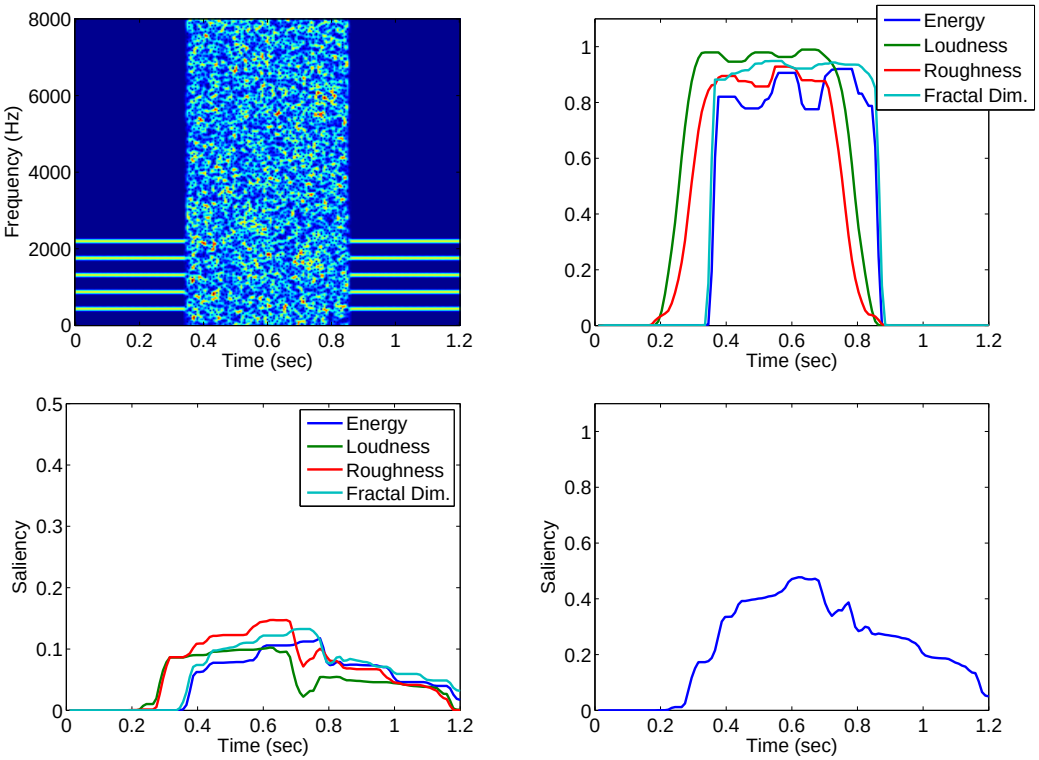
Στο Σχήμα 5.2 φαίνεται το φασματογράφημα ενός τέτοιου ήχου. Αρχικά υπάρχει ένας τόνος με πέντε αρμονικές συνιστώσες, κάποια στιγμή διακόπτεται από θόρυβο για διάρκεια 500 ms, και στη συνέχεια επανέρχεται. Τα σημεία εμφάνισης και διακοπής του θορύβου είναι αρκετά προεξέχοντα. Στο ίδιο Σχήμα φαίνονται οι καμπύλες των χαρακτηριστικών. Όλες οι καμπύλες λαμβάνουν μέγιστη τιμή κατά την χρονική διάρκεια ύπαρξης του θορύβου. Ομοίως οι καμπύλες σημαντικότητας έχουν τιμές κοντά στο μέγιστο στα σημεία που υπάρχει θόρυβος.

Αποσυγχρονισμός αρμονικής

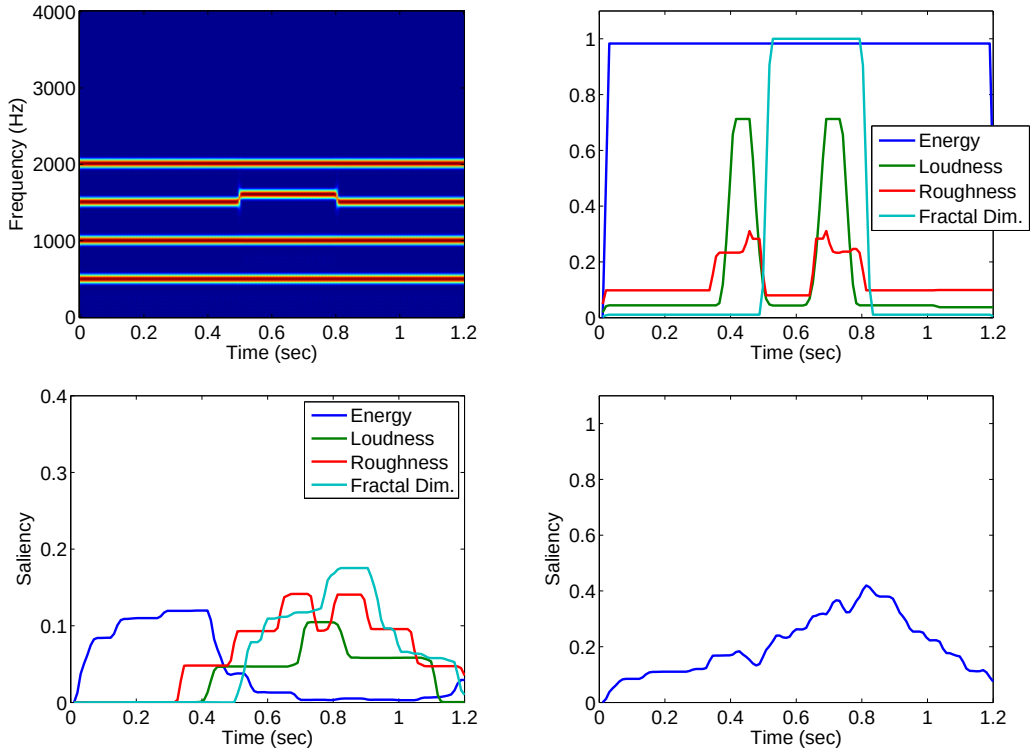
Ο αποσυγχονισμός μίας αρμονικής (mistuned harmonic) μιας συστοιχίας αρμονικών είναι ένα φαινόμενο το οποίο έχει παρατηρηθεί ότι έλκει την ανθρώπινη προσοχή [22, 47]. Υπάρχει η τάση η αποσυγχρονισμένη αρμονική να δημιουργεί την αίσθηση ότι ακούγονται δύο ηχητικά ροές (*streams*) αντί για ένα όταν η αποσυγχρόνιση διαρκεί μερικές εκατοντάδες msec. Ωστόσο, η αποσυγχρόνιση για μεγάλο χρονικό διάστημα οδηγεί στην ενσωμάτωση της αποσυχρονισμένης αρμονικής στις υπόλοιπες, και την μη διάκριση της.



Σχήμα 5.1: Γραμμική αύξηση συχνότητας. Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω: φασματογράφημα, καμπύλες χαρακτηριστικών, καμπύλες σημαντικότητας χαρακτηριστικών, και καμπύλη σημαντικότητας του ηχητικού σήματος (δείτε την έγχρωμη έκδοση).

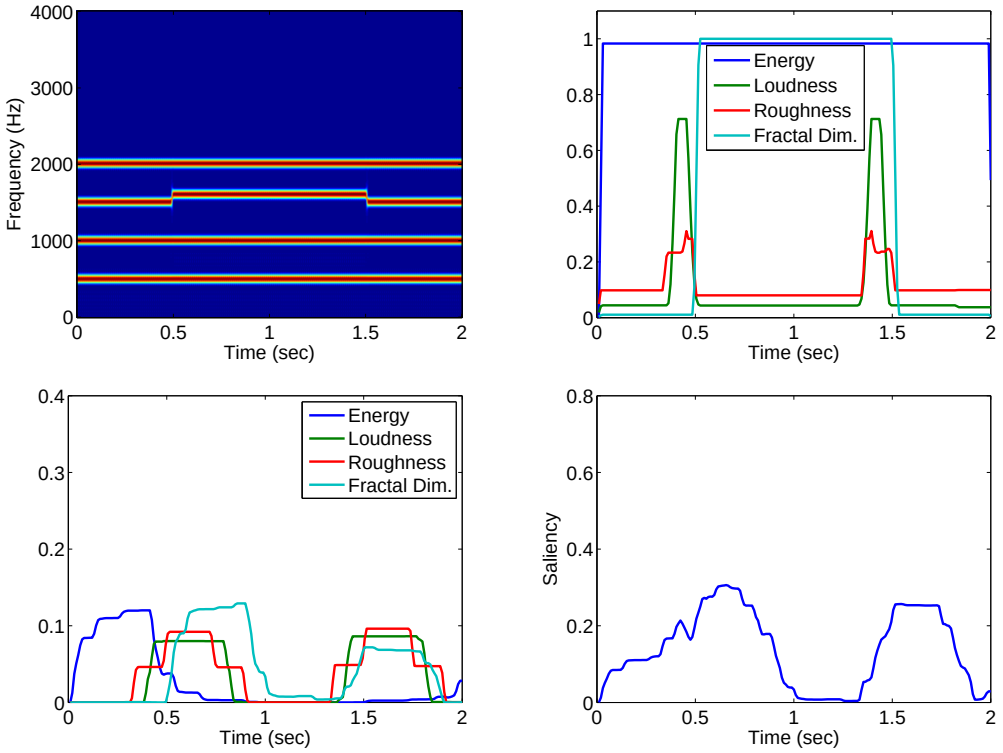


Σχήμα 5.2: Διακοπή συμπλέγματος αρμονικών από θόρυβο και επανεμφάνιση τους. Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω: φασματογράφημα, καμπύλες χαρακτηριστικών, καμπύλες σημαντικότητας χαρακτηριστικών, και καμπύλη σημαντικότητας του ηχητικού σήματος (δείτε την έγχρωμη έκδοση).



Σχήμα 5.3: Αποσυγχρονισμός αρμονικής μιας συστοιχίας αρμονικών για μικρή χρονική διάρκεια. Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω: φασματογράφημα, καμπύλες χαρακτηριστικών, καμπύλες σημαντικότητας χαρακτηριστικών, και καμπύλη σημαντικότητας του ηχητικού σήματος (δείτε την έγχρωμη έκδοση)

Έγινε έλεγχος της ικανότητας του μοντέλου να ανιχνεύει αποσυγχρονισμένες αρμονικές. Σε μια συστοιχία από τέσσερις αρμονικές (πολλαπλάσια των 500 Hz), μεταβλήθηκε η 3η αρμονική κατά 100 Hz για χρονικό διάστημα 400 ms, και έπειτα επανήλθε στην αρχική της συχνότητα. Στο Σχήμα 5.3 φαίνεται το φασματογράφημα του σήματος, οι χάρτες χαρακτηριστικών και οι καμπύλες σημαντικότητας τους. Το χαρακτηριστικό της ενέργειας παραμένει σταθερό κατά τη μεταβολή, στα loudness και roughness παρατηρούνται κορυφές λόγω της συνύπαρξης σε παράθυρα ανάλυσης συχνοτικού περιεχομένου με και χωρίς αποσυγχρονίση, ενώ η fractal διάσταση λαμβάνει μέγιστη τιμή καθ' όλη την διάρκεια λόγω της τάσης να λαμβάνει στην πράξη μεγαλύτερες τιμές σε τόνους μεγαλύτερης συχνότητας. Στην έξοδο του μοντέλου η ενέργεια έχει τονίσει την αρχή (onset) της σκηνής ενώ τα υπόλοιπα χαρακτηριστικά την μεταβολή της συχνότητας. Η τελική καμπύλη σημαντικότητας έδωσε κάποια σημαντικότητα στην αποσυγχρόνιση, αλλά όχι την μέγιστη. Αντίθετα μέγιστη τιμή σημαντικότητας δόθηκε στο συγχρονισμό και πάλι με τις υπόλοιπες αρμονικές.



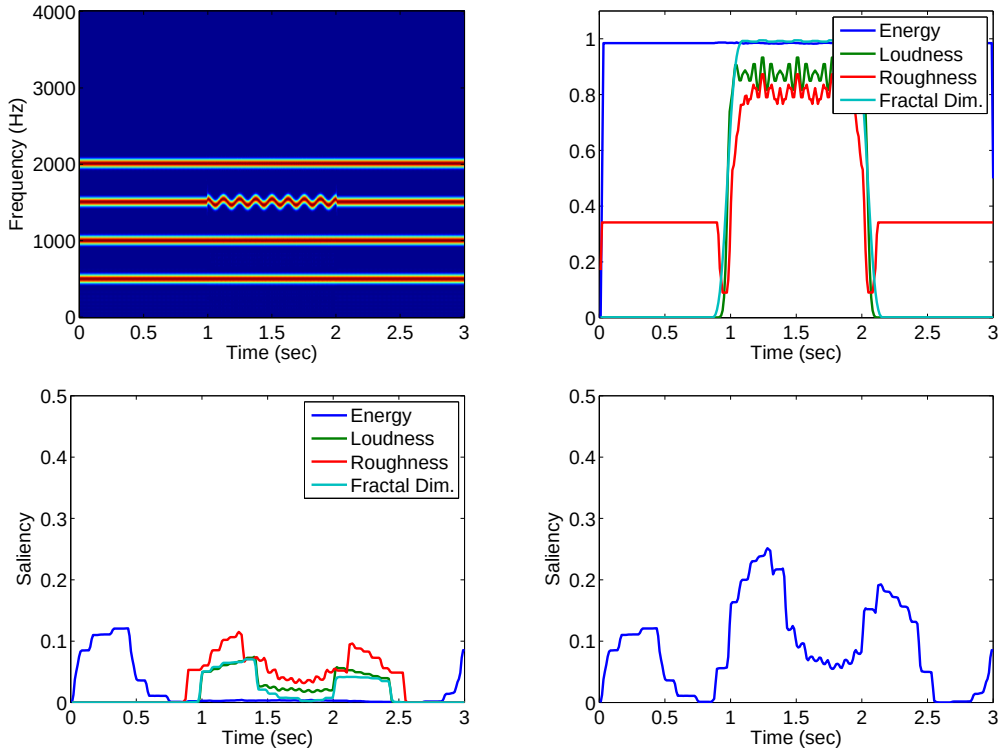
Σχήμα 5.4: Αποσυγχρονισμός αρμονικής μιας συστοιχίας αρμονικών για μεγάλη χρονική διάρκεια. Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω: φασματογράφημα, καμπύλες χαρακτηριστικών, καμπύλες σημαντικότητας χαρακτηριστικών, και καμπύλη σημαντικότητας του ηχητικού σήματος (δείτε την έγχρωμη έκδοση).

Δοκιμάστηκε στην συνέχεια η αύξηση της διάρκειας της αποσυγχρόνισης για τον έλεγχο του φαινομένου της ενσωμάτωσης της αποσυχρονισμένης στις υπόλοιπες αρμονικές. Η διάρκεια της αποσυγχόνισης έγινε ένα δευτερόλεπτο. Το μοντέλο ήταν σε θέση να ανιχνεύσει τόσο τις δύο μεταβολές αλλά και να θέσει στο υπόβαθρο την αποσυγχόνιση έπειτα από κάποιο χρονικό διάστημα. Στο Σχήμα 5.4 φαίνονται τα αποτελέσματα.

Διαμορφωμένος τόνος

Ένα πολύ γνωστό φαινόμενο στην ψυχοακουστική βιβλιογραφία είναι η τάση των διαμορφωμένων τόνων κατά πλάτος ή συχνότητα να ξεχωρίζουν σε περιβάλλον σταθερών τόνων, να εντάσσονται σε άλλο ηχητικό ρεύμα και να γίνονται άμεσα αντιληπτοί [43, 44]. Θα ελέγξουμε εάν το μοντέλο είναι σε θέση να ανιχνεύσει διαμορφώσεις κατά συχνότητα.

Σε μια συστοιχία 4 αρμονικών διαμορφώνουμε κατά συχνότητα την 3η αρμονική για χρονική διάρκεια ενός δευτερολέπτου, και έπειτα την επαναφέρουμε σε σταθερή κατάσταση. Χρησιμοποιούμε συχνότητα διαμόρφωσης 8 Hz ενώ ο δείκτης (βάθος) διαμόρφω-



Σχήμα 5.5: Διαμόρφωση τόνου κατά συχνότητα για κάποιο χρονικό διάστημα. Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω: φασματογράφημα, καμπύλες χαρακτηριστικών, καμπύλες σημαντικότητας χαρακτηριστικών, και καμπύλη σημαντικότητας του ηχητικού σήματος (δείτε την έγχρωμη έκδοση).

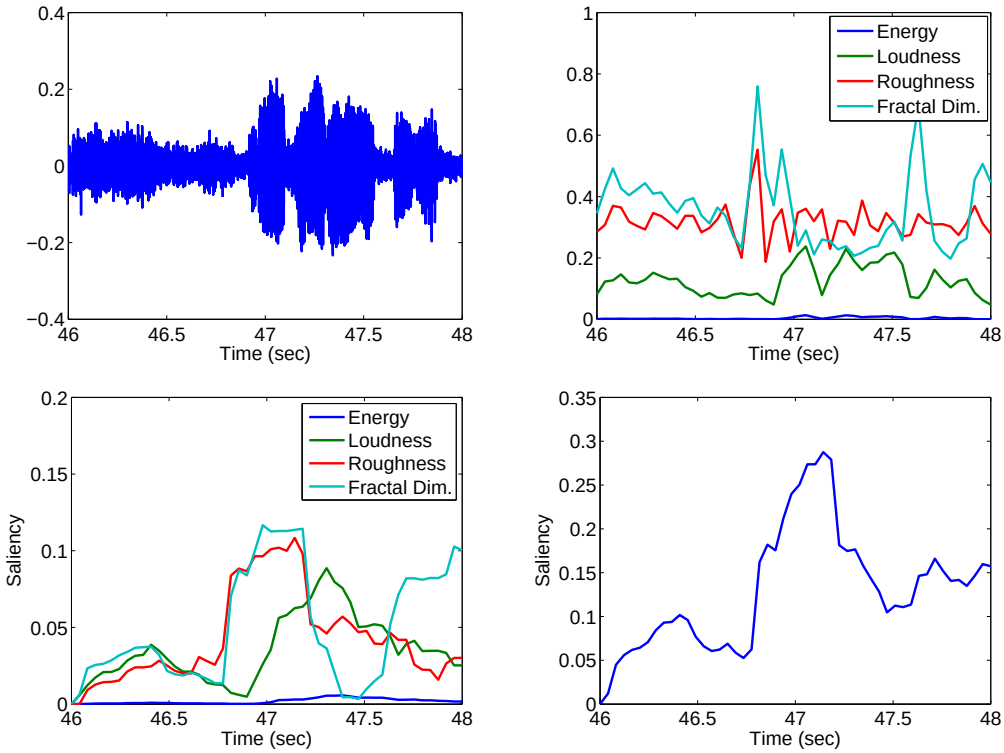
σης ισούται με 0.625. Στο Σχήμα 5.5 φαίνεται το φασματογράφημα του ηχητικού ερεθίσματος καθώς και η έξοδος του μοντέλου. Τα χαρακτηριστικά που συνεισφέρουν στην ανίχνευση των διαμορφώσεων είναι τα loudness, roughness, και fractal διάσταση.

