
Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων
και Μηχανικων Υπολογιστων

Τομεας Σηματων, Ελεγχου και Ρομποτικης


# Επεξεργασία Σημάτων Μουσικής και Εφαρμογές Αναγνώρισης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ


**ΑΘΑΝΑΣΙΑ Χ. ΖΛΑΤΙΝΤΣΗ**

Διπλωματούχος Μηχανικός: Master of Science in Media Technology
Royal Institute of Technology (KTH), Stockholm, Sweden.

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


Αθήνα, Δεκέμβριος 2013





Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) - Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος ΙΙ . Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων
και Μηχανικων Υπολογιστων

Τομεας Σηματων, Ελεγχου και Ρομποτικης

# Επεξεργασία Σημάτων Μουσικής και Εφαρμογές Αναγνώρισης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

**ΑΘΑΝΑΣΙΑ Χ. ΖΛΑΤΙΝΤΣΗ**

Διπλωματούχος Μηχανικός: Master of Science in Media Technology
Royal Institute of Technology (KTH), Stockholm, Sweden

**Συμβουλευτική Επιτροπή:** Καθ. Πέτρος Μαραγκός (Επιβλέπων)
Καθ. Γεώργιος Καραγιάννης
Καθ. Στέφανος Κόλλιας

Εγκρίθηκε από την επταμελή επιτροπή στις 20/12/ 2013 :

| | | |
|---|---|---|
| Π. Μαραγκός | Γ. Καραγιάννης | Σ. Κόλλιας |
| Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. |
| Κ. Τζαφέστας | Γ. Ποταμιάνος | |
| Επικ. Καθηγήτης Ε.Μ.Π. | Αναπλ. Καθηγητής Παν/μιο Θεσσαλίας | |
| Α. Πικράκης | Ε.-Σ. Φωτεινέα | |
| Λεκτ. Παν/μιο Πειραιώς | Ερευν. Β', ΙΕΛ, ΕΚ-Αθηνά | |

Αθήνα, Δεκέμβριος 2013




Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) - Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος ΙΙ . Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

. . . . . .

**ΑΘΑΝΑΣΙΑ Χ. ΖΛΑΤΙΝΤΣΗ**
Διδάκτωρ Μηχανικός Ε.Μ.Π.

# Περιεχόμενα

# Κατάλογος Σχημάτων

# Κατάλογος Πινάκων

# ΠΡΟΛΟΓΟΣ

*Σα βγεις στον πηγαιμό για την Ιθάκη, να εύχεσαι νάναι μακρύς ο δρόμος,*
*γεμάτος περιπέτειες, γεμάτος γνώσεις. Τους Λαιστρυγόνας και τους Κύκλωπας,*
*του θυμωμένο Ποσειδώνα μη φοβάσαι, τέτοια στον δρόμο σου ποτέ σου δεν θα βρεις,*
*αν μέν᾽ η σκέψις σου υψηλή, αν εκλεκτή συγκίνησις το πνεύμα και το σώμα σου αγγίζει. . . .*
*Πάντα στον νου σου νάχεις την Ιθάκη. Το φθάσιμον εκεί είν᾽ ο προορισμός σου.*
*Αλλά μη βιάζεις το ταξείδι διόλου. Καλλίτερα χρόνια πολλά να διαρκέσει·*
*και γέρος πια ν᾽ αράξεις στο νησί, πλούσιος με όσα κέρδισες στον δρόμο*[1] . . .

Ο δρόμος για την ολοκλήρωση της εργασίας αυτής ήταν όντως μακρύς και κάποιες φορές δύσκολος με πολλές χαρούμενες στιγμές αλλά και απογοητεύσεις. Πάνω απ᾽ όλα όμως αποτέλεσε ένα σημαντικό ταξίδι και όπως κάθε ταξίδι ήταν γεμάτο καινούριες γνώσεις, πρωτόγνωρες εμπειρίες και εκπλήξεις.

Στον δρόμο αυτό προς την «*Ἰθακη*» μου υποστηρικτής ο καθηγητής μου κ. Πέτρος Μαραγκός, τον οποίο και ευχαριστώ θερμά για την ευκαιρία που μου έδωσε να ασχοληθώ με την έρευνα σε τόσο υψηλό επίπεδο. Έθεσε τα θεμέλια της γνώσης και της δίψας για συνεχή αναζήτηση νέων εξωτικών μονοπατιών και αποτέλεσε έμπνευση καθ᾽ όλη τη διάρκεια της εκπόνησης της διατριβής αυτής. Στάθηκε καθοδηγητής όλα αυτά τα χρόνια, σύμβουλος και συμπαραστάτης, μετατρέποντας την όποια δυσκολία και απογοήτευση σε δημιουργία. Ένα θερμό ευχαριστώ επίσης στον Αλέξανδρο Ποταμιάνο του οποίου οι συμβουλές σε πολλές περιπτώσεις υπήρξαν καθοριστικές.