5.1.2 Καμπύλες σημαντικότητας φυσικών ήχων

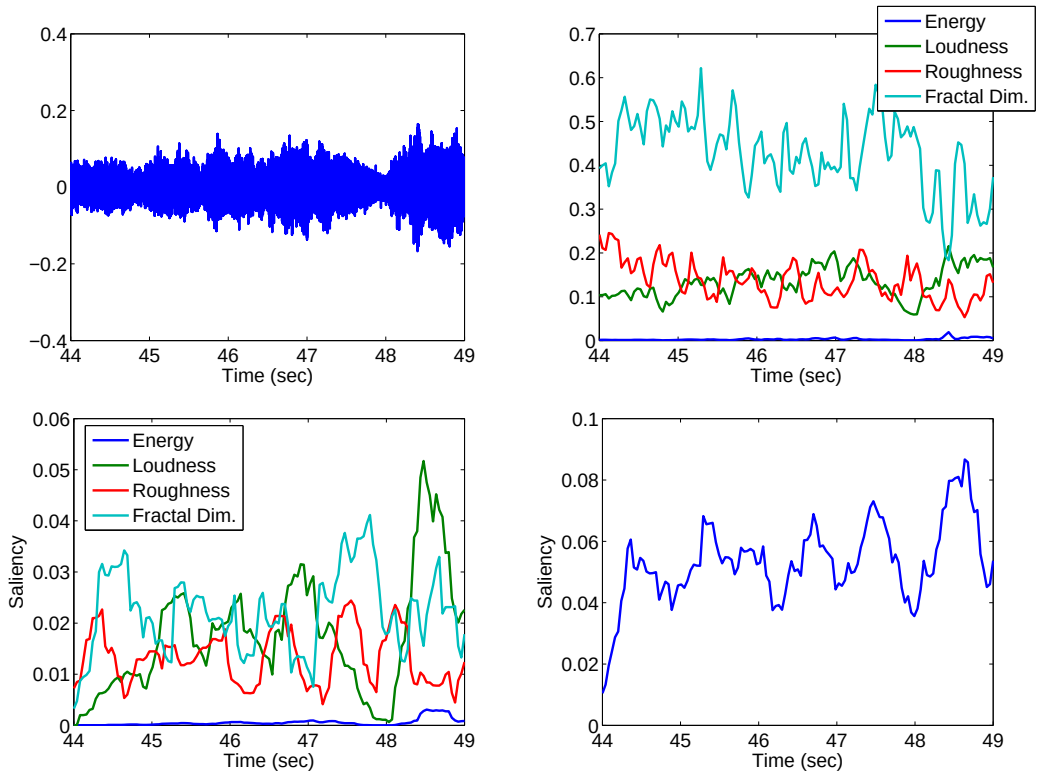
Μετά τον έλεγχο του μοντέλου σε συνθετικούς ήχους με απλή δομή, πραγματοποιείται έλεγχος σε πραγματικούς ήχους, με τους οποίους έρχεται σε επαφή καθημερινά ο άνθρωπος, και είναι αυτοί που θα καλείται να ταξινομεί σε κατηγορίες σημαντικότητας.

Η πρώτη σκηνή που δοκιμάστηκε αποτελείται από ήχο ανθρώπινης φωνής. Το μοντέλο έδωσε μέτρια σημαντικότητα στη σκηνή, όπως φαίνεται και στο Σχήμα 5.6, καθώς η ένταση της φωνής ήταν σχεδόν σταθερή, σε συνήθη επίπεδα συνομιλίας.

Μία δεύτερη σκηνή που δοκιμάστηκε αποτελείται από μουσική (Σχήμα 5.7. Το μοντέλο έδωσε μικρή σημαντικότητα στην σκηνή. Να παρατηρηθεί η περιοδικότητα των χαρακτηριστικών, καθώς και της τελικής καμπύλης σημαντικότητας που πηγάζει από περιοδικότητα στο ερέθισμα.



Σχήμα 5.6: Απόσπασμα ομιλίας από την ταινία Gladiator, με την φράση "he's cleverer than I thought". Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω: κυματομορφή, καμπύλες χαρακτηριστικών, καμπύλες σημαντικότητας χαρακτηριστικών, και καμπύλη σημαντικότητας του ηχητικού σήματος (δείτε την έγχρωμη έκδοση).



Σχήμα 5.7: Απόσπασμα μουσικής από την ταινία Lord of The Rings. Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω: κυματομορφή, καμπύλες χαρακτηριστικών, καμπύλες σημαντικότητας χαρακτηριστικών, και καμπύλη σημαντικότητας του ηχητικού σήματος (δείτε την έγχρωμη έκδοση).

5.2 Αξιολόγηση Μοντέλου μέσω Πειράματος των Kayser et al

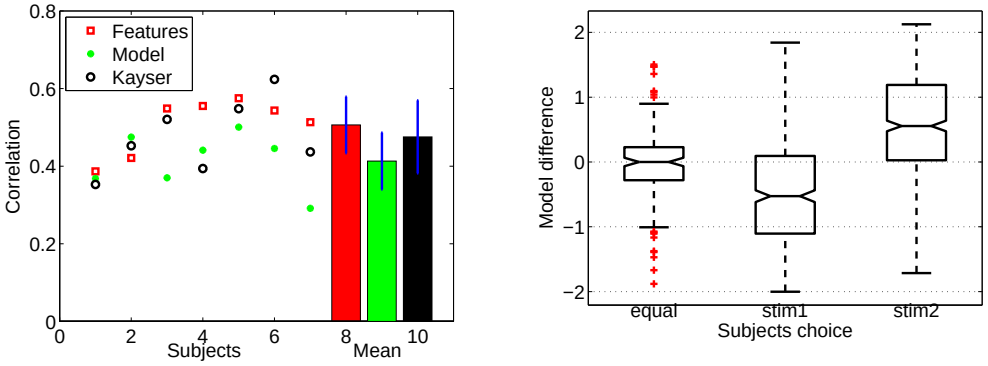
Όπως περιγράφηκε και σε προηγούμενη ενότητα ένας από τους τρόπους που οι Kayser et al [31] αξιολόγησαν το υπολογιστικό τους μοντέλο, είναι μέσω της σύγκρισης σημαντικότητας δύο ηχητικών σκηνών. Επειδή οι συγγραφείς παρέχουν τα πειραματικά τους δεδομένα για αυτό το πείραμα, θα τα χρησιμοποιήσουμε για να ελεγχθεί το υπολογιστικό μοντέλο που αναπτύχθηκε αλλά και για σύγκριση με τα δικά τους αποτελέσματα.

Το πείραμα αποτελούνταν από την ταυτόχρονη παρουσίαση ζευγών ηχητικών σκηνών σε άτομα, τα οποία έπρεπε να απαντήσουν ποια ηχητική σκηνή τους φάνηκε περισσότερο σημαντική ή εάν τους φάνηκαν εξίσου σημαντικές. Οι σκηνές αποτελούνταν από ήχους της φύσης (κελάηδισμα πουλιών), ήχους μηχανών, ενώ στο υπόβαθρο υπήρχε θόρυβος (*bubble noise*). Κάθε σκηνή είχε διάρκεια 3 sec, ενώ υπήρχαν συνολικά 52 διαφορετικές σκηνές η κάθε μία σε τρία θορυβώδη υπόβαθρα, ίδια για όλες τις σκηνές. Τα ζεύγη σύγκρισης είχαν το ίδιο θορυβώδες υπόβαθρο.Το πλήθος των ζευγών που παρουσιάστηκε σε κάθε άτομο ποικίλει από 150-160 (ίδια ζεύγη για κάθε άτομο), και το πλήθος των ατόμων είναι 7.

Για τη σύγκριση της σημαντικότητας των σκηνών, αρχικά πραγματοποιήθηκε εξαγωγή χαρακτηριστικών σε όλες τις σκηνές και υπολογίσθηκε η έξοδος του μοντέλου για κάθε χαρακτηριστικό. Για την επιλογή της πιο σημαντικής σκηνής με βάση το μοντέλο, ακολουθήθηκε η ίδια διαδικασία με αυτή των Kayser et al. Επιλέγονταν η σκηνή της οποίας η καμπύλη σημαντικότητας είχε μεγαλύτερο μέγιστο. Έπειτα υπολογίσθηκε η συσχέτιση με την επιλογή των χρηστών. Στον Πίνακα 5.1 φαίνεται η συσχέτιση κάθε χαρακτηριστικού/εξόδου του μοντέλου, με την επιλογή των χρηστών, καθώς και άνω φράγμα της p-τιμής(p-value), της πιθανότητας η τιμή συσχέτισης να έχει προέλθει από τύχη όταν η πραγματική συσχέτιση είναι μηδέν.

Μέγιστη συσχέτιση έχει το loudness (0.53), ενώ ακολουθούν fractal διάσταση (0.41), ενέργεια (0.34), και roughness (0.31). Οι p-τιμές είναι αρκετά χαμηλές και η βεβαιότητα ότι υπάρχει συσχέτιση είναι υψηλή. Συνδυάζοντας γραμμικά με ίσα βάρη τα χαρακτηριστικά επιτυγχάνεται συσχέτιση 0.51 ± 0.07 (μέση τιμή $\pm$ τυπική απόκλιση) , η οποία είναι υψηλότερη από το 0.47 ± 0.10 που επιτυγχάνει το μοντέλο των Kayser et al. Η έξοδος του μοντέλου έχει χαμηλότερη συσχέτιση από ότι τα χαρακτηριστικά, και συνολικά επιτυγχάνει 0.41 ± 0.04 η οποία είναι χαμηλότερη από τους Kayser et al. Η πτώση της απόδοσης οφείλεται κυρίως στα χαρακτηριστικά των roughness, και fractal διάστασης.

Στο Σχήμα 5.8α΄ φαίνεται η συσχέτιση με χρήση του συνδυασμού των χαρακτηριστικών και της εξόδου του μοντέλου, για κάθε άτομο που συμμετείχε στο πείραμα. Φαίνεται, επίσης, η συσχέτιση του μοντέλου των Kayser et al με κάθε άτομο. Ο συνδυασμός των χαρακτηριστικών έχει συστηματικά υψηλότερη συσχέτιση από τους Kayser et al, ενώ η έξοδος του μοντέλου για ορισμένα άτομα είναι ελαφρώς μεγαλύτερη, και σε άλλα αρκετά μικρότερη. Στο Σχήμα 5.8β΄ φαίνονται διαγράμματα πλαισίου (*box plots*) για την διαφορά σημαντικότητας της δεύτερης σκηνής από την πρώτη όπως υπολογίσθηκε από τον συνδυασμό των χαρακτηριστικών. Τα δεδομένα είναι ομαδοποιημένα σε τρεις κατηγορίες με



(α΄) (β΄)

Σχήμα 5.8: Αριστερά: συσχέτιση συνδυασμού χαρακτηριστικών, εξόδου του μοντέλου, και μοντέλο των Kayser et al με τις επιλογές των χρηστών στο πείραμα σύγκρισης σκηνών. Δεξιά: box plots για την διαφορά σημαντικότητας της σκηνής 2 - σκηνής 1 όπως υπολογίσθηκε από τον συνδυασμό χαρακτηριστικών.

Πίνακας 5.1: Μέση συσχέτιση μεταξύ ταξινόμησης χρηστών και των χαρακτηριστικών / έξοδο μοντέλου στο πείραμα σύγκρισης σκηνών των Kayser et al, και άνω φράγματα για τις p τιμές.

Feature	En	Loud	Rough	Frd	Combine
Corr	0.34 / 0.36	0.53 / 0.49	0.31 / 0.27	0.41 / 0.16	0.51 / 0.41
p Value	$6 \cdot 10^{-4}$	$5 \cdot 10^{-9}$	$2 \cdot 10^{-2}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-7}$

βάση ποια σκηνή θεώρησαν οι χρήστες πιο προεξέχουσα ή αν τις θεώρησαν εξίσου προεξέχουσες. Στην περίπτωση που οι χρήστες επέλεξαν κάποια σκηνή, οι μέσοι (medians) καθώς και τα μεσαία τεταρτημόρια (interquartiles) είναι εμφανώς μετατοπισμένα προς την σκηνή που επέλεξαν οι χρήστες. Όταν οι χρήστες θεώρησαν εξίσου σημαντικές τις δύο σκηνές, ο μέσος ήταν κοντά στο μηδέν και τα μεσαία τεταρτημόρια αρκετά μικρότερου εύρους.

Η συσχέτιση κάθε χαρακτηριστικού με την ανθρώπινη ταξινόμηση μπορεί να χρησιμοποιηθεί επίσης ως βάρος για τον συνδυασμό των χαρακτηριστικών και τη δημιουργία της τελικής καμπύλης σημαντικότητας. Είναι ένας από τους συνδυασμούς που δοκιμάζεται.

5.3 Κατωφλίωση των Καμπυλών

Όπως και στην περίπτωση των χαρτών σημαντικότητας, αρχικά πραγματοποιείται έλεγχος του μοντέλου εξαγωγής καμπύλης σημαντικότητας με χρήση κατωφλίου. Γίνεται επιλογή ενός σταθερού κατωφλίου και τα σημεία της καμπύλης τα οποία βρίσκονται πάνω

από αυτό ταξινομούνται ως σημαντικά. Τα υπόλοιπα ταξινομούνται ως μη-σημαντικά.

Εκτός από την έξοδο του μοντέλου, καμπύλη σημαντικότητας δημιουργείται επίσης με χρήση άμεσα των χαρακτηριστικών που εξήχθησαν από το σήμα. Ο συνδυασμός τους γίνεται με τον ίδιο τρόπο που συνδυάζονται οι καμπύλες σημαντικότητας για την δημιουργία της τελικής καμπύλης, αρχικά κλιμακώνονται στο διάστημα $[0, 1]$ και έπειτα συνδυάζονται γραμμικά.

Στον Πίνακα 5.2 φαίνονται τα αποτελέσματα ταξινόμησης με χρήση κάθε χαρακτηριστικού ξεχωριστά αλλά και συνδυάζοντας τα γραμμικά με ίσα βάρη. Αριστερά της μπάρας είναι το αποτέλεσμα με χρήση της εξόδου του μοντέλου ενώ δεξιά χρησιμοποιώντας άμεσα τα χαρακτηριστικά. Παρατηρήθηκε ότι ήταν αρκετά ωφέλιμο τα χαρακτηριστικά και οι καμπύλες να φιλτραριστούν με φίλτρο μέσου (median) πριν συνδυαστούν και γίνει η ταξινόμηση. Το μήκους του φίλτρου είναι 50 δείγματα (2 sec.). Παρόμοια αποτελέσματα λήφθηκαν συνδυάζοντας τα χαρακτηριστικά με βάρη ανάλογα της συσχέτισης τους με την ανθρώπινη θεώρηση της σημαντικότητας, από το πείραμα σύγκρισης σκηνών των Kayser et al.

Να σημειώσουμε ότι οι τιμές που φαίνονται στον Πίνακα 5.2 (αλλά και σε κάθε σχετικό πίνακα) υπολογίστηκαν επιλέγοντας για κάθε χαρακτηριστικό κάποιο κατώφλι και ταξινομώντας. Το κατώφλι επιλέχθηκε έτσι ώστε να λαμβάνεται υψηλό accuracy. Θα μπορούσε να τεθεί ένα χαμηλότερο κατώφλι και να επιτευχθεί υψηλότερο recall, θυσιάζοντας ενδεχομένως κάποια ακρίβεια στην ταξινόμηση. Για παράδειγμα, στα roughness, fractal διάσταση εμφανίζονται υψηλές τιμές recall διότι παρατηρήθηκε ότι υψηλότερες τιμές κατωφλίου δεν αύξαναν την ακρίβεια της ταξινόμησης, και έτσι δεν επιλέχθηκαν. Αντίθετα, για τα χαρακτηριστικά ενέργεια και loudness, αύξηση του κατωφλίου οδηγούσε σε αύξηση της ακρίβειας και επιλέχθηκε ένα κατώφλι, ώστε να επιτυγχάνεται ταυτόχρονα υψηλό accuracy και υψηλό recall.

Για να ελεγχθεί η δυνατότητα κάθε χαρακτηριστικού να διακρίνει τις δύο κλάσεις σημαντικότητας, υπολογίσθηκαν ROC καμπύλες οι οποίες φαίνονται στο Σχήμα 5.9 καθώς και το εμβαδόν τους (Πίνακας 5.3). Μέγιστο εμβαδόν με τη διαγώνιο έχει το χαρακτηριστικό της ενέργειας και ακολουθούν τα loudness, roughness, και fractal διάσταση. Επίσης, τα χαρακτηριστικά χωρίς την χρήση του μοντέλου έχουν μεγαλύτερη διακριτική ικανότητα από ότι με αυτό, σύμφωνα με τις καμπύλες.