Θα ήθελα επίσης να ευχαριστήσω τα μέλη της επταμελούς εξεταστικής επιτροπής της διατριβής, τους κ. Γεώργιο Καραγιάννη, Στέφανο Κόλλια, Κωνσταντίνο Τζαφέστα, Γεράσιμο Ποταμιάνο, Άγγελο Πικράκη και την κ. Ευίτα Φωτεινέα για τα σχόλια και τις παρατηρήσεις ως προς το περιεχόμενο, τις ιδέες αλλά και τις μελλοντικές κατευθύνσεις της εργασίας αυτής.

Δεν θα μπορούσα να παραλείψω τους συναδέλφους του εργαστηρίου, συνεργάτες και φίλοι μαζί. Ξεκινώντας από τα πιο παλιά μέλη, ένα ιδιαίτερο ευχαριστώ στον Γιώργο Ευαγγελόπουλο στον οποίο χρωστάω πολλά, η βοήθεια του ήταν ανεκτίμητη και οι συμβουλές του ανυπολόγιστης αξίας για τα επόμενα χρόνια. Στον Βασίλη που μου έμαθε

[1] Κ.Π. Καβάφης, Τα Ποιήματα (1897-1933)

να αναζητώ και να ψάχνω, στον πάντα χαμογελαστό Νάσο ο οποίος στάθηκε δίπλα μου σε κάθε δυσκολία. Στους Γιώργο, Σταμάτη, Τάσο και Δημήτρη, όλοι τους πρόσφεραν απλόχερα τις συμβουλές και τις γνώσεις τους. Ο Σταύρος ότι και αν χρειάστηκα πάντα ήταν πρόθυμος να βοηθήσει. Ο Ισίδωρος ο πρώτος που συναντάς στο εργαστήριο το πρωί και ο τελευταίος που χαιρετάς το βράδυ. Και φυσικά ο Νώντας του οποίου τη φιλία εκτίμησα ιδιαίτερα. Στα νέα μέλη του εργαστηρίου, κάτι που δεν τα κάνει λιγότερο σημαντικά, την Αντιγόνη, τον Πέτρο και τον Παναγιώτη. Ένα ευχαριστώ και στην πάντα χαμογελαστή Μόνικα, που έστω και αν ήρθε για μικρό χρονικό διάστημα, εκτός της βοήθεια της που ήταν σημαντική, έφερε χαρά, ζωντάνια αλλά και ζεστασιά στον χώρο του εργαστηρίου, αλλά και στην αγαπημένη μου Δέσποινα, μία και μοναδική, ακούραστη, πάντα πρόθυμη.

Η παρούσα διατριβή πραγματοποιήθηκε στο εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου & Επεξεργασίας Σήματος (CVSP) του Ε.Μ.Π., στα πλαίσια του προγράμματος Ηράκλειτος ΙΙ, μέσω του επιχειρησιακού προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ).

Επιπλέον θα ήθελα να ευχαριστήσω πολλούς καλούς φίλους. Τον Ηλία για τη συμβολή του ειδικά τον τελευταίο χρόνο. Οι συμβουλές του άπειρες για την ομαλή διεκπεραίωση αυτής της εργασίας, κρίμα που δεν τις ακολουθούσα όμως πάντα. Η συμπαράσταση του επίσης καθοριστική το τελευταίο τρίμηνο, μιας και αποδείχθηκε πολλές φορές «σωτήρας» σε αξιοπερίεργες καταστάσεις. Την αγαπημένη μου Κατερίνα για τη συνεχή της συμπαράσταση, πάντα ακούραστη, πάντα δίπλα, τον Θοδωρή για τις πολλές και ενδιαφέρουσες μουσικές συζητήσεις αλλά και τον μικρό Άγγελο που μας έφερε ιδιαίτερη χαρά τον τελευταίο χρόνο. Τον Τόλη για τη βοήθεια του με τη δημιουργία των αυτόματων περιλήψεων αλλά και για το καταπληκτικό του χιούμορ που όλα αυτά τα χρόνια μας κάνει και γελάμε. Τη Λουκία, τον Πάνο και τον μικρό Άρη, την Άντα, τον Γιωργάκη και φυσικά τη Λένια για τη βοήθεια της με πολύτιμα σχόλια και διορθώσεις. Τέλος, ευχαριστώ τον αγαπημένο μου φίλο Νίκο για τις πολύτιμες μουσικές του γνώσεις και ιδέες, τον επιβλέποντα της διπλωματικής μου εργασίας Kjetil Falkenberg Hansen, καθώς και όλους τους «μουσικούς συνοδοιπόρους» των χρόνων αυτών.

Τέλος την εργασία αυτή την αφιερώνω στους ανθρώπους που υποστήριξαν την κάθε μου προσπάθεια· στον πατέρα μου που μου έμαθε πως μπορώ να κατορθώσω τα πάντα με «*υπομονή και επιμονή*», στη μητέρα μου όχι μόνο γιατί από πολύ μικρή ηλικία μου έμαθε την *Ιθάκη*, αλλά γιατί από την ίδια ηλικία μου δίδαξε το πραγματικό της νόημα και στον αγαπημένο μου αδελφό. Τέλος ευχαριστώ με όλη μου την καρδιά τον Παναγιώτη! Βρέθηκε δίπλα μου από την αρχή και άντεξε ως και το τέλος· συνοδοιπόρος, συμπαραστάτης, σύμβουλος, *competitor in crime*. . . . . .

*Νάνσυ Ζλατίντση*
*Δεκέμβριος 2013*

# ΠΕΡΙΛΗΨΗ

Η διδακτορική αυτή έρευνα ασχολείται με το θέμα της ψηφιακής επεξεργασίας μουσικών σημάτων και την ανάλυσή τους με υπολογιστικές μεθόδους με στόχο την εξαγωγή χρήσιμης πληροφορίας για την αναγνώρισή τους. Συγκεκριμένα μελετάμε και αναπτύσσουμε αποτελεσματικούς αλγορίθμους, με τη χρήση μη-γραμμικών μοντέλων, για την επεξεργασία των σημάτων μουσικής, την κατανόηση μουσικών φαινομένων και τη μοντελοποίηση τους. Εστιάζουμε στη διερεύνηση και την ανάλυση των σχέσεων μεταξύ των μουσικών οργάνων για την κατανόηση της λειτουργίας και των χαρακτηριστικών τους. Εξετάζουμε τα γνωρίσματα των διαφορετικών ειδών μουσικής, ενώ επιπλέον αξιολογούμε την αποτελεσματικότητα των μη-γραμμικών μοντέλων για την ανίχνευση σημαντικών μουσικών και γενικά ακουστικών γεγονότων.

Η ανάλυση αυτή συνεισφέρει στην έρευνα και στην τεχνολογία αιχμής που σχετίζεται με την αυτόματη κατηγοριοποίηση μουσικής μέσω των διαφορετικών αυτών πλαισίων, αλλά και στη δημιουργία περιλήψεων των ηχητικών σημάτων. Τέτοιες εφαρμογές στις μέρες μας συναντώνται ευρέως σε εφαρμογές από το λογισμικό υπολογιστών έως τα κινητά τηλέφωνα τρίτης γενιάς. Λόγω της πληθώρας των ηχητικών, μουσικών, αλλά και πολυμεσικών δεδομένων, η χρησιμότητα της μελέτης αυτής διαφαίνεται σε εφαρμογές όπως η αυτόματη αναζήτηση μουσικής με βάση το είδος, η αναγνώριση βασικών δομών της μουσικής, όπως για παράδειγμα τα μουσικά όργανα, και η δημιουργία περιλήψεων.

Με βάση το πλαίσιο αυτό προτείνουμε νέα χαρακτηριστικά για τη μοντελοποίηση των σημάτων μουσικής. Η πειραματική αξιολόγηση τους τεκμηριώνει τη δυναμική των μεθόδων που ακολουθούμε καθώς τα αποτελέσματα παρουσιάζονται ιδιαίτερα ενθαρρυντικά. Συγκεκριμένα, η έρευνα αυτή δείχνει πως τα προτεινόμενα χαρακτηριστικά δύνανται να περιγράψουν σημαντικά φαινόμενα των μουσικών σημάτων όπως για παράδειγμα τις μικρο-μεταβολές των δομών τους. Επιπλέον, αναπαραστάσεις που βασίζονται στις μακροδομές των σημάτων επιφέρουν μείωση της πολυπλοκότητας του συστήματος κατηγοριοποίησης, εφόσον ικανοποιητικά αποτελέσματα επιτυγχάνονται με απλούστερα στατιστικά μοντέλα. Τέλος, η εισαγωγή ιδεών όπως η «μουσική» συστοιχία φίλτρων επιδεικνύει ιδιαίτερη διακριτική ικανότητα στην κατηγοριοποίηση των μουσικών σημάτων.