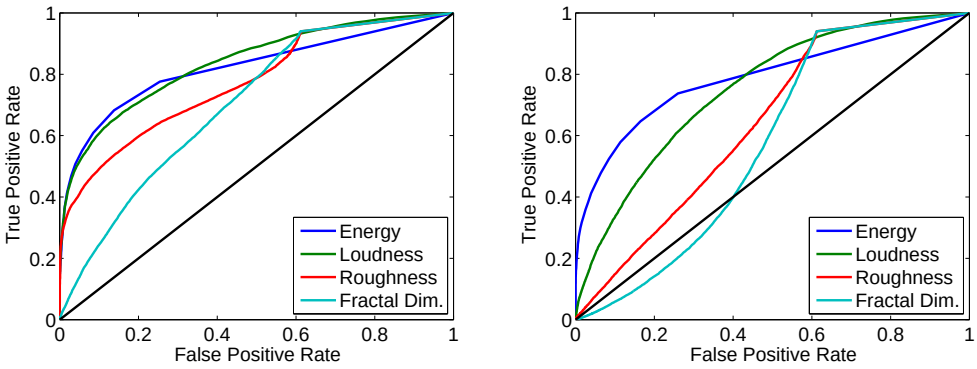
5.4 Ταξινόμηση με χρήση SVM

Έχει παρατηρηθεί ότι υπάρχει σημαντική επικάλυψη μεταξύ των ήχων που θεωρούνται σημαντικοί από τους χρήστες και αυτών που δεν θεωρούνται σημαντικοί. Ως συνέπεια, κατά την εξαγωγή χαρακτηριστικών από τις σκηνές, παρατηρείται αντίστοιχα σημαντική επικάλυψη μεταξύ των δύο κατηγοριών με αποτέλεσμα να γίνεται δύσκολος ο διαχωρισμός τους και η ορθή ταξινόμηση τους.

Προσπάθεια υπέρβασης αυτών των εμποδίων, γίνεται μέσω της χρήσης των *Support Vector Machines* (SVM) ταξινομητών. Τα SVM έχουν το πλεονέκτημα ότι μπορούν να απεικονίσουν τα δεδομένα σε έναν χώρο αυθαίρετης διάστασης, και με κατάλληλη επιλογή του πυρήνα να γίνουν διαχωρίσιμα ώστε να ταξινομηθούν σε αυτό τον χώρο με ένα

Πίνακας 5.2: Αποτελέσματα ταξινόμησης με χρήση κατωφλίου στα χαρακτηριστικά του χρονικού χάρτη σημαντικότητας. Αριστερά της πλάγιας μπάρας η έξοδος του μοντέλου, και δεξιά τα χαρακτηριστικά.

Feature	Acc(%)	Prec(%)	Rec(%)
Energy	73.1 / 76.0	71.9 / 76.8	80.7 / 78.0
Loudness	68.9 / 74.1	68.2 / 73.2	77.0 / 80.6
Roughness	65.0 / 65.1	62.3 / 62.6	85.3 / 84.5
Fractal Dim.	64.2 / 65.9	61.8 / 63.6	84.0 / 82.9
Loud.+FrDim.	64.6 / 66.3	63.5 / 64.5	77.2 / 80.5
Rough.+FrDim.	64.0 / 65.0	61.8 / 62.9	83.4 / 82.1
En.+Loud.	69.2 / 74.2	67.1 / 73.0	81.8 / 81.1
En.+Loud.+Rough.	65.4 / 71.2	65.4 / 72.8	72.9 / 72.5
En.+Loud.+Rough.+FrDim.	64.7 / 66.3	63.0 / 64.5	80.3 / 80.5



Σχήμα 5.9: Αριστερά: καμπύλες ROC για τα χαρακτηριστικά, και δεξιά για την έξοδο του μοντέλου.

Πίνακας 5.3: Εμβαδόν μεταξύ καμπύλης ROC και διαγωνίου για τα χαρακτηριστικά (πρώτη γραμμή), και την έξοδο του μοντέλου (δεύτερη γραμμή).

Feature	En.	Loud	Rough	FrD
AUC	0.3219	0.3378	0.2741	0.2024
AUC	0.2899	0.2527	0.1417	0.0767

υπέρ-επίπεδο.

Τα SVM ταξινομούν δεδομένα με χρήση υπερ-επιπέδων σε έναν χώρο αυθαίρετης διάστασης. Στηρίζονται στην θεωρία στατιστικής μάθησης (*statistical learning theory*) [63]. Έστω ένα πρόβλημα ταξινόμησης σε δύο κατηγορίες, όπου τα δεδομένα είναι γραμμικά διαχωρίσιμα, δηλαδή μπορούν να διαχωρισθούν με ένα επίπεδο. Στόχος των SVM είναι να μεγιστοποιήσουν την απόσταση των δεδομένων από το επίπεδο, διατηρώντας τα ταυτόχρονα διαχωρίσιμα δηλαδή τα δεδομένα που ανήκουν στην ίδια κλάση να βρίσκονται στην ίδια πλευρά του επιπέδου. Η συνθήκη της μεγιστοποίησης της απόστασης αυξάνει την πιθανότητα της ορθής ταξινόμησης νέων δεδομένων καθώς αυτά είναι πιθανό να βρίσκονται κοντά στα υπάρχοντα.

Δοθέντων σημείων $x_i \in \mathbb{R}^n,\ i = 1, 2, \ldots l$, όπου κάθε ένα ανήκει σε μία εκ των δύο κατηγοριών y_i, $y_i \in \{1, -1\}$, τα SVM βρίσκουν επίπεδο w ώστε κάθε νέο δεδομένο να ταξινομείται με βάση το πρόσημο της ποσότητας $w^T x$, $sgn(w^T x)$. Η εύρεση του w μπορεί να γίνει λύνοντας το ακόλουθο πρόβλημα βελτιστοποίησης [4, 11]:

$$\min_{w,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{5.1α΄}$$

$$y^{(i)}(w^T \phi(x_i)) \geq 1 - \xi_i \tag{5.1β΄}$$

$$\xi_i \geq 0,\ i = 1, 2, \ldots, l \tag{5.1γ΄}$$

Ο πρώτος όρος της εξίσωσης 5.1α΄ σε συνδυασμό με τις συνθήκες 5.1β΄, 5.1γ΄ διασφαλίζει την μεγιστοποίηση της απόστασης των δεδομένων από το επίπεδο. Τα διανύσματα εκπαίδευσης που απέχουν απόσταση ίση με 1 από το επίπεδο καλούνται διανύσματα υποστήριξης (*support vectors*). Ο δεύτερος όρος είναι όρος σφάλματος και επιτρέπει σε μερικά από τα δεδομένα εκπαίδευσης να απέχουν από το επίπεδο λιγότερο από ότι τα διανύσματα υποστήριξης και πιθανώς να ταξινομηθούν σε λάθος κατηγορία. Για κάθε διάνυσμα που απέχει απόσταση μικρότερη από 1, αυξάνεται η τιμή του συναρτησιακού ελαχιστοποίησης κατά $C \cdot \xi_i,\ C > 0$. Το να επιτρέπεται σε μερικά από τα δεδομένα να απέχουν απόσταση μικρότερη του 1, οδηγεί συνήθως σε υψηλότερα ποσοστά ορθής ταξινόμησης καθώς ο ταξινομητής γίνεται λιγότερο ευάλωτος σε θόρυβο, και έχει μικρότερη εξάρτηση από τα δεδομένα εκπαίδευσης.

Η φύση του προβλήματος ταξινόμησης πιθανώς να είναι τέτοια που τα δεδομένα να μην είναι γραμμικά διαχωρίσιμα στον χώρο που βρίσκονται, ή να μην είναι διαχωρίσιμα λόγω άλλων παραγόντων. Έχει φανεί αρκετά αποδοτικό σε τέτοιες περιπτώσεις, τα δεδομένα να απεικονίζονται σε έναν χώρο μεγαλύτερης διάστασης, πριν πραγματοποιηθεί εκπαίδευση. Αυτό γίνεται μέσω της συνάρτησης ϕ. Για δύο διανύσματα $x,\ y\ \in \mathbb{R}^n$, η συνάρτηση δύο μεταβλητών K, με $K(x, y) = \phi(x)^T \phi(y)$, καλείται πυρήνας του ταξινομητή. Είναι εύκολο να δειχθεί ότι για την λύση του προβλήματος 5.1α΄ δεν απαιτείται ο υπολογισμός της συνάρτησης ϕ για κάθε διάνυσμα, αλλά μόνο του πυρήνα για ζεύγη διανυσμάτων υποστήριξης, το οποίο μειώνει σημαντικά το υπολογιστικό κόστος.

Κατά την επίλυση προβλημάτων βελτιστοποίησης, αρκετά συχνά γίνεται χρήση της δυαδικότητας Lagrange [5], η οποία επιτρέπει την επίλυση ενός προβλήματος το οποίο

έχει την ίδια λύση με το αρχικό αλλά μικρότερο υπολογιστικό κόστος. Η δυαδική μορφή του προβλήματος 5.1 είναι η ακόλουθη:

$$\min_{\alpha} \quad \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \tag{5.2α΄}$$

$$y^T\alpha = 0 \tag{5.2β΄}$$

$$0 \leq \alpha_i \leq C,\ i = 1, 2, \ldots, l \tag{5.2γ΄}$$

όπου $e = [1, \ldots, 1]^T$ διάνυσμα με όλα τα στοιχεία ίσα με μονάδα, και Q ένας l-επί-l θετικά ημιορισμένος πίνακας με $Q_{ij} = y_i y_j K(x_i, x_j)$. Λόγω της μεγάλης συνήθως διάστασης του διανύσματος w, η λύση του δυαδικού προβλήματος είναι αποδοτικότερη.

5.4.1 Δημιουργία διανυσμάτων

Για να γίνει ταξινόμηση με χρήση των SVM ταξινομητών, απαιτείται η δημιουργία διανυσμάτων από τις σκηνές. Τα διανύσματα δομούνται από τα χαρακτηριστικά που περιγράφηκαν στο Κεφάλαιο **??**. Για τη δημιουργία διανυσμάτων επιλέγεται αρχικά ένα χρονικό παράθυρο κάποιας διάρκειας και μερικά από τα χαρακτηριστικά. Κάθε διάνυσμα αποτελείται από την συνένωση (concatenation) των δειγμάτων κάθε χαρακτηριστικού εντός του χρονικού παραθύρου. Γίνεται πειραματισμός με επιλογή διαφορετικών χρονικών παραθύρων τα οποία κυμαίνονται από 40 ms, όπου μόνο ένα δείγμα από κάθε χαρακτηριστικό χρησιμοποιείται σε κάθε διάνυσμα, έως 2 sec.

Η εκπαίδευση των SVM γίνεται χρησιμοποιώντας δεδομένα από πέντε ταινίες και γίνεται έλεγχος της απόδοσης σε μία. Για κάθε ταινία στην οποία γίνεται έλεγχος, εκπαιδεύεται και διαφορετικό SVM. Τα αποτελέσματα που παρουσιάζονται σε επόμενες ενότητες, είναι μέσοι όροι πάνω σε όλους τους δυνατούς συνδυασμούς από δεδομένα εκπαίδευσης.

5.4.2 Ταξινόμηση με γραμμικό πυρήνα

Αρχικά δοκιμάστηκε η ταξινόμηση με χρήση γραμμικού πυρήνα για τα SVM, δηλαδή όπου $\phi(x) = x$. Για την εύρεση του επιπέδου w λύνεται αντί της εξίσωσης 5.3 η ακόλουθη εξίσωση:

$$\min_{w} \frac{1}{2}w^T w + C \cdot \sum_{i=1}^{l} \max(0, 1 - y_i w^T x_i)^l \tag{5.3}$$

όπου $l = 1$ ή 2, αντιστοιχεί στην χρήση l_1 ή l_2 νόρμας, αντίστοιχα. Στην περίπτωση της l_1 νόρμας, η δυαδική μορφή της εξίσωσης 5.3 είναι ίδια με αυτήν της 5.1 όταν χρησιμοποιείται γραμμικός πυρήνας. Στην περίπτωση της l_2 νόρμας, οι όροι α_i δεν είναι άνω φραγμένοι, και στην διαγώνιο του πίνακα Q αρκεί να προστεθεί ο όρος $D = 1/(2C)$. Η λύση της εξίσωσης 5.3 έχει το πλεονέκτημα του μικρότερου υπολογιστικού κόστους και της ταχύτερης προσέγγισης του βέλτιστου διανύσματος w.

Σε πρώτο στάδιο γίνεται δοκιμή με χρήση παραθύρου 40 ms, δηλαδή λαμβάνοντας ένα δείγμα από κάθε χαρακτηριστικό για κάθε συνιστώσα του διανύσματος. Στο Σχήμα

5.10 φαίνεται η κατανομή των συνιστωσών σε δύο και τρεις διαστάσεις, διανυσμάτων αντίστοιχου μήκους, για κάποια από τα χαρακτηριστικά από την έξοδο του μοντέλου. Τα σημεία ίδιου χρώματος αντιστοιχούν στην ίδια κλάση σημαντικότητας, όπου με κόκκινο συμβολίζονται τα σημαντικά και με μπλε τα μη-σημαντικά όπως σημειώθηκαν από τους χρήστες. Αυτός ο χρωματικός κώδικας διατηρείται σε όλη την έκταση της εργασίας και δεν θα αναφέρεται στη συνέχεια. Στο ίδιο σχήμα φαίνεται, επίσης, το επίπεδο που υπολογίσθηκε από το SVM για τη διάκριση τους. Η επικάλυψη των δύο κλάσεων είναι εμφανής. Μικρή διάκριση φαίνεται να υπάρχει στη διάσταση του loudness για τα χαρακτηριστικά του σχήματος. Εξίσου σημαντική επικάλυψη παρατηρήθηκε και μεταξύ άλλων χαρακτηριστικών. Μέγιστη διακριτική ικανότητα ως προς τις κλάσεις σημαντικότητας φαίνεται πως έχουν η ενέργεια και το loudness, και ακολουθούν τα roughness και fractal διάσταση.

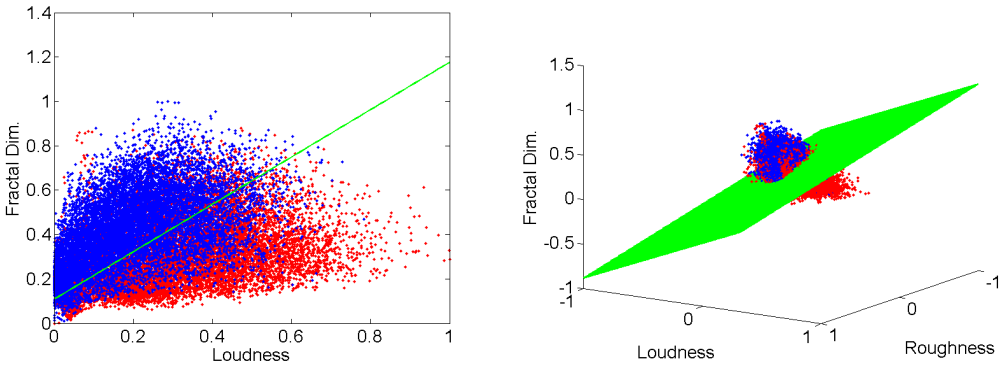
Στον Πίνακα 5.4 φαίνονται τα αποτελέσματα ταξινόμησης για κάποιους συνδυασμούς χαρακτηριστικών. Μέγιστη ακρίβεια επιτεύχθηκε με χρήση των loudness και roughness, ενώ μέγιστο recall με χρήση ενέργειας και loudness. Το χαρακτηριστικό που φαίνεται να συνεισφέρει περισσότερο στην ακρίβεια της ταξινόμησης είναι το loudness. Όταν υπάρχει στο σύνολο των χαρακτηριστικών λαμβάνεται πάντα η μέγιστη ακρίβεια και υψηλό recall, ανεξάρτητα από το ποια είναι τα υπόλοιπα χαρακτηριστικά. Με χρήση της ενέργειας αντί του loudness λαμβάνεται επίσης υψηλό recall, αλλά περίπου 5% μικρότερη ακρίβεια σε σύγκριση με αυτό. Η χρήση των roughness και fractal διάστασης χωρίς την ενσωμάτωση ενέργειας ή loudness, δίνει πολύ χαμηλή ακρίβεια (κοντά στο επίπεδο της τύχης), και δείχνει ότι από μόνα τους αυτά τα δύο χαρακτηριστικά δεν είναι αρκετά για την διάκριση των δύο κλάσεων. Ο ρόλος τους, είναι κυρίως να συμπληρώνουν τα άλλα δύο χαρακτηριστικά.

Δοκιμάστηκε στη συνέχεια μεταβολή του χρονικού παραθύρου στο οποίο λαμβάνονται δείγματα από τα χαρακτηριστικά για την δημιουργία διανυσμάτων και πραγματοποιήθηκε ταξινόμηση. Στον Πίνακα 5.5 φαίνονται τα ποσοστά ταξινόμησης με χρήση όλων των χαρακτηριστικών για χρονικά παράθυρα που κυμαίνονται από 40 ms έως 3 sec. Με μεταβολή του μήκους του παραθύρου δεν παρατηρήθηκε ιδιαίτερη μεταβολή στην κλάση που ταξινομείται κάθε ηχητική σκηνή. Τα αποτελέσματα για μεγαλύτερα παράθυρα είναι κοντά σε αυτά που επιτεύχθηκαν με παράθυρο 40 ms (ένα μόνο δείγμα ανά χαρακτηριστικό). Παρόμοια συμπεριφορά στην ταξινόμηση παρατηρήθηκε και για άλλους συνδυασμούς χαρακτηριστικών. Απαιτείται κάποια άλλη επεξεργασία από την απλή συνένωση των συνιστωσών των χαρακτηριστικών εντός του χρονικού παραθύρου.

5.4.3 Ταξινόμηση με Gaussian (RBF) πυρήνα

Στην περίπτωση του Gaussian πυρήνα η εφαρμογή της συνάρτησης ϕ, απεικονίζει τα δεδομένα σε έναν χώρο άπειρης διάστασης. Ο πυρήνας K έχει την εξής μορφή: $K(i,j) = \phi(x_i)^T\phi(x_j) = \exp(-\gamma||x_i - x_j||^2)$, όπου η θετική σταθερά γ επιλέγεται από εμάς. Έχει δειχθεί ότι ο γραμμικός πυρήνας είναι ειδική περίπτωση ενός Gaussian πυρήνα με άλλες παραμέτρους [32].

Για την ταξινόμηση με RBF πυρήνα, αρχικά πρέπει να προσδιορισθεί η παράμετρος γ. Η παράμετρος γ μπορεί να θεωρηθεί ως ένα μέτρο της συσχέτισης μεταξύ των διανυ-



Σχήμα 5.10: Διανύσματα χαρακτηριστικών για μήκος παραθύρου 40 ms, και διαχωριστικό επίπεδο όπως υπολογίστηκε από το SVM γραμμικού πυρήνα. Με κόκκινο σημειώνονται τα σημαντικά και με μπλε τα μη-σημαντικά σύμφωνα με την σημείωση των χρηστών.