# ABSTRACT

This thesis lays in the area of signal processing and analysis of music signals using computational methods for the extraction of effective representations for automatic recognition. We explore and develop efficient algorithms using nonlinear methods for the analysis of the structure of music signals, which is of importance for their modeling. Our main research directions deals with the analysis of the structure and the characteristics of musical instruments in order to gain insight about their function and properties. We study the characteristics of the different genres of music. Finally, we evaluate the effectiveness of the proposed nonlinear models for the detection of perceptually important music and audio events.

The approach we follow contributes to state-of-the-art technologies related to automatic computer-based recognition of musical signals and audio summarization, which nowadays are essential in everyday life. Because of the vast amount of music, audio and multimedia data in the web and our personal computers, the use of this study could be shown in applications such as automatic genre classification, automatic recognition of music's basic structures, such as musical instruments, and audio content analysis for music and audio summarization.

The above mentioned applications require robust solutions to information processing problems. Toward this goal, the development of efficient digital signal processing methods and the extraction of relevant features is of importance. In this thesis we propose such methods and algorithms for feature extraction with interesting results that render the descriptors of direct applicability. The proposed methods are applied on classification experiments illustrating that they can capture important aspects of music, such as the micro-variations of their structure. Descriptors based on macro-structures may reduce the complexity of the classification system, since satisfactory results can be achieved using simpler statistical models. Finally, the introduction of a "music" filterbank appears to be promising for automatic genre classification.

# ΕΚΤΕΝΗΣ ΠΕΡΙΛΗΨΗ

Η διδακτορική αυτή έρευνα ασχολείται με το θέμα της ψηφιακής επεξεργασίας μουσικών σημάτων και την ανάλυσή τους με υπολογιστικές μεθόδους με στόχο την εξαγωγή χρήσιμης πληροφορίας για την αναγνώρισή τους. Συγκεκριμένα μελετώνται και αναπτύσσονται αποτελεσματικοί αλγόριθμοι, με τη χρήση μη-γραμμικών μοντέλων, για την επεξεργασία των σημάτων μουσικής, την κατανόηση μουσικών φαινομένων και τη μοντελοποίησή τους. Οι κυρίες κατευθύνσεις της έρευνας αφορούν τη διερεύνηση και την ανάλυση των σχέσεων μεταξύ των μουσικών οργάνων για την κατανόηση της λειτουργίας και των δομών τους, την εξέταση των χαρακτηριστικών των διαφορετικών ειδών μουσικής, και τέλος την αξιολόγηση της αποτελεσματικότητας των μοντέλων αυτών για την ανίχνευση σημαντικών μουσικών και γενικά ακουστικών γεγονότων.

Η ανάλυση αυτή συνεισφέρει στην έρευνα και στην τεχνολογία αιχμής που σχετίζεται με την αυτόματη κατηγοριοποίηση μουσικής μέσω των διαφορετικών αυτών πλαισίων, αλλά και στη δημιουργία περιλήψεων των ηχητικών σημάτων. Τέτοιες εφαρμογές στις μέρες μας συναντώνται ευρέως στην καθημερινότητα, επιζητώνται δε πλέον, όχι μόνο από ανθρώπους εξειδικευμένης ή ανώτερης τεχνολογικής και μουσικολογικής γνώσης, αλλά και από το ευρύ κοινό, σε εφαρμογές από το λογισμικό υπολογιστών έως τα κινητά τηλέφωνα τρίτης γενιάς. Ακριβώς λόγω αυτής της πληθώρας των ηχητικών, μουσικών, αλλά και πολυμεσικών δεδομένων, η χρησιμότητα της συγκεκριμένης έρευνας διαφαίνεται σε εφαρμογές όπως η αυτόματη αναζήτηση μουσικών κομματιών με βάση το είδος, η αναγνώριση βασικών δομών της μουσικής, όπως για παράδειγμα τα μουσικά όργανα, και η γρήγορη αναζήτηση πληροφοριών του περιεχομένου των δεδομένων με τη βοήθεια συνοπτικών ηχητικών αποσπασμάτων. Συγκεκριμένα, τα ερευνητικά πεδία της έρευνας αυτής συνοψίζονται στη συνέχεια:

*Επεξεργασία των μουσικών σημάτων διαφορετικών μουσικών οργάνων στα πλαίσια της φράκταλ γεωμετρίας.*

Ξεκινώντας από το πρωταρχικό πρόβλημα της διερεύνησης των δομών των σημάτων μουσικής, εξετάσαμε την πολυπλοκότητα και την ανομοιογένεια των μουσικών σημάτων με μετρήσεις βασισμένες στη Φράκταλ (*Fractal*) θεωρία, όπου και διερευνήθηκε η φράκταλ διάσταση των μουσικών σημάτων σε πολλαπλές κλίμακες (*multiscale fractal*

*dimension, MFD*). Από διάφορες μελέτες έχει προκύψει πως οι μουσικοί ήχοι εμφανίζουν φράκταλ δομές. Στη συγκεκριμένη έρευνα εστιάσαμε στα διαφορετικά μεταβατικά στάδια των μουσικών σημάτων (π.χ. στην αρχή της μουσικής νότας (*attack*) και στη σταθερή κατάσταση *(sustain)*) για την εύρεση ακριβώς αυτής της πληροφορίας που βοηθά στην κατηγοριοποίησή τους. Παράλληλα, εφαρμόσαμε τον αλγόριθμο μορφολογικής κάλυψης σε συνθετικά σήματα για να αξιολογήσουμε διάφορες παρατηρήσεις μας, ενώ ενσωματώσαμε τη φράκταλ διάσταση σε πολλαπλές κλίμακες ως χαρακτηριστικό σε πειράματα αναγνώρισης μουσικών οργάνων, παρατηρώντας πως τα MFD μπορούν να διακρίνουν διάφορα χαρακτηριστικά τους. Η αξιολόγηση διενεργήθηκε με τη χρήση αλγορίθμων της αναγνώρισης προτύπων, όπως τα Γκαουσιανά μοντέλα (*Gaussian Mixture Models, GMM*) και τα κρυφά Μαρκοβιανά μοντέλα (*hidden Markov models, HMM*), σε σύγκριση αλλά και σε συνδυασμό με βασικά χαρακτηριστικά, όπως τα Mel Frequency Cepstral Coefficients (MFCC), με πολύ καλά αποτελέσματα αναγνώρισης των διαφορετικών μουσικών οργάνων και μείωση σφάλματος έως και 32%.

*Ανάλυση των μουσικών σημάτων διαφορετικών μουσικών οργάνων με μη-γραμμικά μοντέλα διαμορφώσεων.*

Έχοντας ως κύριο στόχο την αναζήτηση των δομών αυτών που διαχωρίζουν τη μουσική, συνεχίσαμε εξετάζοντας τα μοντέλα διαμορφώσεων (*amplitude and frequency modulation models, AM-FM*), όπου εφαρμόζεται η πολυζωνική ανίχνευση τους σε σήματα μουσικής, τα οποία αναλύονται από συστοιχίες φίλτρων. Συγκεκριμένα, μελετήσαμε και επεκτείναμε τον αλγόριθμο διαχωρισμού ενέργειας *ESA (Energy Separation Algorithm)*, για την εξαγωγή εύρωστων αναπαραστάσεων, όπως το μέσο στιγμιαίο πλάτος και η μέση στιγμιαία συχνότητα, και τη μοντελοποίηση των μικροδομών των ήχων των μουσικών οργάνων. Παρατηρώντας την ακρίβεια του μοντέλου και τη δυνατότητα των χαρακτηριστικών για τη συγκεκριμένη εφαρμογή συνεχίσαμε με την ανάλυση και εφαρμογή του Επαναληπτικού-ESA (*Iterative-ESA*). Βάσει της θεώρησης αυτής, καταλήξαμε σε σημαντικά συμπεράσματα τα οποία σχετίζονται με τη δυνατότητα του αλγορίθμου για τη δημιουργία καλύτερων εκτιμήσεων των χαρακτηριστικών, πιο εύστοχο προσδιορισμό των δομών της μουσικής ενώ επίσης διαπιστώσαμε πως η μέθοδος αυτή θα μπορούσε πιθανώς να χρησιμοποιηθεί για την εκτίμηση του αρμονικού περιεχομένου των μουσικών ήχων. Η ανάλυση καθώς και η πειραματική αξιολόγηση των προτεινόμενων μεθόδων και χαρακτηριστικών υλοποιήθηκε με τη χρήση GMM και HMM, σε σχέση με βασικά χαρακτηριστικά τα οποία επιλέχθηκαν για την καλή τους απόδοση, οδηγώντας σε πολύ καλά αποτελέσματα αναγνώρισης με μείωση λάθους έως και 60%.