Πίνακας 5.4: Ταξινόμηση με SVM γραμμικού πυρήνα και παράθυρο διάρκειας ίσο με 40 ms με χρήση των χαρακτηριστικών.

Feature	Acc(%)	Prec(%)	Rec(%)
En.+Loud.	74.3	76.3	76.7
En+Rough	71.9	76.4	72.9
En+FrD	72.7	77.1	72.5
Loud+Rough	75.5	78.5	74.6
Loud+FrD	74.6	77.2	75.5
Rough+FrD	61.9	64.8	73.6
En+Loud+R	75.3	78.1	75.1
En+L+F	74.5	76.9	75.9
E+R+F	72.3	75.8	73.9
L+R+F	75.3	78.3	74.5
All four	75.1	77.9	74.9

Πίνακας 5.5: Αποτελέσματα ταξινόμησης με χρήση όλων των χαρακτηριστικών στις συνιστώσες διανυσμάτων, για διαφορετικά μήκη παραθύρων.

Win(sec)	Dims.	Acc(%)	Prec(%)	Rec(%)
0.04	4	75.1	77.90	74.9
0.2	20	75.0	78.1	74.7
0.6	60	74.6	78.0	75.1
1	100	74.7	78.4	75.6
1.6	160	74.4	78.2	76.0
2	200	75.0	79.0	76.7
3	300	74.6	79.2	76.4

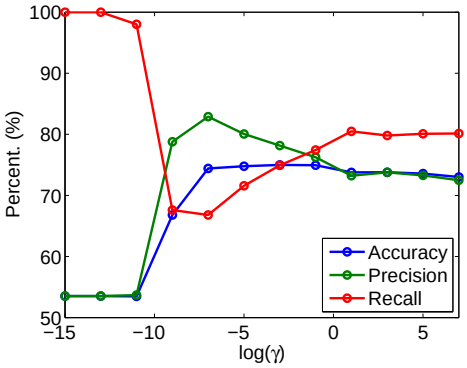
σμάτων. Για τον προσδιορισμό της παραμέτρου πραγματοποιήθηκε ταξινόμηση με χρήση διάφορων τιμών της, χρησιμοποιώντας κάποια από τα χαρακτηριστικά ως συνιστώσες των διανυσμάτων. Έγινε χρήση τιμών που είναι δυνάμεις του δύο, από 2^{-15} έως 2^{7}. Στο Σχήμα 5.11 φαίνεται η μεταβολή των μέτρων accuracy, precision, recall ως συνάρτηση του γ χρησιμοποιώντας όλα τα χαρακτηριστικά. Τα accuracy, precision λαμβάνουν μέγιστη τιμή για τιμές του γ κοντά στη μονάδα. Για πολύ μικρές τιμές της παραμέτρου, όλα τα δεδομένα ταξινομούνται ως σημαντικά και φαίνεται πως δεν είναι κατάλληλες για ταξινόμηση. Επιλέχθηκε $\gamma = 1$, και όλα τα αποτελέσματα που παρουσιάζονται στη συνέχεια εξήχθησαν με χρήση αυτής της τιμής.

Στον Πίνακα 5.6 φαίνονται αποτελέσματα ταξινόμησης για κάποιους συνδυασμούς χαρακτηριστικών. Τα συμπεράσματα είναι παρόμοια με αυτά του γραμμικού πυρήνα. Η ύπαρξη του loudness μεταξύ των χαρακτηριστικών δίνει την μέγιστη ακρίβεια ταξινόμησης. Με την ενέργεια λαμβάνεται μέτρια απόδοση, ενώ η χρήση μόνο roughness και fractal διάστασης δεν αποδίδει. Μέγιστο recall επιτεύχθηκε και πάλι με χρήση ενέργειας και loudness, ενώ μέγιστη ακρίβεια με loudness και roughness. Η προσθήκη της ενέργειας στο σύνολο των χαρακτηριστικών όταν υπάρχει το loudness φαίνεται πως δεν ωφελεί. Ως συνέπεια αυτών, με χρήση όλων των χαρακτηριστικών επιτεύχθηκε παρόμοια απόδοση με χρήση υποσύνολο τους. Να σημειωθεί επίσης ότι υψηλό ποσοστό των σημείων ($\approx$ 50%) λαμβάνονται ως διανύσματα υποστήριξης για κάθε συνδυασμό χαρακτηριστικών που δοκιμάστηκε. Αυτό οφείλεται στην σημαντική επικάλυψη των δύο κλάσεων.

Σε σύγκριση με τον γραμμικό πυρήνα δεν επιτεύχθηκε βελτίωση στην ταξινόμηση. Στο Σχήμα 5.12 φαίνονται διαχωριστικές καμπύλες και επιφάνειες για κάποιους συνδυασμούς χαρακτηριστικών, τόσο με χρήση Gaussian πυρήνα όσο και άλλων. Οι διαχωριστικές καμπύλες είναι κοντά μεταξύ τους και οδηγούν σε όμοιο αποτέλεσμα ταξινόμησης. Η χρήση διαφορετικών πυρήνων φαίνεται πως δεν επηρεάζει ιδιαίτερα την ταξινόμηση.

5.5 Ιστογραφικά Χαρακτηριστικά

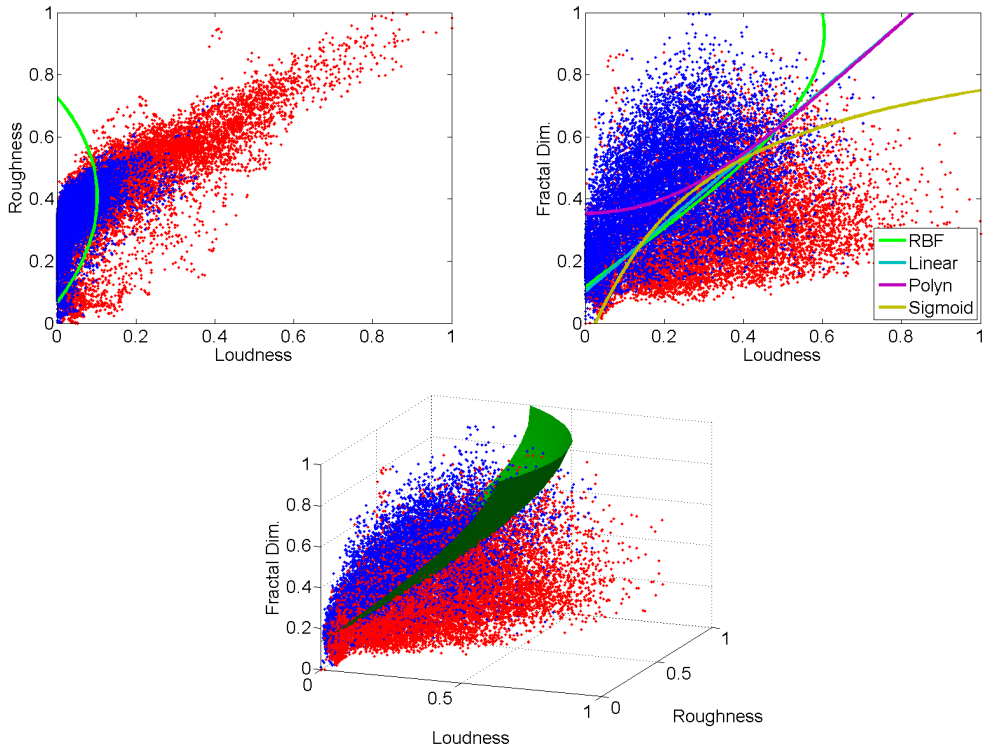
Με χρήση των χαρακτηριστικών που δοκιμάστηκαν στις προηγούμενες ενότητες, επιτεύχθηκε μία μέση απόδοση δημιουργώντας καμπύλες σημαντικότητας και λαμβάνοντας



Σχήμα 5.11: Απόδοση ταξινόμησης με χρήση SVM και RBF πυρήνα για διάφορες τιμές της παραμέτρου γ.

Πίνακας 5.6: Ταξινόμηση με SVM Gaussian πυρήνα και παράθυρο διάρκειας ίσο με 40 ms.

Feature	Acc(%)	Prec(%)	Rec(%)
En.+Loud.	71.7	74.6	75.0
En+Rough	65.4	70.0	70.9
En+FrD	59.9	72.8	64.1
Loud+Rough	73.6	77.3	74.2
Loud+FrD	72.8	77.6	72.8
Rough+FrD	62.5	69.1	67.1
E+R+F	64.6	70.6	69.6
L+R+F	73.4	78.1	72.6
All four	72.8	75.5	75.9



Σχήμα 5.12: Διαχωριστικές επιφάνειες με χρήση SVM Gaussian πυρήνα για διανύσματα χαρακτηριστικών σε δύο και τρεις διαστάσεις.

απόφαση μέσω κάποιου κατωφλίου, αλλά και διανυσμάτων από τα χαρακτηριστικά και εκπαιδεύοντας ένα SVM κάποιου πυρήνα. Οι διαφορετικοί συνδυασμοί χαρακτηριστικών είχαν μια μικρή συνήθως διαφορά απόδοσης μεταξύ τους, ενώ οι διαφορετικοί πυρήνες SVM είχαν την ίδια απόδοση. Σε κάθε περίπτωση η μέγιστη απόδοση που επιτεύχθηκε με βάση τα μέτρα που χρησιμοποιούνται δεν μεταβαλλόταν ιδιαίτερα. Σε αυτή την ενότητα αναπτύσσεται ένας διαφορετικός τρόπος χειρισμού των χαρακτηριστικών που εξάχθηκαν. Από τα χαρακτηριστικά υπολογίζονται ιστογράμματα, τα οποία χρησιμοποιούνται στη συνέχεια για ταξινόμηση με χρήση των SVM.

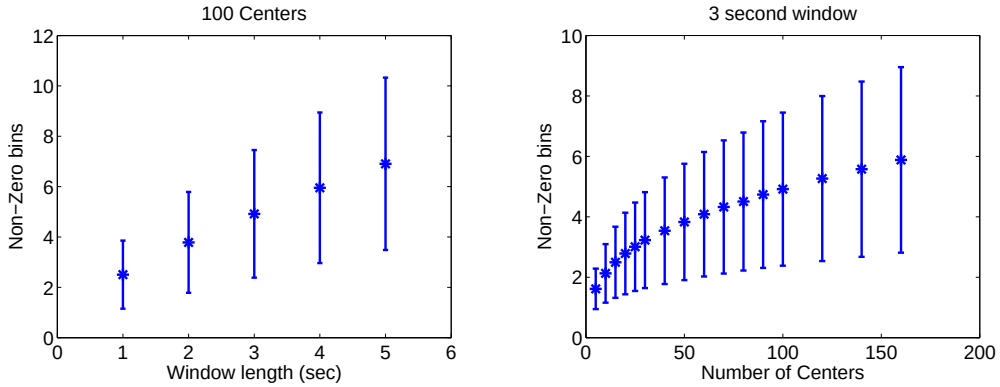
Ιστογραφικά χαρακτηριστικά έχουν χρησιμοποιηθεί ευρέως στην επεξεργασία εικόνων, για την καταγραφή της κατανομής ιδιοτήτων της εικόνας, όπως η ένταση (luminance), το χρώμα, αλλά και ανωτέρου επιπέδου χαρακτηριστικά που απαιτούν κάποια επεξεργασία αρχικά της εικόνας, όπως το σχήμα αντικειμένων. Στην επεξεργασία ηχητικών σημάτων έχουν χρησιμοποιηθεί επιτυχώς σε εφαρμογές κατηγοριοποίησης του είδους μουσικής (genre classification), στην αυτόματη εύρεση ηχητικών δεδομένων (audio retrieval), και σε αναγνώριση φωνής.

Σε αυτή την ενότητα, στις θέσεις (*bins*) των ιστογραμμάτων αντιστοιχίζονται διανύσματα. Τα διανύσματα αποτελούν το λεγόμενο λεξικό (*dictionary*) στο οποίο θα "προβληθούν" τα δεδομένα. Έχουν αναπτυχθεί διάφορες μέθοδοι για την εύρεση του λεξικού [67, 70, 65] ανάλογα με την εφαρμογή, οι οποίες οδηγούν σε διαφορετικές αναπαραστάσεις των δεδομένων και επιτυγχάνουν διαφορετικές επιδόσεις. Για την εύρεση του λεξικού θα χρησιμοποιηθεί ο αλγόριθμος k-Means, και οι λέξεις θα είναι τα κέντρα του.

Αρχικά τα δεδομένα χωρίζονται σε εκπαίδευση και επαλήθευση. Εκτελείται ο k-Means στα δεδομένα εκπαίδευσης για την εύρεση των κέντρων και κβαντισμού του χώρου (*vector-quantization*). Κάθε κέντρο αντιστοιχίζεται σε μία θέση του ιστογράμματος. Έπειτα θεωρείται χρονικό παράθυρο κάποιας διάρκειας και όλα τα διανύσματα δεδομένων εντός του παραθύρου. Για κάθε χρονικό παράθυρο δημιουργείται και ένα ιστόγραμμα. Το ύψος της μπάρας του ιστογράμματος σε κάθε θέση ισούται με το πλήθος των διανυσμάτων εντός του χρονικού παραθύρου των οποίων το πλησιέστερο κέντρο είναι αυτό που αντιστοιχεί στην θέση. Το ιστόγραμμα κανονικοποιείται ώστε το άθροισμα των τιμών του να ισούται με μονάδα. Δηλαδή, το ύψος της μπάρας σε κάθε θέση ισούται με τον λόγο των σημείων εντός του χρονικού παραθύρου που αντιστοιχίζονται στη θέση.

Ιστογράμματα υπολογίζονται με αυτή την διαδικασία για τα δεδομένα εκπαίδευσης και επαλήθευσης. Ωστόσο, μόνο τα δεδομένα εκπαίδευσης χρησιμοποιούνται για την εύρεση των κέντρων. Κάθε φορά που αλλάζουν τα δεδομένα εκπαίδευσης, αλλάζουν τα κέντρα και συνεπώς τα ιστογράμματα.

Τα διανύσματα χαρακτηριστικών έχουν συνιστώσες τα τέσσερα χαρακτηριστικά που περιγράφηκαν στο προηγούμενο Κεφάλαιο, δηλαδή ενεργεία, loudness, roughness, και fractal διάσταση. Η συχνότητα δειγματοληψίας των χαρακτηριστικών είναι 25 Hz, με συνέπεια διαδοχικά δείγματα να είναι υψηλά συσχετισμένα μεταξύ τους. Αυτό έχει ως επακόλουθο να έχουν ίδια πλησιέστερα κέντρα και τα ιστογράμματα που δημιουργούνται να είναι αρκετά αραιά (*sparse*), έχοντας σε μικρό ποσοστό των θέσεων τους μη-μηδενικά ύψη. Η αραιή αναπαράσταση που δημιουργείται ευνοεί την ταξινόμηση με SVM, τόσο ως προς την ορθότητα ταξινόμησης αλλά και στη μείωση υπολογιστικού κόστους. Στο



Σχήμα 5.13: Μέση τιμή του πλήθους των μη-μηδενικών θέσεων ως συνάρτηση: του μήκους το χρονικού παραθύρου (αριστερά), του πλήθους των θέσεων των ιστογραμμάτων (δεξιά). Οι κατακόρυφες γραμμές δείχνουν την τυπική απόκλιση.

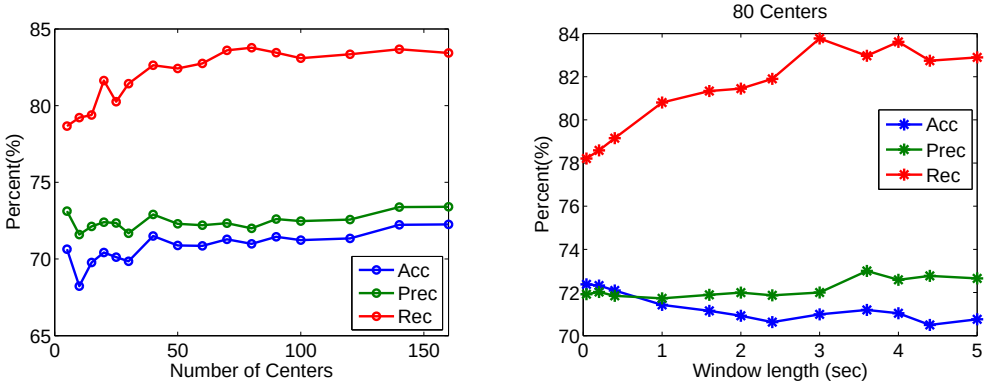
Σχήμα 5.13 φαίνεται η μέση τιμή του πλήθους των θέσεων που τα ιστογράμματα έχουν μη-μηδενικά ύψη συναρτήσει του μήκους του χρονικού παραθύρου και του αριθμού των κέντρων του k-means. Είναι εμφανές ότι μικρό πλήθος θέσεων των ιστογραφικών διανυσμάτων έχουν μη-μηδενική τιμή, που συνήθως είναι μικρότερο του 20% του μήκους του διανύσματος.

5.5.1 Μεταβολή του αριθμού των κέντρων

Μία παράμετρος της διαδικασίας είναι ο αριθμός των κέντρων για τον κβαντισμό του χώρου, που ταυτίζεται με το πλήθος των θέσεων των ιστογραμμάτων. Εάν M είναι ο αριθμός των κέντρων και N η διάσταση τους, συνήθως $M >> N$. Μικρός αριθμός κέντρων έχει ως συνέπεια μεγάλο σφάλμα κβάντισης, και μικρό αριθμό θέσεων στο ιστόγραμμα. Τα ιστογράμματα αποτυπώνουν μια πολύ γενική κατανομή των διανυσμάτων στον χώρο, καθώς με λίγα κέντρα ακόμη και διανύσματα που διαφέρουν σημαντικά πιθανώς να έχουν το ίδιο πλησιέστερο κέντρο και καταλήγουν στην ίδια θέση του ιστογράμματος. Αυτό θα κάνει δυσδιάκριτες τις δύο κλάσεις σημαντικότητας και οδηγεί σε μειωμένη απόδοση ταξινόμησης.

Αύξηση του αριθμού των κέντρων οδηγεί σταδιακά σε μείωση του σφάλματος κβάντισης και στην δημιουργία μικρότερων γειτονιών γύρω από τα κέντρα. Τα κέντρα γίνονται πιο αντιπροσωπευτικά των διανυσμάτων στην περιοχή, και τα ιστογράμματα αρχίζουν να απεικονίζουν την κατανομή τους στον χώρο. Ωστόσο, πολύ μεγάλος αριθμός κέντρων οδηγεί σε αύξηση του μεγέθους των ιστογραμμάτων και στην συμπερίληψη λεπτομερειών που δεν συμβάλλουν στην ταξινόμηση και διάκριση των κλάσεων. Τα ιστογράμματα επηρεάζονται έντονα από τις θέσεις των κέντρων και μικρή μετακίνηση τους οδηγεί σε αλλαγή τοπικά των τιμών τους.

Τα διανύσματα που χρησιμοποιούνται έχουν 4 διαστάσεις. Έγινε πειραματισμός με τον αριθμό των κέντρων να μεταβάλλεται από 5 έως 160. Στο Σχήμα 5.14α΄ φαίνεται η



(α΄) (β΄)

Σχήμα 5.14: Μεταβολή της απόδοσης γραμμικού SVM συναρτήσει του αριθμού των κέντρων για χρονικό παράθυρο διάρκειας 3 second (αριστερά), και του μήκους του χρονικού παραθύρου με χρήση 80 κέντρων (δεξιά).

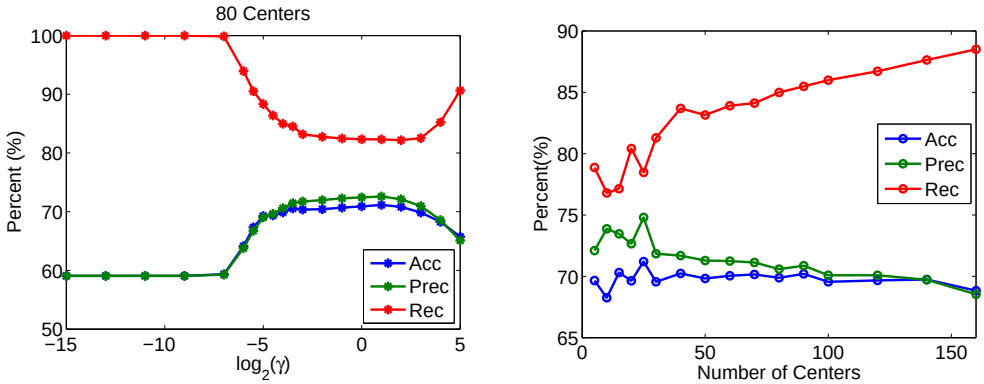
απόδοση συναρτήσει του αριθμού των κέντρων όταν η ταξινόμηση πραγματοποιείται με SVM γραμμικού πυρήνα για χρονικό παράθυρο διάρκειας 3 δευτερολέπτων. Αύξηση του αριθμού των κέντρων οδηγεί σε αύξηση αρχικά του accuracy, και έπειτα σταθεροποίηση του. Το recall αυξάνεται επίσης, και λαμβάνει μέγιστη τιμή ($\approx 84\%$) για αριθμό κέντρων ίσο με 80. Σε σύγκριση με την ταξινόμηση με χρήση των χαρακτηριστικών, χωρίς τη δημιουργία ιστογραμμάτων, επιτυγχάνεται αύξηση σχεδόν 10% στο recall, ενώ μια μείωση περίπου 3% στα accuracy και precision.

Για αριθμό κέντρων ίσο με 80, που επιτεύχθηκε μέγιστο recall με χρήση γραμμικού πυρήνα, δοκιμάστηκε ταξινόμηση με RBF πυρήνα για διάφορες τιμές της παραμέτρου γ, για να εξετασθεί εάν μπορεί να επιτευχθεί μεγαλύτερη απόδοση ταξινόμησης. Με επιλογή του γ μπορεί να υπάρξει περαιτέρω αύξηση του recall, χωρίς μεγάλη μείωση των accuracy και precision, όπως φαίνεται και στις καμπύλες του Σχήματος 5.15α΄. Για παράδειγμα, για $\log_2(\gamma) = -5$, επιτυγχάνεται αύξηση 4% στο recall το οποίο φθάνει στο 88%, ενώ η μείωση του accuracy είναι μόλις 1% σε σχέση με τον γραμμικό πυρήνα. Η ταξινόμηση με σταθερό $\gamma = 2^{-4}$ και μεταβολή του αριθμού των κέντρων είχε ως έξοδο καμπύλες του Σχήματος 5.15β΄. Αύξηση του αριθμού των κέντρων οδηγεί σε αύξηση του recall και μικρή μείωση της ακρίβειας ταξινόμησης.

5.5.2 Μεταβολή του μήκους του παραθύρου

Η άλλη παράμετρος η οποία εισέρχεται στην διαδικασία είναι το μήκος του παραθύρου στο οποίο θεωρούνται τα διανύσματα χαρακτηριστικών και υπολογίζεται κάθε ιστόγραμμα. Το παράθυρο αντιπροσωπεύει επίσης το χρονικό εύρος για το οποίο λαμβάνεται απόφαση για την ύπαρξη σημαντικών ηχητικών γεγονότων.

Στο Σχήμα 5.14β΄ φαίνεται η μεταβολή της απόδοσης για γραμμικό πυρήνα SVM συναρτήσει του μήκους του παραθύρου το οποίο κυμαίνεται από 40 ms (ένα διάνυσμα ανά



(α΄) (β΄)

Σχήμα 5.15: Μεταβολή της απόδοσης για RBF SVM συναρτήσει της παραμέτρου γ για αριθμό κέντρων ίσο με 80 (αριστερά), συναρτήσει του αριθμού του κέντρων για $\gamma = 2^{-4}$ και χρονικό παράθυρο διάρκειας 3 δευτερολέπτων (δεξιά).

παράθυρο), έως 5 sec (125 διανύσματα ανά παράθυρο). Αύξηση του μήκους του χρονικού παραθύρου οδηγεί σε αύξηση του recall και μικρή μείωση της ακρίβειας ταξινόμησης. Μέγιστο recall επιτυγχάνεται για χρονικό παράθυρο διάρκειας 3 sec.

Για παράθυρο διάρκειας 40 ms (περίοδος δειγματοληψίας χαρακτηριστικών), κάθε ιστόγραμμα δομείται από ένα μόνο διάνυσμα χαρακτηριστικών. Σε όλες τις θέσεις του έχει μηδενικά εκτός από μία στην οποία έχει τιμή μονάδα, που είναι η αυτή που αντιστοιχεί στο πλησιέστερο κέντρο του διανύσματος. Τα διανύσματα χαρακτηριστικών μετασχηματίζονται στα διανύσματα $(0, \ldots, 0, 1, 0, \ldots, 0)$, τα οποία είναι μήκους M, όπου M το μέγεθος του λεξικού, και κάθε ένα αντιπροσωπεύεται από το πλησιέστερο του κέντρο. Σε σύγκριση με την χρήση άμεσα των διανυσμάτων χαρακτηριστικών για ταξινόμηση, παρατηρείται μια αύξηση 3% στο recall, και αντίστοιχη μείωση στο accuracy.

Κεφάλαιο 6

Επεκτάσεις

Εισαγωγή

Σε αυτό το κεφάλαιο χρησιμοποιούνται τα χαρακτηριστικά MFCC και AM-FM της βιβλιογραφίας, που έχουν γνωρίσει επιτυχία στην επίλυση άλλων προβλημάτων, και εξετάζεται η απόδοση τους στο πρόβλημα ανίχνευσης σημαντικών ηχητικών γεγονότων. Δείχνεται ότι η απόδοση τους είναι εφάμιλλη αυτής που επιτεύχθηκε με τα χαρακτηριστικά των προηγούμενων κεφαλαίων. Επίσης, δοκιμάζεται μία υψηλότερου επιπέδου προσέγγιση και το ηχητικό σήμα χωρίζεται σε δύο σύνολα, όπου το ένα αποτελείται από τα σημεία στα οποία εμφανίζεται φωνή και το άλλο από αυτά στα οποία δεν εμφανίζεται. Γίνεται χρήση διαφορετικών χαρακτηριστικών σε κάθε σύνολο και ξεχωριστή ταξινόμηση. Με χρήση αυτή της προσέγγισης παρατηρείται βελτίωση των αποτελεσμάτων.

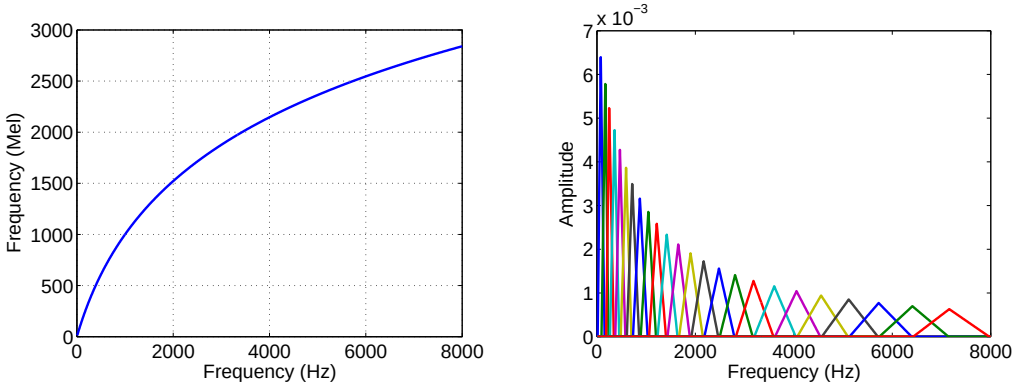
6.1 Mel Ενέργειες και MFCC

Τα χαρακτηριστικά MFCC (*Mel Frequency Cepstral Coefficients*) προκύπτουν από τον χώρο cepstrum του σήματος [14]. Έχουν χρησιμοποιηθεί επιτυχώς σε πολλές εφαρμογές μοντελοποίησης της χρονικής εξέλιξης των σημάτων όπως σε αναγνώριση φωνής και μουσικής, και αναγνώριση ομιλητή. Σε αυτήν την εργασία θα εξετασθεί η ικανότητα τους να διαχωρίσουν τις δύο κλάσεις σημαντικότητας.

Ο χώρος cepstrum ενός σήματος προκύπτει εφαρμόζοντας αντίστροφο μετασχηματισμό Fourier στο λογάριθμο του μέτρου του φάσματος του. Πιο συνοπτικά, για ένα διακριτό σήμα με μετασχηματισμό Fourier $X(e^{j\omega})$, το cepstrum δίνεται από την ακόλουθη σχέση:

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log(|X(e^{j\omega}|) e^{j\omega n} d\omega \tag{6.1}$$

Μια σημαντική ιδιότητα του χώρου του cepstrum είναι η μετατροπή της συνέλιξης στον χρόνο σε άθροιση στο cepstrum.



(α΄) (β΄)

Σχήμα 6.1: Αριστερά: καμπύλη που συνδέει γραμμική και mel κλίμακα συχνοτήτων. Δεξιά: Συστοιχία φίλτρων στο πεδίο της συχνότητας κεντραρισμένα με βάση την mel κλίμακα.

Για τον υπολογισμό των MFCC αρχικά πραγματοποιείται προεπεξεργασία του σήματος, που αποτελείται από αφαίρεση της μέσης τιμής, προέμφαση μέσω της ακόλουθης σχέσης:

$$x_{proemph} = x[n] - 0.97x[n-1] \tag{6.2}$$

και τέλος φιλτράρισμα με παράθυρο Hamming.

Στην συνέχεια υπολογίζεται το φάσμα του σήματος και χωρίζεται σε συχνοτικές μπάντες με βάση την κλίμακα *Mel*. Η κλίμακα Mel υπολογίζεται ώστε η αίσθηση της μεταβολής του pitch για δύο τόνους να είναι σταθερή εάν απέχουν το ίδιο σε αυτή την κλίμακα. Δηλαδή, η αίσθηση μεταβολής του pitch είναι σταθερή με την διαφορά της Mel συχνότητας. Η αντιστοίχιση της με την γραμμική κλίμακα συχνοτήτων γίνεται μέσω της ακόλουθης σχέσης:

$$m = 2595 \log(\frac{f}{700} + 1) \tag{6.3}$$

όπου f η συχνότητα σε Hz και m η συχνότητα σε Mel. Στο Σχήμα 6.1α΄ φαίνεται το γράφημα που συνδέει την γραμμική με την Mel κλίμακα συχνοτήτων. Ο χωρισμός του συχνοτικού άξονα σε μπάντες γίνεται ώστε οι κεντρικές συχνότητες των φίλτρων να απέχουν σταθερή απόσταση στην κλίμακα Mel, και οι συχνότητες αποκοπής κάθε μπάντας να είναι ίσες με τις κεντρικές συχνότητες των δύο μπάντων που βρίσκονται εκατέρωθεν της. Χρησιμοποιούνται φίλτρα τριγωνικού σχήματος, τα οποία φαίνονται στο 6.1β΄.

Ακολουθεί ο υπολογισμός της ενέργειας του σήματος σε κάθε μπάντα:

$$E(i) = \sum_k V_i[k]^2 |X[k]|^2 \tag{6.4}$$

όπου V_i η απόκριση συχνότητας του i φίλτρου, με $i = 0, 1, \ldots, Q-1$. Χρησιμοποιούμε $Q = 24$ φίλτρα, τα οποία καλύπτουν εύρος συχνοτήτων 8 kHz. Η έξοδος αυτού του

σταδίου είναι γνωστές στη βιβλιογραφία ως Mel-ενέργειες. Χρησιμοποιούνται επίσης σε εφαρμογές ανίχνευσης και αναγνώρισης φωνής παράλληλα με τα MFCC.

Στο επόμενο στάδιο λαμβάνεται ο λογάριθμος των Mel ενεργειών, G. Θα γίνει χρήση των συντελεστών G για ταξινόμηση σε κλάσεις σημαντικότητας. Για λόγους συντομίας θα αναφερόμαστε στη συνέχεια σε αυτούς με τον όρο Mel ενέργειες, παραλείποντας το λογάριθμο. Τέλος, υπολογίζεται ο διακριτός μετασχηματισμός συνημιτόνου (*Discrete Cosine Transform*) για αποσυσχέτιση τους:

$$C(n) = \frac{1}{Q}\sum_{i=0}^{Q-1} G(i)\cos\left(\frac{2\pi}{Q}(i-1/2)n\right),\ n = 0, 1, \ldots, N_c - 1,\ N_c \leq Q \qquad (6.5)$$

όπου το διάνυσμα C είναι τα MFCC χαρακτηριστικά του πλαισίου. Συνήθως διατηρείται ένα υποσύνολο των δυνατών συντελεστών που μπορούν να παραχθούν ($N_c < Q$) ώστε να αφαιρεθεί η εξάρτηση από ιδιότητες του ήχου από τον οποίο εξήχθησαν (π.χ. το pitch). Σε αυτήν την εργασία διατηρούνται $N_c = 14$ συντελεστές, συμπεριλαμβανομένου του συντελεστή ενέργειας (μηδενικού συντελεστή).

6.1.1 Υπολογισμός παραγώγων

Μεγάλη επιτυχία σε εφαρμογές αναγνώρισης φωνής, έχει η ενσωμάτωση στο διάνυσμα χαρακτηριστικών πρώτων και δεύτερων παραγώγων των Mel ενεργειών και των MFCC. Θα εξετασθεί εδώ εάν συνεισφέρουν στη διάκριση των κλάσεων σημαντικότητας. Οι παράγωγοι υπολογίζονται χρησιμοποιώντας ένα χρονικό παράθυρο διάρκειας N, ως εξής:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \qquad (6.6)$$

όπου c η παράμετρος της οποίας η παράγωγος πρέπει να υπολογισθεί. Για τον υπολογισμό δεύτερης παραγώγου, λαμβάνεται η παράγωγος της πρώτης παραγώγου με χρήση της ανωτέρας σχέσης.

6.2 Ταξινόμηση με Mel Ενέργειες και MFCC

Σε πρώτο στάδιο πραγματοποιείται ταξινόμηση με χρήση των Mel ενεργειών και των MFCC, χρησιμοποιώντας ως ταξινομητή τα SVM σε αναλογία με το προηγούμενο κεφάλαιο. Από τα δύο σύνολα χαρακτηριστικών έγινε επίσης εξαγωγή πρώτων και δεύτερων παραγώγων με τον τρόπο που περιγράφηκε στην προηγούμενη ενότητα, χρησιμοποιώντας παράθυρο $N_1 = 10$ και $N_2 = 5$, αντίστοιχα. Σε αυτό το σύνολο χαρακτηριστικών έγινε διαφορετική προ-επεξεργασία από τα χαρακτηριστικά του προηγούμενου κεφαλαίου. Οι συνιστώσες των διανυσμάτων κανονικοποιήθηκαν ώστε να έχουν μέση τιμή μηδέν και διακύμανση μονάδα. Ωστόσο η κλιμάκωση στο $[0, 1]$ έδωσε εφάμιλλα αποτελέσματα.

Πίνακας 6.1: Αποτελέσματα ταξινόμησης με χρήση log-mel ενεργειών και MFCC, με SVM γραμμικού πυρήνα. Ο δείκτης 0 στα MFCC δηλώνει την χρήση της μηδενικής συνιστώσας (συνιστώσα ενέργειας).

Feature	Acc(%)	Prec(%)	Rec(%)
LogMelen	73.8	76.2	76.1
LogMelen+D	74.1	76.6	76.6
LogMelen+D+A	74.3	76.7	76.9
MFCC	57.3	60.1	68.5
MFCC0	73.7	76.1	76.4
MFCC0+D	73.7	76.1	76.4
MFCC0+D+A	74.7	77.3	77.0

Στον Πίνακα 6.1 φαίνονται αποτελέσματα ταξινόμησης με χρήση γραμμικού πυρήνα SVM, Mel ενεργειών και MFCC. Τόσο οι Mel ενέργειες όσο και τα MFCC είχαν παρόμοια απόδοση με τα χαρακτηριστικά του προηγούμενου κεφαλαίου. Λαμβάνεται μέσο accuracy της τάξης του 74% και recall περίπου 77%. Η ενσωμάτωση παραγώγων έδωσε πολύ μικρή αύξηση των ποσοστών (< 1%) και για τις Mel ενέργειες και για τα MFCC. Επίσης, τα MFCC χωρίς την χρήση του μηδενικού συντελεστή (χωρίς τον δείκτη 0 στον Πίνακα), είχαν 16% μικρότερη ακρίβεια από ότι με την χρήση του, το οποίο αναδεικνύει για ακόμη μία φορά την ανάγκη ύπαρξης ενός χαρακτηριστικού το οποίο συσχετίζεται υψηλά με την ενέργεια του σήματος.