*Ανάλυση των μουσικών σημάτων διαφορετικών ειδών μουσικής με μη-γραμμικά μοντέλα*

*διαμορφώσεων.*

Για το πρόβλημα της ανάλυσης και κατηγοριοποίησης των διαφορετικών ειδών μουσικής, λαμβάνοντας υπόψη τα αποτελέσματα της έρευνας που προηγήθηκε, αξιολογήσαμε το μοντέλο διαμόρφωσης πλάτους και συχνότητας (AM-FM) για τη μοντελοποίηση των μικροδομών και των μακροδομών των μουσικών σημάτων. Εξάγαμε χαρακτηριστικά όπως το μέσο στιγμιαίο πλάτος, η μέση στιγμιαία συχνότητα και το ποσοστό διαμόρφωσης της συχνότητας, επεκτείναμε τον αλγόριθμο διαχωρισμού ενέργειας ενώ παράλληλα προτείναμε τη δημιουργία συστοιχίας φίλτρων εστιασμένων στα μουσικά σήματα για την εξαγωγή πιο εύρωστων αναπαραστάσεων. Επιπλέον, αξιολογήσαμε διαφορετικές μορφές αναπαραστάσεων των χαρακτηριστικών, όπως για παράδειγμα περιγραφείς βασισμένους στην ανάλυση βραχέος χρόνου ή στις μακροδομές των σημάτων μουσικής. Τα αποτελέσματα της ανάλυσης και των πειραμάτων αναγνώρισης ήταν ιδιαίτερα ενθαρρυντικά με μείωση σφάλματος ως και 28% όταν τα προτεινόμενα χαρακτηριστικά συνδυάζονται με τα MFCC.

Σημαντικά συμπέρασμα της διερεύνησης αυτής είναι τα εξής: (α) οι διαμορφώσεις μπορούν να περιγράψουν σημαντικά φαινόμενα των σημάτων μουσικής, όπως για παράδειγμα τις μικρο-μεταβολές που συμβαίνουν λόγω των δομών της (π.χ. μελωδία, ρυθμός κ.ά.). (β) Η χρήση αναπαραστάσεων οι οποίες βασίζονται στις μακροδομές των μουσικών σημάτων επιφέρουν σημαντική μείωση της πολυπλοκότητας του συστήματος κατηγοριοποίησης, εφόσον επιτυγχάνουν ικανοποιητικά αποτελέσματα χρησιμοποιώντας απλούστερα μοντέλα τύπου GMM. (γ) Η εισαγωγή της «μουσικής» συστοιχίας φίλτρων έχει ως αποτέλεσμα τη δημιουργία συνόλων χαρακτηριστικών με ιδιαίτερα αξιόλογη διακριτική ικανότητα στην κατηγοριοποίηση των διαφορετικών ειδών μουσικής.

*Δημιουργία αναπαραστάσεων Bag-of-Words για την ανάλυση των μουσικών σημάτων διαφορετικών ειδών μουσικής.*

Για το πρόβλημα της κατηγοριοποίησης διαφορετικών ειδών μουσικής αρχικά προτάθηκε ο συνδυασμός διαφόρων χαρακτηριστικών, για παράδειγμα AM-FM, fractals και MFCC), με αποτέλεσμα ικανοποιητικά αποτελέσματα. Ακολουθώντας μια άλλη προσέγγιση, βασιστήκαμε σε ιδέες της Όρασης Υπολογιστών, διατυπώνοντας μία διαφορετική διαδικασία για την εξαγωγή χαρακτηριστικών μέσω του μοντέλου *Bag-of-Words* (BoW), καθιστώντας έτσι δυνατή την εισαγωγή εναλλακτικών αναπαραστάσεων των μουσικών σημάτων. Ακολουθώντας την προσέγγιση αυτή επιτυγχάνεται η δημιουργία ενός *«μουσικού λεξικού»*, το οποίο χρησιμοποιείται για την περιγραφή του κάθε μουσικού κομματιού βάσει της συχνότητας των *«μουσικών λέξεων»* που περιλαμβάνει. Οι αναπαραστάσεις που δημιουργούνται αξιολογούνται με τη χρήση των Support Vector Machines (SVMs) με μείωση λάθους ως και 11% για τον συνδυασμό των μη-γραμμικών

χαρακτηριστικών σε σχέση με τα MFCC και 16% σε συνδυασμό με τα MFCC. Με τη μέθοδο αυτή αντιμετωπίζουμε διάφορα προβλήματα πολυπλοκότητας κατά την κατηγοριοποίηση, λόγω των νέων συμπαγών αναπαραστάσεων.

*Μελέτη των μη-γραμμικών μοντέλων διαμορφώσεων για την ανίχνευση σημαντικών ακουστικών γεγονότων και τη δημιουργία ηχητικών περιλήψεων.*

Βασιζόμενοι σε προηγούμενη ερευνητική μας εργασία, εξετάσαμε την καταλληλότητα του μη-γραμμικού μοντέλου διαμόρφωσης πλάτους και συχνότητας (AM-FM) σε θέματα ανίχνευσης σημαντικών ακουστικών γεγονότων. Συγκεκριμένα, εξάγουμε χαρακτηριστικά όπως το μέσο στιγμιαίο πλάτος, η μέση στιγμιαία συχνότητα και η μέση τιμή της Teager ενέργειας και προτείνουμε υπολογιστικές μεθόδους για την ενδοτροπική σύμμειξή (*intramodal fusion*) τους και τη δημιουργία αναπαραστάσεων σημαντικότητας (*saliency*), οι οποίες και αποτελούν κριτήριο επιλογής αντιληπτικά σημαντικών ηχητικών γεγονότων για τη δημιουργία ηχητικών συνόψεων. Παράλληλα, επεκτείναμε τον αλγόριθμο δημιουργίας περιλήψεων προτείνοντας αλγόριθμο βελτίωσης του, ο οποίος βασίζεται σε ιδέες της μαθηματικής μορφολογίας, με αποτέλεσμα αντιληπτικά ποιοτικές περιλήψεις. Η αξιολόγηση των προτεινόμενων αλγορίθμων για την εξαγωγή γεγονότων και τη δημιουργία αυτόματων περιλήψεων διεξάγεται με εκτενείς ποσοτικές αξιολογήσεις.

Κατά τη διάρκεια της διδακτορικής αυτής διατριβής ασχοληθήκαμε επίσης με το θέμα της ανίχνευσης σημαντικών γεγονότων σε πολυμεσικά δεδομένα, αξιοποιώντας τα προηγούμενα αποτελέσματα μας και επεκτείνοντας προ-υπάρχουσα ερευνητική μας εργασία για τη δημιουργία περιλήψεων από ταινίες. Οι κατευθύνσεις με τις οποίες ασχοληθήκαμε αφορούν διεξοδικές ποσοτικές αξιολογήσεις του αλγόριθμου δημιουργίας περιλήψεων, έλεγχο της αποδοτικότητας των αυτόματων περιλήψεων με ποιοτικά κριτήρια και αξιολόγηση από χρήστες, και διερεύνηση καινούριων αναπαραστάσεων των σημάτων καθώς και μεθόδων σύμμειξής τους.

*Ανάπτυξη συστηματικής βάσης δεδομένων από ταινίες «MovieSum Database».*

Αλγόριθμοι ανίχνευσης γεγονότων και παραγωγής περιλήψεων μπορούν να βελτιωθούν σημαντικά όταν υπάρχουν οι κατάλληλες συλλογές δεδομένων για την εκπαίδευση, την προσαρμογή και την αξιολόγηση των παραμέτρων τους. Λόγω της ενασχόλησης μας με το θέμα της δημιουργίας συνόψεων σε ηχητικά αλλά και πολυμεσικά δεδομένα, αναπτύξαμε μία συστηματική βάση δεδομένων από ταινίες. Η βάση αυτή επισημειώνεται με τη μονοτροπική και την πολυτροπική σημαντικότητα του βίντεο καθώς και με τη σημασιολογική πληροφορία και αποτελεί σημαντικό εργαλείο για τις αξιολογήσεις των αλγορίθμων ανίχνευσης ηχητικών/πολυμεσικών γεγονότων και των ηχητικών και πολυμεσικών περιλήψεων.