6.3 Χαρακτηριστικά Διαμόρφωσης

Τα χαρακτηριστικά διαμόρφωσης ενός σήματος βασίζονται στο μοντέλο των AM-FM διαμορφώσεων [39, 40], όπου το σήμα γράφεται ως ένας γραμμικός συνδυασμός από διαμορφωμένα κατά πλάτος και συχνότητα σήματα. Τα χαρακτηριστικά αυτά έχουν χρησιμοποιηθεί για την ανίχνευση σημαντικών ακουστικών και οπτικών γεγονότων και την αυτόματη δημιουργία περιλήψεων σε ταινίες [18]. Από τα AM-FM σήματα είναι δυνατός ο υπολογισμός της στιγμιαίας ενέργειας, του στιγμιαίου πλάτους και της στιγμιαίας συχνότητας τους με χρήση του τελεστή Teager-Kaiser.

Για συνεχή μονοδιάστατα σήματα ο ενεργειακός τελεστής Teager-Kaiser ορίζεται ως εξής [28]:

$$\Psi[x(t)] = (x'(t))^2 - x(t)x''(t) \tag{6.7}$$

Για διακριτά σήματα υπάρχουν διάφορες εκδοχές του τελεστή. Η εκδοχή που θα χρησιμοποιηθεί σε αυτή την εργασία είναι η ακόλουθη:

$$\Psi_d[x(n)] = x^n(n) - x(n-1)x(n+1) \tag{6.8}$$

Ο τελεστής Teager-Kaiser χαρακτηρίζεται ενεργειακός διότι εάν εφαρμοσθεί σε ένα σύστημα απλών αρμονικών ταλαντώσεων, υπολογίζει την ενέργεια του. Για ένα ημιτονικό

σήμα $x(t) = A\cos(\omega t + \phi)$, η έξοδος του τελεστή είναι $\Psi[x(t)] = \omega^2 A^2$, που ισούται με την ενέργεια του απλού αρμονικού ταλαντωτή που παράγει το σήμα.

Η εφαρμογή του τελεστή σε ένα AM-FM σήμα της μορφής:

$$x(t) = \alpha(t)\cos\left(\int_0^t \omega(\tau)d\tau\right) \tag{6.9}$$

όπου $\alpha(t),\ \omega(t)$ είναι τα χρονικά μεταβαλλόμενα πλάτος και συχνότητα αντίστοιχα, προσεγγίζει την μεταβαλλόμενη ενέργεια της πηγής, όπως έχει δειχθεί στο [39]:

$$\Psi[x(t)] \approx \omega^2(t)\alpha^2(t) \tag{6.10}$$

Λόγω της ιδιότητας του τελεστή Ψ να προσεγγίζει την ενέργεια για σήματα στενής ζώνης, καθίσταται δυνατή η αποδιαμόρφωση τους σε πλάτος και συχνότητα μέσω του αλγορίθμου AM-FM αποδιαμόρφωσης (*Energy Separation Algorithm-ESA*) ως εξής [39]:

$$\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \approx \omega(t), \quad \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \approx \alpha(t) \tag{6.11}$$

Για διακριτά AM-FM σήματα:

$$x(n) = A(n)\cos(\int_0^n \Omega(k)dk) \tag{6.12}$$

αναπτύχθηκε ο διακριτός ESA αλγόριθμος (*Discrete ESA - DESA*):

$$\arccos\left(1 - \frac{\Psi_d[x(n) - x(n-1)] + \Psi_d[x(n+1) - x(n)]}{4\Psi_d[x(n)]}\right) \approx \Omega(n)$$

$$\sqrt{\frac{\Psi_d[x(n)]}{\sin^2(\Omega(n))}} \approx A(n)$$

Για να εφαρμοσθεί ο αλγόριθμος αποδιαμόρφωσης σε ένα ευρυζωνικό (*wideband*) σήμα, απαιτείται το φιλτράρισμα του με συστοιχία ζωνοπερατών στενών φίλτρων, και η εφαρμογή του στην έξοδο κάθε φίλτρου. Χρησιμοποιούνται ζωνοπερατά μονοδιάστατα Gabor φίλτρα που η κρουστική τους απόκριση είναι της μορφής:

$$h(t) = e^{-\alpha^2 t^2}\cos(\omega_c t) \tag{6.13}$$

όπου ω_c η κεντρική συχνότητα του φίλτρου, και α το συχνοτικό τους εύρος.

Εάν K είναι ο αριθμός των Gabor φίλτρων που χρησιμοποιούνται, στην έξοδο του κάθε ενός υπολογίζεται η στιγμιαία ενέργεια μέσω της εφαρμογής του διακριτού τελεστή Teager-Kaiser, Ψ_d. Στο χρονικό παράθυρο m του σήματος x, έστω μήκους N σημείων, αντιστοιχίζεται η μέγιστη μέση στιγμιαία ενέργεια (*mean Multiband Teager Energy-MTE*) από την έξοδο των φίλτρων, ως εξής:

$$MTE(m) = \max_{1\le k\le K} \frac{1}{N}\sum_{n=1}^{N} \Psi_d[(h_k * x)(n)] \tag{6.14}$$

Τέλος, εάν j είναι το φίλτρο για το οποίο λαμβάνεται η μέγιστη μέση στιγμιαία ενέργεια, εφαρμόζεται ο αλγόριθμος αποδιαμόρφωσης στην έξοδο του και υπολογίζεται το αντίστοιχο μέσο στιγμιαίο πλάτος (*mean Multiband Instant Amplitude-MIA*), και η μέση στιγμιαία συχνότητα (*mean Multiband Instant Frequency-MIF*):

$$MIA(m) = \frac{1}{N}\sum_{n=1}^{N}|A_j(n)|, \quad MIF(m) = \frac{1}{N}\sum_{n=1}^{N}|\Omega_j(n)| \tag{6.15}$$

6.4 Ταξινόμηση με Χαρακτηριστικά Διαμόρφωσης

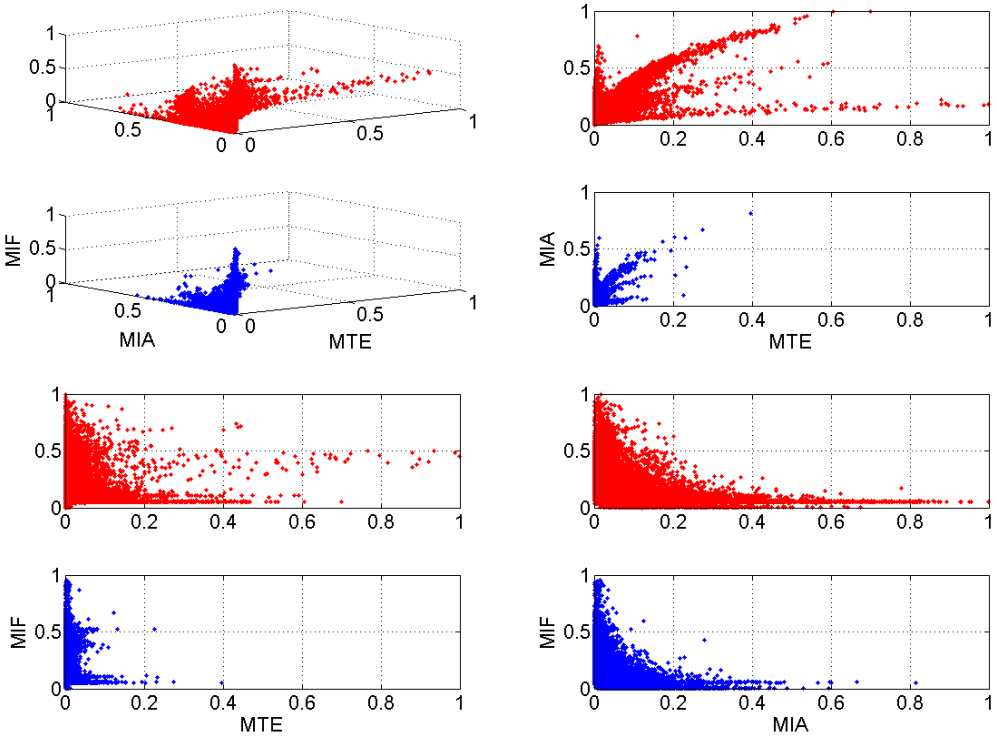
Πραγματοποιήθηκε εξαγωγή AM-FM χαρακτηριστικών από τα αποσπάσματα ταινιών της βάσης με χρήση της μεθόδου που περιγράφηκε στην προηγούμενη ενότητα. Για την εξαγωγή των χαρακτηριστικών χρησιμοποιούνται χρονικά παράθυρα διάρκειας 30 ms, με 20 ms επικάλυψη μεταξύ τους. Σε κάθε παράθυρο υπολογίζονται τρία μεγέθη: η ενέργεια Teager-Kaiser (MTE), το στιγμιαίο πλάτος (MIA), και η στιγμιαία συχνότητα (MIF). Τα χαρακτηριστικά για κάθε ταινία κλιμακώνονται ώστε να λαμβάνουν τιμές στο διάστημα $[0, 1]$. Επίσης, εφαρμόζεται φιλτράρισμα μέσου (median) για την ομαλοποίηση της μεταβολής τους συναρτήσει του χρόνου.

Η προσέγγιση που ακολουθείται σε αυτή την ενότητα περιλαμβάνει τη δημιουργία διανυσμάτων με συνιστώσες τα χαρακτηριστικά, και ταξινόμηση με χρήση μεθόδων μηχανικής μάθησης. Συγκεκριμένα, για την ταξινόμηση τους χρησιμοποιείται ο αλγόριθμος πλησιέστερων γειτόνων (*k Nearest Neighbor - kNN*). Στο Σχήμα 6.2 φαίνεται η κατανομή των χαρακτηριστικών στον χώρο για την ταινία Chicago για τις δύο κλάσεις σημαντικότητας. Η επικάλυψη μεταξύ των δύο κλάσεων είναι σημαντική. Κάποια διάκριση παρατηρείται στο μέγεθος της ενέργειας και στο πλάτος, με τα σημεία που έχουν σημειωθεί ως σημαντικά να λαμβάνουν μεγαλύτερες τιμές σε αυτά τα μεγέθη από τα μη-σημαντικά. Αντίθετα, οι δύο κλάσεις δεν διακρίνονται ως προς την στιγμιαία συχνότητα.

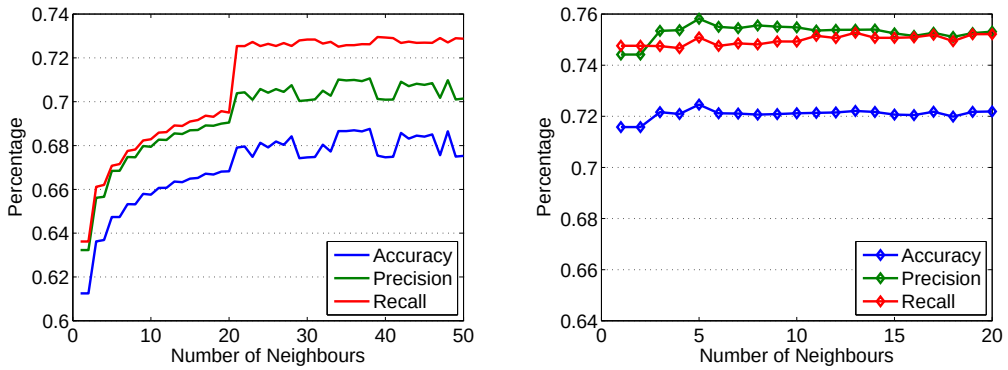
Αναπαριστώντας κάθε χρονικό παράθυρο με τα μεγέθη MTE, MIA, και MIF έγινε ταξινόμηση τους με χρήση του αλγόριθμου kNN για διάφορες τιμές του k. Στο Σχήμα 6.3 φαίνεται η μέση τιμή των accuracy, precision, recall για τις ταινίες της βάσης ως συνάρτηση του αριθμού των γειτόνων. Χωρίς την εφαρμογή median φιλτραρίσματος παρατηρείται μία αύξηση στην απόδοση με αύξηση του αριθμού γειτόνων, ενώ με την εφαρμογή του η απόδοση είναι σχεδόν σταθερή. Το median φιλτράρισμα ευνοεί την ταξινόμηση και την επίτευξη υψηλότερων ποσοστών. Στον Πίνακα 6.2 φαίνονται τα αποτελέσματα ταξινόμησης για κάθε ταινία της βάσης για την τιμή του k που επιτεύχθηκε το μέγιστο recall με την εφαρμογή median φιλτραρίσματος.

Για την ταινία Departed, που επιτεύχθηκε αρκετά μικρότερο accuracy και precision από τις υπόλοιπες ταινίες, η ενέργεια και το πλάτος των δύο κλάσεων σημαντικότητας δεν διέφεραν σημαντικά, όπως φαίνεται και στο Σχήμα 6.4.

Τέλος πραγματοποιήθηκε ταξινόμηση των σκηνών με χρήση ενός μόνο χαρακτηριστικού από τα MTE, MIA, MIF, και έπειτα με χρήση δύο για την δημιουργία του διανύσματος χαρακτηριστικών, με σκοπό να εξετασθεί η συμβολή κάθε συνιστώσας στην διάκριση των



Σχήμα 6.2: Κατανομή AM-FM χαρακτηριστικών στον χώρο, όπως υπολογίστηκαν από ηχητικά δεδομένα της βάσης.

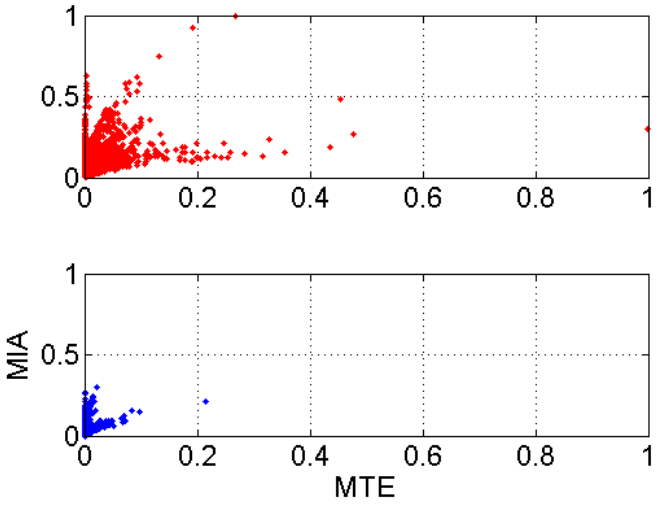


(α΄) (β΄)

Σχήμα 6.3: Απόδοση των χαρακτηριστικών AM-FM ταξινομώντας με αλγόριθμο kNN συναρτήσει του αριθμού γειτόνων k, 6.3α΄ χωρίς την εφαρμογή median φιλτραρίσματος, 6.3β΄ με median φιλτράρισμα.

Πίνακας 6.2: Αποτελέσματα ταξινόμησης με χρήση AM-FM χαρακτηριστικών και kNN αλγορίθμου για k=13.

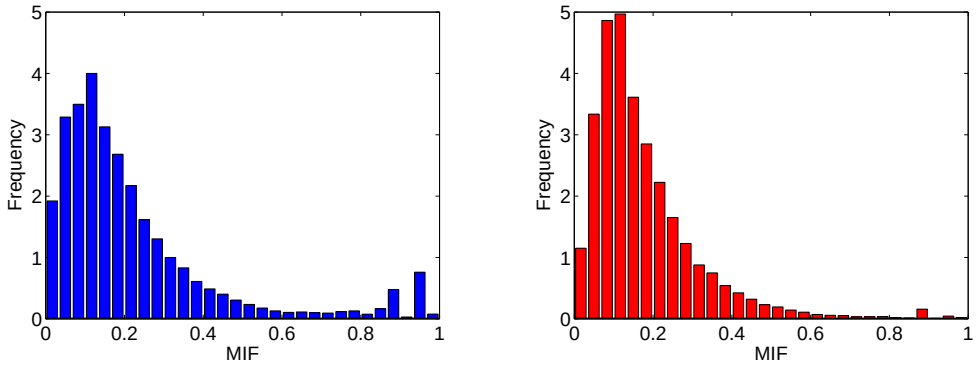
Movie	Accuracy(%)	Precision(%)	Recall(%)
CHI	68.1	65.5	94.3
CRA	74.0	76.4	77.9
DEP	71.0	51.1	76.8
FNE	71.8	70.6	81.1
GLA	78.0	85.4	76.4
LOR	70.9	98.1	51.3
Average	72.2	75.4	75.3



Σχήμα 6.4: Κατανομή MTE-MIA χαρακτηριστικών στο επίπεδο για την ταινία Departed.

Πίνακας 6.3: Μέσα ποσοστά επιτυχίας για όλες τις ταινίες της βάσης με διάφορους συνδυασμούς AM-FM χαρακτηριστικών.

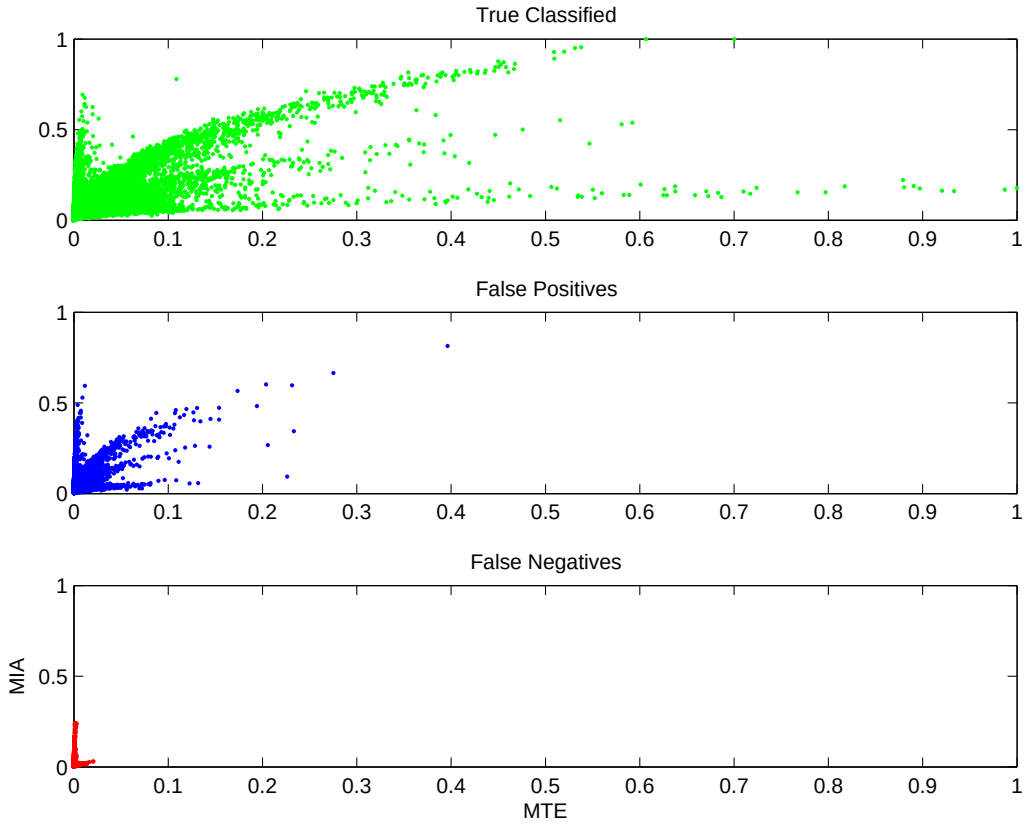
Features	Accuracy(%)	Precision(%)	Recall(%)
MTE	70.0	73.2	74.5
MIA	71.6	71.7	79.5
MIF	56.9	55.7	89.8
MTE+MIA	71.3	74.4	74.8
MTE+MIF	72.8	78.0	70.9
MIA+MIF	72.4	75.0	76.1
MTE+MIA+MIF	72.2	75.4	75.3



Σχήμα 6.5: Ιστόγραμμα τιμών για όλες τις ταινίες της βάσης MovieSum του χαρακτηριστικού MIF για τα μη-σημαντικά και σημαντικά γεγονότα, αριστερά και δεξιά αντίστοιχα.

δύο κλάσεων. Παρατηρήθηκε, ότι τα MTE, MIA είναι ικανά από μόνα τους να διακρίνουν τις δύο κλάσεις ενώ το MIF δεν είναι. Τα σημεία στην MIF διάσταση κατανέμονται με παρόμοιο τρόπο όπως φαίνεται και στο Σχήμα 6.5. Ωστόσο, συνδυάζοντας το MIF με κάποιο από τα άλλα δύο χαρακτηριστικά η απόδοση βελτιώνετε. Τα αποτελέσματα ταξινόμησης για κάθε συνδυασμό φαίνονται στον Πίνακα 6.3.

Παρατηρείται, επίσης, συμφωνία στην κλάση στην οποία ταξινομείται κάθε σκηνή με χρήση των διαφορετικών διανυσμάτων. Εξαιρείται η περίπτωση που το διάνυσμα χαρακτηριστικών αποτελείται μόνο από το MIF. Ενδεικτικά τα MTE - MIA ταξινομούν στην ίδια κλάση το 85.3% των σκηνών, τα MIA - MIF το 60.8%, τα MTE - ALL το 89.8%, τα MTE+MIF - ALL το 88.8%, και τα MIA+MIF - ALL το 97.1%, όπου ALL = MTE+MIA+MIF.



Σχήμα 6.6: Χαρακτηριστικά MTE, MIA για την ταινία Chicago χωρισμένα με βάση την ορθότητα ταξινόμησης από τον kNN αλγόριθμο.

6.5 Διαχωρισμός φωνής-μη φωνής

Τα χαρακτηριστικά που έχουν δοκιμαστεί, είτε αναφέρονται σε αυτή την εργασία είτε όχι, για τον διαχωρισμό των δύο κλάσεων σημαντικότητας της βάσης δεδομένων που χρησιμοποιείται, δεν έχουν αποδώσει ιδιαίτερα υψηλά ποσοστά αναγνώρισης. Όλες οι κατηγορίες χαρακτηριστικών που έχουν δοκιμαστεί έχουν απόδοση περίπου στο 75% για όλα τα μέτρα που χρησιμοποιούνται, ενώ μόνο η μέθοδος των ιστογραμμάτων απέδωσε συστηματικά υψηλότερο recall. Σε αυτή την ενότητα δοκιμάζεται μια υψηλότερου επιπέδου προσέγγιση και χωρίζεται το ηχητικό σήμα σε δύο σύνολα, που το ένα αποτελείται από τα σημεία στα οποία υπάρχει ανθρώπινη φωνή ενώ το άλλο από τα σημεία στα οποία δεν υπάρχει. Έπειτα πραγματοποιείται επεξεργασία και ταξινόμηση ξεχωριστά σε κάθε σύνολο.

Ο διαχωρισμός των σημείων που υπάρχει φωνή από αυτά στα οποία δεν υπάρχει, γίνεται με βάση τους υπότιτλους που παρέχονται μαζί με την κάθε ταινία. Τα σημεία στα οποία υπάρχουν υπότιτλοι θεωρούνται σημεία φωνής, ενώ τα υπόλοιπα θεωρούνται σημεία μη-φωνής. Να σημειωθεί ότι τα όρια τα οποία λαμβάνονται από τους υπότιτλους δεν είναι πολύ ακριβή, καθώς αρκετές φορές οι υπότιτλοι εμφανίζονται αφού έχει αρχίσει η ομιλία, και εξαφανίζονται μετά το τέλος της. Αυτή η διαφορά χρονισμού είναι συνήθως της τάξης των εκατοντάδων millisecond, το οποίο σημαίνει ότι επηρεάζει μερικά πλαίσια των χαρακτηριστικών που εξάγονται. Ωστόσο, η χρήση των υποτίτλων είναι ένας εύκολος και γρήγορος τρόπος να γίνει μία αρχική ταξινόμηση με ικανοποιητική ακρίβεια. Επίσης, παρατηρείται το φαινόμενο της αδράνειας στην ανθρώπινη προσοχή, και ενώ η ηχητική διέγερση (όπως ομιλία) έχει παύσει, οι χρήστες έχουν σημειώσει ως σημαντική κάποια επιπλέον διάρκεια.

Στην περίπτωση ηχητικών δεδομένων που δεν υπάρχουν υπότιτλοι, διαχωρισμός μπορεί να γίνει με κάποιον αλγόριθμο ανίχνευσης φωνής (*voice activity detection-VAD*) ο οποίος μπορεί να δώσει αρκετά υψηλά ποσοστά ανίχνευσης ανάλογα με την περίπτωση των δεδομένων. Επίσης, οι VAD αλγόριθμοι σε συνδυασμό με τους υπότιτλους, μπορούν να χρησιμοποιηθούν για την εύρεση ακριβέστερων ορίων φωνής στις ταινίες.

Στον Πίνακα 6.4 φαίνονται κάποια στατιστικά στοιχεία για τους υπότιτλους κάθε ταινίας αλλά και συνολικά. Η συχνότητα εμφάνισης υποτίτλων ποικίλει μεταξύ των ταινιών από 34% έως 78%, ενώ συνολικά στο 54% της διάρκειας των ταινιών εμφανίζονται υπότιτλοι. Επίσης, τα σημεία στα οποία εμφανίζονται υπότιτλοι έχουν σημειωθεί σε μικρότερο ποσοστό σημαντικά (SubSal), από τα σημεία χωρίς υποτίτλους (notSubSal).

Για την ταξινόμηση των σημείων σε κλάσεις σημαντικότητας θα χρησιμοποιηθούν τα χαρακτηριστικά που περιγράφηκαν σε προηγούμενες ενότητες. Για την ταξινόμηση όλου του μήκους των ταινιών, πραγματοποιείται ταξινόμηση σε κάθε σύνολο ξεχωριστά και υπολογίζεται η απόδοση για το σύνολο, και στη συνέχεια υπολογίζεται η συνολική απόδοση συνδυάζοντας γραμμικά τις επιμέρους αποδόσεις με βάρη ανάλογα του μεγέθους των συνόλων (Πίνακας 6.4). Εξετάζεται για τα δύο σύνολα δεδομένων (φωνής - μη φωνής) ποια χαρακτηριστικά συμβάλλουν στη διάκριση των δύο κλάσεων. Λόγω της ασυμμετρίας που υπάρχει στη διάρκεια των υποτίτλων κάθε ταινίας και μεταξύ των ταινιών, για τον διαχωρισμό σε δεδομένα εκπαίδευσης-επαλήθευσης (όπου απαιτείται), θεωρεί-

Πίνακας 6.4: Στατιστικά υποτίτλων για τις ταινίες της βάσης.

Movie	Dur(mins)	Subt(mins)	Percent(%)	SubSal(%)	notSubSal(%)
CHI	30.14	23.55	78.15	53.03	74.68
CRA	26.62	12.01	45.12	41.59	68.33
DEP	30.47	19.79	64.93	24.07	40.93
FNE	30.29	19.69	64.99	45.70	68.03
GLA	30.05	10.35	34.45	37.88	71.58
LOR	37.56	14.22	37.86	44.22	67.38
Total	185.13	99.61	53.80	41.17	65.14

ται το ηχητικό σήμα από όλες τις ταινίες για κάθε σύνολο και πραγματοποιείται 5-fold cross-validation σε αυτό (ξεχωριστά για κάθε σύνολο).

6.5.1 Χρήση χαρακτηριστικών χρονικού χάρτη σημαντικότητας

Αρχικά πραγματοποιήθηκε ταξινόμηση των σημείων κάθε συνόλου με χρήση των χαρακτηριστικών του Κεφαλαίου 4 (loudness, roughness, κ.τ.λ.) και κατωφλίωσης των καμπυλών σημαντικότητας. Στον Πίνακα 6.5 φαίνονται τα αποτελέσματα από την ταξινόμηση στα σύνολα φωνής/μη-φωνής. Για τα δύο σύνολα επιτυγχάνεται παρόμοια ακρίβεια, αλλά τα precision και recall στο σύνολο της μη-φωνής είναι τουλάχιστον 6% υψηλότερα από το σύνολο φωνής για τους περισσότερους συνδυασμούς χαρακτηριστικών. Και στα δύο σύνολα δεδομένων, βέλτιστη απόδοση έχει η καμπύλη που δημιουργείται από ενέργεια και loudness. Ιδιαίτερα για το σύνολο μη-φωνής αυτός ο συνδυασμός χαρακτηριστικών έχει αρκετά υψηλότερη απόδοση από κάθε άλλο, με το accuracy να είναι στο 81% και το recall στο 88%, ενώ περαιτέρω μείωση του κατωφλίου δεν οδηγούσε σε ιδιαίτερη πτώση του accuracy, με το recall να αυξάνεται. Φαίνεται πως η ενέργεια είναι χαρακτηριστικό στο οποίο βασίζονται σε σημαντικό βαθμό οι χρήστες για την επισημείωση των τμημάτων μη-φωνής. Αντίθετα στα σημεία φωνής είναι λιγότερο σημαντική.

Συνδυάζοντας τις καλύτερες αποδόσεις σε κάθε σύνολο λαμβάνεται accuracy 78% και recall 81%, τα οποία είναι υψηλότερα 2 και 3% αντίστοιχα, από τα βέλτιστα που επιτεύχθηκαν χωρίς τον διαχωρισμό σε σύνολα φωνής και μη-φωνής με χρήση κατωφλίου.

6.5.2 Χρήση MFCC για το σύνολο φωνής

Λόγω της μεγάλης επιτυχίας που έχουν τα MFCC χαρακτηριστικά στην μοντελοποίηση του τρόπου εξέλιξης της ανθρώπινης ομιλίας, θα χρησιμοποιηθούν για την ανίχνευση σημαντικών γεγονότων στο σύνολο φωνής.

Αρχικά δοκιμάζονται τα MFCC στο σύνολο της φωνής. Στο διάνυσμα χαρακτηριστικών ενσωματώνονται και πρώτες και δεύτερες παράγωγοι, υπολογισμένες με χρήση πα-

Πίνακας 6.5: Κατωφλίωση καμπυλών χαρακτηριστικών έπειτα από διαχωρισμό σε σημεία φωνής / μη-φωνής.

Feature	Acc(%)	Prec(%)	Rec(%)
En.+Loud.	**75.6** / **81.3**	**69.2** / **84.2**	**74.7** / **88.2**
En+Rough	72.9 / 66.3	66.7 / 72.7	69.6 / 78.2
En+FrD	62.8 / 64.3	54.5 / 72.8	63.6 / 73.1
Loud+Rough	74.4 / 72.4	68.0 / 78.6	72.5 / 79.9
Loud+FrD	70.4 / 71.5	62.6 / 77.7	71.0 / 77.9
Rough+FrD	66.4 / 64.5	58.1 / 74.5	69.0 / 70.3
E+R+F	70.4 / 65.2	64.0 / 72.3	65.9 / 76.6
L+R+F	72.7 / 69.6	66.3 / 77.1	70.2 / 76.6
All four	72.4 / 70.0	64.9 / 77.2	73.0 / 77.4
Best	**78.3**	**76.3**	**81.0**

Πίνακας 6.6: Αποτελέσματα ταξινόμησης με χρήση log-Mel ενεργειών και MFCC, με SVM γραμμικού πυρήνα στο σύνολο φωνής. Ο δείκτης 0 στα MFCC δηλώνει την χρήση της μηδενικής συνιστώσας (συνιστώσα ενέργειας).

Feature	Acc(%)	Prec(%)	Rec(%)
Melen+D+A	75.6	69.3	75.2
MFCC+D+A	61.9	56.6	46.8
MFCC0+D+A	73.9	69.1	69.0

ραθύρων όπως προηγουμένως. Η ταξινόμηση γίνεται και πάλι χρησιμοποιώντας γραμμικό SVM. Όπως φαίνεται και στον Πίνακα 6.6 η απόδοση δεν βελτιώθηκε περιορίζοντας τα δεδομένα στο σύνολο φωνής. Επίσης με χρήση του μηδενικού συντελεστή για τα MFCC λαμβάνεται και πάλι αρκετά μεγαλύτερη ακρίβεια ταξινόμησης από ότι χωρίς αυτόν.

Κεφάλαιο 7

Επίλογος

7.1 Σύνοψη του προβλήματος και συμβολή της εργασίας

Η παρούσα εργασία αποτελεί μια προσπάθεια να προσεγγισθεί η ανθρώπινη ακουστική αντίληψη. Ειδικότερα επιχειρήθηκε η μοντελοποίηση και προσέγγιση της ακουστικής σημαντικότητας η οποία αντιστοιχεί στον κάτωθεν μηχανισμό της προσοχής. Τα σημαντικά σημεία ενός ηχητικού σήματος είναι αυτά τα οποία έλκουν αυθόρμητα την ανθρώπινη προσοχή ανεξαρτήτως σημασιολογικού περιεχόμενου. Οι χρήστες δεν καταβάλλουν προσπάθεια για να στρέψουν την προσοχή τους προς ένα σημαντικό ηχητικό γεγονός και δεν μπορούν να το αποφύγουν. Είναι οι ιδιότητες του ήχου που τον κάνουν σημαντικό στο περιβάλλον που εμφανίζεται, και η χαμηλού επιπέδου προσέγγιση που ακολουθείται σε αυτή την εργασία είχε στόχο την διερεύνηση αυτών των ιδιοτήτων.

Προς αυτή την κατεύθυνση εξετάσθηκε η ανίχνευση σημαντικών γεγονότων σε ηχητικά ερεθίσματα από ταινίες. Η βάση δεδομένων στην οποία πραγματοποιήθηκαν τα πειράματα αποτελούνταν από έξι αποσπάσματα ταινιών του Hollywood συνολικής διάρκειας 180 λεπτών (3 ωρών). Ως δεδομένα αλήθειας χρησιμοποιήθηκαν επισημειώσεις χρηστών στη βάση, που περιείχαν για κάθε χρονική στιγμή των ερεθισμάτων εάν η ηχητική σκηνή τους τράβηξε την προσοχή ή όχι. Ο έλεγχος των μοντέλων γίνεται με βάση αυτές τις επισημειώσεις.

Για την μοντελοποίηση της ακουστικής σημαντικότητας πραγματοποιήθηκε εξαγωγή χαρακτηριστικών χαμηλού επιπέδου. Αρχικά, έγινε εξαγωγή χαρακτηριστικών από το φασματογράφημα του σήματος. Χειριζόμενοι το φασματογράφημα ως εικόνα πραγματοποιήσαμε φιλτράρισμα του με τρία Gabor φίλτρα για την ενίσχυση σημείων με υψηλή ενέργεια, σημείων όπου παρατηρείται μεταβολή κατά μήκος του χρονικού άξονα, και σημείων που παρατηρείται μεταβολή κατά μήκος του συχνοτικού άξονα. Ο συνδυασμός αυτών των τριών χαρτών, έπειτα από κάποια ενδιάμεσα στάδια, οδηγούσε σε τελικό χάρτη και έπειτα καμπύλη σημαντικότητας για ανίχνευση προεξεχόντων γεγονότων. Από την ταξινόμηση που πραγματοποιήθηκε διαπιστώθηκε ότι κάθε ένας από τους χάρτες ήταν σε θέση να ανιχνεύσει σημαντικά γεγονότα. Ο συνδυασμός των χαρτών δεν βελτίωσε την επίδοση

ταξινόμησης. Αυτό έρχεται σε αντίθεση σε κάποιο βαθμό με τα αποτελέσματα στο [31] όπου ισχυρίζονται ότι μόνο ο χάρτης ενέργειας δεν αρκεί για την ανίχνευση της σημαντικότητας. Ωστόσο το πλαίσιο ταξινόμησης των συγγραφέων ήταν διαφορετικό.

Σε συνδυασμό με τους χάρτες σημαντικότητας, χρησιμοποιήθηκε η έννοια του *gist* μιας σκηνής από την βιβλιογραφία ανάλυσης εικόνων, και έγινε εξαγωγή διανυσμάτων από αυτούς και αναπαράσταση κάθε ηχητικής σκηνής. Η ταξινόμηση των διανυσμάτων έγινε με μεθόδους μηχανικής μάθησης. Η χρήση αυτής της αναπαράστασης έδωσε υψηλότερα ποσοστά ταξινόμησης σε σύγκριση με την κατωφλίωση της καμπύλης σημαντικότητας.

Εξετάσθηκε επίσης, η αντικατάσταση του φασματογραφήματος με χαρακτηριστικά σε μία διάσταση και η προσαρμογή όλων των σταδίων ώστε να χειρίζονται μονοδιάστατες καμπύλες. Έγινε εξαγωγή βραχέως χρόνου χαρακτηριστικών, στο πεδίο του χρόνου και της συχνότητας, τα οποία αποτέλεσαν και την είσοδο του μοντέλου. Από τα αποτελέσματα ταξινόμησης φάνηκε ότι για την επιτυχή ανίχνευση σημαντικών γεγονότων είναι απαραίτητη η ύπαρξη ενός χαρακτηριστικού που να συσχετίζεται υψηλά με την ενέργεια του σήματος. Επομένως, η ενέργεια βραχέως χρόνου ή το loudness, είναι απαραίτητο να υπάρχουν μεταξύ των χαρακτηριστικών. Ο συνδυασμός loudness και fractal διάστασης είχε τη βέλτιστη απόδοση σε αρκετές περιπτώσεις, που δείχνει ότι η προσθήκη ενός χαρακτηριστικού που εξαρτάται από την συχνότητα του σήματος βοηθά στην ταξινόμηση. Ακόμη, η αντικατάσταση του φασματογραφήματος με τα μονοδιάστατα χαρακτηριστικά οδήγησε σε αύξηση της απόδοσης.

Στην συνέχεια δοκιμάστηκε η δημιουργία ιστογραμμάτων από τα χαρακτηριστικά τα οποία απεικονίζουν πως αυτά κατανέμονται στον χρόνο, σε αναλογία με την μέθοδο bag-of-words στην όραση υπολογιστών, κάτι που δεν έχει επιχειρηθεί ξανά στην βιβλιογραφία ανίχνευσης σημαντικών ηχητικών γεγονότων. Με αυτή την μέθοδο υπήρξε μια επιπλέον αύξηση στην απόδοση με την ανίχνευση μεγαλύτερου μέρους από τα σημαντικά γεγονότα.

Δοκιμάστηκε επίσης η χρήση κάποιων καθιερωμένων χαρακτηριστικών στην βιβλιογραφία, όπως τα MFCC και τα AM-FM χαρακτηριστικά. Διαπιστώθηκε ότι τα MFCC για να έχουν υψηλή απόδοση απαιτείται η ενσωμάτωση στο διάνυσμα χαρακτηριστικών του συντελεστή ενέργειας. Για τα AM-FM φαίνεται πως μόνο οι συνιστώσες του στιγμιαίου πλάτους και της ενέργειας συνεισφέρουν στην διάκριση των δύο κλάσεων σημαντικότητας. Η συνιστώσα της στιγμιαίας συχνότητας δεν συμβάλλει ιδιαίτερα από μόνη της αλλά μόνο σε συνδυασμό με τις άλλες δύο συνιστώσες.

Τέλος, ακολουθήθηκε μια υψηλότερου επιπέδου προσέγγιση και έγινε διαχωρισμός του ηχητικού σήματος σε σύνολα φωνής και μη-φωνής. Διαπιστώθηκε ότι για το σύνολο φωνής ο συνδυασμός των χαρακτηριστικών της ενέργειας και του loudness προσεγγίζουν αρκετά καλά τις ανθρώπινες θεωρήσεις σημαντικότητας, και επιτεύχθηκε μια αύξηση της τάξης του 10% σε σύγκριση με τις υπόλοιπες αποδόσεις. Η χρήση των MFCC για το σύνολο φωνής δεν έφερε κάποια βελτίωση στην απόδοση ταξινόμησης. Συνολικά για όλη τη διάρκεια των ηχητικών δεδομένων, παρατηρήθηκε αύξηση της απόδοσης περίπου 3% για τα μέτρα accuracy και recall. Η προσέγγιση διαχωρισμού στα δύο σύνολα δεν έχει εφαρμοσθεί ξανά στην βιβλιογραφία, αποτελεί μια καινοτομία αυτής της εργασίας (αν και σε αρκετά πρώιμο στάδιο), και έχει προοπτικές για περαιτέρω εξέλιξη.

Συνοπτικά η εργασία εξέτασε τα ακόλουθα θέματα:

- Σύντομη εισαγωγή της έννοιας της προσοχής του ανθρώπου, παρουσίαση υπαρχόντων υπολογιστικών μοντέλων και μεθόδων αξιολόγησης τους.
- Αναλυτική περιγραφή του μοντέλου των Kayser et al [31] και του τρόπου αξιολόγησης του. Περιγραφή του μοντέλου ακουστικού φάσματος των Shamma et al [69].
- Συνδυασμό και χρήση των δύο προηγούμενων μοντέλων για την ανίχνευση σημαντικών σημείων σε ηχητικά σήματα από ταινίες. Εισαγωγή της έννοιας του *gist* μίας σκηνής, και εξαγωγής διανυσμάτων από τους χάρτες σημαντικότητας για ταξινόμηση με αλγορίθμους μηχανικής μάθησης.
- Τροποποίηση κάθε σταδίου του μοντέλου των Kayser et al ώστε να χειρίζεται μονοδιάστατα χαρακτηριστικά αντί για διδιάστατους χάρτες. Εξαγωγή μονοδιάστατων χαρακτηριστικών βραχέως χρόνου για να δοθούν ως είσοδος στο τροποποιημένο μοντέλο.
- Δημιουργία ιστογραμμάτων από τα χαρακτηριστικά τα οποία χειριζόμενα ως διανύσματα ταξινομήθηκαν με αλγορίθμους μηχανικής μάθησης.
- Υπολογισμό MFCC και AM-FM χαρακτηριστικών και σύγκριση της απόδοσης τους με τα υπόλοιπα χαρακτηριστικά.
- Διαχωρισμό των ηχητικών δεδομένων σε δύο σύνολα όπου στο ένα εμφανίζεται ανθρώπινη φωνή ενώ στο άλλο δεν εμφανίζεται, και χρήση διαφορετικών χαρακτηριστικών σε κάθε ένα για ταξινόμηση των σημείων του.

7.2 Μελλοντικές Κατευθύνσεις

Το πεδίο της ανίχνευσης σημαντικών ή προεξεχόντων γεγονότων βρίσκεται ακόμη σε πρώιμο στάδιο και οι υπολογιστικές μέθοδοι απαιτείται να εξελιχθούν ώστε να προσεγγίσουν περισσότερο τις πειραματικές μετρήσεις. Όπως τονίσθηκε και στο Κεφάλαιο 2 ένας βασικός λόγος (και ίσως ο κύριος) που καθυστερεί η ανάπτυξη τόσο των υπολογιστικών μοντέλων όσο και της μάθησης από βιολογική σκοπιά των μηχανισμών της ακουστικής προσοχής, είναι η έλλειψη ενός πιο "αντικειμενικού" τρόπου αξιολόγησης τους και συγκέντρωσης πειραματικών μετρήσεων. Τα ερεθίσματα των ταινιών, και άλλες μέθοδοι που έχουν εφαρμοσθεί είναι χρήσιμα για την εξαγωγή κάποιων συμπερασμάτων αλλά εμφανίζουν ένα μικρό μόνο μέρος της αλήθειας. Μία λύση στο πρόβλημα θα ήταν η χρήση τεχνικών απεικόνισης της δραστηριότητας του ανθρώπινου εγκεφάλου, όπως fMRI, ενώ οι χρήστες εκτίθενται στα ηχητικά ερεθίσματα. Συσχετίζοντας την δραστηριότητα στον εγκέφαλο με την εμφάνιση ερεθισμάτων θα μπορούσε να εξαχθούν συμπεράσματα για την σημαντικότητα τους. Η επιλογή αυτή, θα παρείχε πιο αξιόπιστα δεδομένα. Έχει ωστόσο το μειονέκτημα, ότι δεν μπορεί να χρησιμοποιηθεί ευρέως από τις ερευνητικές ομάδες λόγω του όγκου και του χρηματικού κόστους του εξοπλισμού που απαιτεί.

Το φαινόμενο της κατεύθυνσης της ακουστικής προσοχής είναι δυναμικό και μεταβάλλεται συνεχώς. Επίσης, το που θα στραφεί η προσοχή κάποια χρονική στιγμή εξαρτάται από το που ήταν στραμμένη προηγούμενες χρονικές στιγμές. Τα χαρακτηριστικά που εξήχθησαν σε αυτή την εργασία ήταν όλα στατικά χωρίς να υπάρχει επιρροή από το τι προηγήθηκε στο σήμα. Μία μελλοντική κατεύθυνση για έρευνα θα ήταν ο υπολογισμός χαρακτηριστικών που να επηρεάζονται από τι ερεθίσματα έχουν προηγηθεί χρονικά. Ισοδύναμα, θα ήταν χρήσιμο το κατώφλι για την ύπαρξη σημαντικού γεγονότος να μεταβάλλεται δυναμικά και να μην μένει σταθερό για όλη την διάρκεια του ηχητικού ερεθίσματος. Για παράδειγμα μια μικρή απόκλιση σε ένα σιωπηλό περιβάλλον θα τραβήξει την προσοχή, ενώ η ίδια απόκλιση σε θορυβώδες περιβάλλον δεν θα γίνει αντιληπτή. Στην πρώτη περίπτωση απαιτείται μικρό κατώφλι για την ανίχνευση της μεταβολής, ενώ στη δεύτερη απαιτείται υψηλό για την απόρριψή της. Με το δυναμικό κατώφλι ή τα δυναμικά χαρακτηριστικά η διάκριση των δύο περιπτώσεων είναι εφικτή.

Τέλος, το πεδίο της έρευνας είναι ανοικτό για την ανάπτυξη άλλων χαρακτηριστικών και αναπαραστάσεων του ηχητικού σήματος τα οποία θα είναι πιο βιολογικά εμπνευσμένα και θα προσεγγίζουν σε μεγαλύτερο βαθμό τις λειτουργίες που πραγματοποιούνται στην άνθρωπο. Ακόμη η ανάπτυξη νέων μεθόδων συνδυασμού των χαρακτηριστικών και η διερεύνηση πως αυτά αλληλεπιδρούν μεταξύ τους θα οδηγήσει στην βαθύτερη κατανόηση του προβλήματος και περαιτέρω ανάπτυξη των μεθόδων για τη λύση του.

Βιβλιογραφία

[1] "Cognimuse project," http://cognimuse.cs.ntua.gr/, accessed: 2015-07-25.

[2] M. Barnsley, *Fractals everywhere*. Academic Press, San Diego, 1988.

[3] R. Bladon and B. Lindblom, "Modeling the judgment of vowel quality differences," *J. Acoust. Soc. Am.*, vol. 69, no. 5, pp. 1414–1422, 1981.

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th annu. workshop Computational learning theory*. ACM, 1992, pp. 144–152.

[5] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[6] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[7] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acust. united Ac.*, vol. 86, no. 1, pp. 117–128, 2000.

[8] M. Campbell and C. Greated, *The musician's guide to acoustics*. Oxford University Press, 1994.

[9] R. P. Carlyon, R. Cusack, J. M. Foxton, and I. H. Robertson, "Effects of attention and unilateral neglect on auditory stream segregation." *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 27, no. 1, p. 115, 2001.

[10] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/25/5/10.1121/1.1907229

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] R. Cusack and R. P. Carlyon, "Perceptual asymmetries in audition," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 29, pp. 713–725, 2003.

[13] R. Cusack, J. Decks, G. Aikman, and R. P. Carlyon, "Effects of location, frequency region, and time course of selective attention on auditory scene analysis." *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 30, no. 4, p. 643, 2004.

[14] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[15] V. D. Delmotte, "Computational auditory saliency," Ph.D. dissertation, Georgia Institute of Technology, 2012.

[16] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. review of neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[17] A. Duchowski, *Eye tracking methodology: Theory and practice*. Springer Science & Business Media, 2007, vol. 373.

[18] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[19] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[20] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention—focusing the searchlight on sound," *Current opinion in neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[21] S. Harding, M. Cooke, and P. König, "Auditory gist perception: an alternative to attentional selection of auditory streams?" in *Proc. WAPCV*. Springer, 2007, pp. 399–416.

[22] W. M. Hartmann, S. McAdams, and B. K. Smith, "Hearing a mistuned harmonic in an otherwise periodic complex tone," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1712–1724, 1990.

[23] ISO, "Acoustics–normal equal-loudness-level contours," International Organization for Standardization, Geneva, Switzerland, ISO, 2003.

[24] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[25] ——, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[26] ——, "Feature combination strategies for saliency-based visual attention systems," *J. of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[27] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[28] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *IEEE Trans. Acoust., Speech, Signal Process.*, 1990, pp. 381–384.

[29] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 5, pp. 1009–1024, 2009.

[30] E. M. Kaya and M. Elhilali, "A temporal saliency map for modeling auditory attention," in *Proc. 46th Annu. Conf. Information Sciences and Systems*, 2012.

[31] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[32] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, 2003.

[33] K. Kim, K.-H. Lin, D.-B. Walther, M.-A. Hasegawa-Johnson, and T.-S. Huang, "Automatic detection of auditory salience with optimized linear filters derived from human annotation," *Pattern Recognition Lett.*, vol. 38, pp. 78–85, 2014.

[34] M. Kipp, "Multimedia annotation, querying and analysis in anvil," *Multimedia Information extraction*, vol. 19, 2010.

[35] A. G. Leventhal, *The neural basis of visual function: vision and visual dysfunction*. CRC Press, 1991, vol. 4.

[36] D. T. Levin and D. J. Simons, "Failure to detect changes to attended objects in motion pictures," *Psychonomic Bulletin & Review*, vol. 4, no. 4, pp. 501–506, 1997.

[37] B. Mandelbrot, *The fractal geometry of nature*. Macmillan, 1983.

[38] P. Maragos, "Fractal signal analysis using mathematical morphology," *Advances in electronics and electron physics*, vol. 88, pp. 199–246, 1994.

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

[40] ——, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1532–1550, 1993.

[41] P. Maragos and A. Potamianos, "Fractal dimensions of speech sounds: Computation and application to automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 105, no. 3, pp. 1925–1932, 1999.

[42] P. Maragos and F.-K. Sun, "Measuring the fractal dimension of signals: morphological covers and iterative optimization," *IEEE Trans. Signal Process.*, vol. 41, no. 1, pp. 108–121, 1993.

[43] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[44] B. C. Moore and S. P. Bacon, "Detection and identification of a single modulated carrier in a complex sound," *J. Acoust. Soc. Am.*, vol. 94, no. 2, pp. 759–768, 1993.

[45] B. C. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acust. united Ac.*, vol. 82, no. 2, pp. 335–345, 1996.

[46] B. C. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Engin. Soc.*, vol. 45, no. 4, pp. 224–240, 1997.

[47] B. C. Moore, B. R. Glasberg, and R. W. Peters, "Thresholds for hearing mistuned partials as separate tones in harmonic complexes," *J. Acoust. Soc. Am.*, vol. 80, no. 2, pp. 479–483, 1986.

[48] A. Oliva, "Gist of the scene," *Neurobiology of Attention*, vol. 696, p. 64, 2005.

[49] J. B. Pierrehumbert, "The phonology and phonetics of english intonation," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[50] R. Plomp, "The ear as a frequency analyzer," *J. Acoust. Soc. Am.*, vol. 36, no. 9, pp. 1628–1636, 1964.

[51] M. C. Potter and E. I. Levy, "Recognition memory for a rapid sequence of pictures." *J. Exp. Psychol.*, vol. 81, no. 1, p. 10, 1969.

[52] K. Robinson and R. D. Patterson, "The duration required to identify the instrument, the octave, or the pitch chroma of a musical note," *Music Perception*, pp. 1–15, 1995.

[53] ——, "The stimulus duration required to identify vowels, their octave, and their pitch chroma," *J. Acoust. Soc. Am.*, vol. 98, no. 4, pp. 1858–1865, 1995.

[54] C. E. Schreiner, H. L. Read, and M. L. Sutter, "Modular organization of frequency integration in primary auditory cortex," *Annu. review neuroscience*, vol. 23, no. 1, pp. 501–529, 2000.

[55] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.*, vol. 66, no. 6, pp. 1647–1652, 1979.

[56] P. G. Schyns and A. Oliva, "From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition," *Psychological science*, vol. 5, no. 4, pp. 195–200, 1994.

[57] W. A. Sethares, *Timbre, tuning, spectrum, scale*. New York: Springer, 1998.

[58] M. Slaney and R. F. Lyon, "On the importance of time-a temporal representation of sound," *Visual Representations of Speech Signals*, pp. 95–116, 1993.

[59] R. Sternberg, *Cognitive psychology*. Cengage Learning, 2008.

[60] E. Terhardt, "On the perception of periodic sound fluctuations (roughness)," *Acta Acust. united Ac.*, vol. 30, no. 4, pp. 201–213, 1974.

[61] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[62] T. Tsuchida and G. W. Cottrell, "Auditory saliency using natural statistics," in *Annu. Meeting Cognitive Science Society*, 2012, pp. 1048–1053.

[63] V.-N. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.

[64] P. N. Vassilakis, "Perceptual and physical properties of amplitude fluctuation and their musical significance," Ph.D. dissertation, University of California, Los Angeles, 2001.

[65] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Conf. Computer Vision, Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3360–3367.

[66] K. Wang and S. Shamma, "Self-normalization and noise robustness in early auditory processing," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 3, pp. 421–435, 1994.

[67] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conf. Computer Vision, Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1794–1801.

[68] W. Yang, "Enhanced modified bark spectral distortion (embsd): An objective speech quality measure based on audible distortion and cognition model," Ph.D. dissertation, Temple University, 1999.

[69] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 824–839, March 1992.

[70] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in neural information processing systems*, 2009, pp. 2223–2231.

[71] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, p. 32, 2008.

[72] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoust. Soc. Am.*, vol. 33, no. 2, p. 248, 1961.