# Λίστα Ακρωνυμίων

**AM** Amplitude Modulation - Διαμόρφωση Πλάτους

**AM-FM** Amplitude - Frequency Modulation - Μοντέλο διαμορφώσεων

**ESA** Energy Separation Algorithm - Αλγόριθμος Διαχωρισμού Ενέργειας

**FM** Frequency Modulation - Διαμόρφωση Συχνότητας

**ESA** Energy Separation Algorithm - Αλγόριθμος Διαχωρισμού Ενέργειας

**FMP** Frequency Modulation Percentage - Ποσοστό διαμόρφωσης της συχνότητας

**GMM** Gaussian Mixture Models - Γκαουσιανά Μοντέλα

**GTZAN** Βάση ηχητικών δεδομένων μουσικών ειδών

**HMM** Hidden Markov Models - Κρυφά Μαρκοβιανά Μοντέλα

**HTK** Hidden Markov Model Toolkit

**IOWA** Iowa University Instrument Samples - Βάση δεδομένων μουσικών οργάνων

**MFCC** Mel-Frequency Cepstral Coefficients - Τύπος χαρακτηριστικών

**MFD** Multiscale Fractal Dimension - Φράκταλ ανάλυση σε πολλαπλές κλίμακες (τύπος χαρακτηριστικών)

**PCA** Principal Component Analysis - Ανάλυση σε κύριες συνιστώσες

**SVM** Support Vector Machines

# Κεφάλαιο 1

# Εισαγωγή

## 1.1 Θέμα και Πρόβλημα

Η μουσική αποτελεί αναπόσπαστο κομμάτι της ζωής των ανθρώπων. Έχει μελετηθεί και εξακολουθεί να μελετάται από διαφορετικές σκοπιές από σχεδόν όλους τους θεωρητικούς και επιστημονικούς τομείς, κάτι που αποδεικνύει περίτρανα την εξέχουσα θέση που κατέχει στην καθημερινότητά μας. Η σπουδαιότητα της μουσικής φαίνεται επίσης από το γεγονός ότι αποτελεί σύμμειξη της τέχνης και της επιστήμης, της λογικής και του συναισθήματος, καθώς και της ψυχολογίας και της φυσιολογίας. Επηρεάζει την καθημερινότητά μας με ποικίλους τρόπους· άλλωστε σε πολλές μελέτες έχει αποδειχθεί πως συνειδητά επιλέγουμε να ακούσουμε μουσική πολλές φορές κατά τη διάρκεια της μέρας όχι απλώς για να περάσουμε την ώρα μας, αλλά για να αλλάξουμε τη διάθεσή μας [73, 164, 183].

Μύθοι και θρύλοι διηγούνται την απαρχή της μουσικής, την πορεία της εξέλιξής της και τη συνύπαρξή της με το ανθρώπινο είδος, ενώ αρχαίοι λαοί αποδίδουν τη «δημιουργία» της στους θεούς τους. Εξάλλου η δύναμη της μουσικής είναι γνωστή από την αρχαιότητα. Όλοι θυμόμαστε την Κίρκη, στην *Οδύσσεια*, να προειδοποιεί τον Οδυσσέα για το τραγούδι των Σειρήνων:

> *Πρώτα στο δρόμο που θα πας, θα φτάσεις στις Σειρήνες, που όλους μαγεύουν τους θνητούς, όσοι κοντά τους φτάσουν. Όποιος ζυγώσει ανύποπτα κι ακούσει τη λαλιά τους, αυτόν πια δε θα τον χαρούν το τρυφερό του ταίρι και τα μικρά του τα παιδιά, στο σπίτι να γυρίσει, ... Μόν' πέρνα από κοντά τους, και πιάσε των συντρόφων σου τ' αυτιά να φράξεις όλων μ' απαλομάλαχτο κερί, κανείς να μην ακούσει. Μα στο κατάρτι πρώτα ολόρθου χεροπόδαρα οι ναύτες να σε δέσουν, κι ας σφίξουν των παλαμαριών τις άκρες από πάνω, χαρούμενα το λάλημα υ' ακούσεις των Σειρήνων. Κι αν σκούζεις στους συντρόφους σου και θέλεις να σε*

*λύσουν, ακόμα τότε πιο γερά να σφίγγουν τα δεσμά σου* [1].

Σε αυτή τη διδακτορική διατριβή, ακριβώς αυτή η δύναμη της μουσικής μάς ώθησε να προσεγγίσουμε το θέμα μελετώντας διαφορετικές πτυχές της μέσω της τεχνολογίας. Πραγματευόμαστε το θέμα α) εξετάζοντας τη δομή και τα χαρακτηριστικά των μουσικών οργάνων, (β) προσδιορίζοντας τα χαρακτηριστικά των διαφορετικών ειδών μουσικής και (γ) αναλύοντας τη δομή των ηχητικών σημάτων για την εύρεση σημαντικών ακουστικών γεγονότων (*audio salient events*). Επικεντρωνόμαστε στη μελέτη δύο βασικών μη-γραμμικών μεθοδολογιών, τις οποίες εξετάζουμε και επεκτείνουμε εστιάζοντας στην καταλληλότητά τους για τις βασικές εφαρμογές της αυτόματης κατηγοριοποίησης της μουσικής μέσω των διαφορετικών αυτών πλαισίων, αλλά και της δημιουργίας περιλήψεων και συνοπτικών ηχητικών αποσπασμάτων των ηχητικών σημάτων. Δεδομένου του ρόλου της μουσικής, η ανάλυση αυτή συνεισφέρει στην τεχνολογία αιχμής. Στις μέρες μας εφαρμογές όπως η αυτόματη αναζήτηση μουσικών κομματιών με βάση το είδος, ή η αναγνώριση βασικών δομών της μουσικής, όπως για παράδειγμα τα μουσικά όργανα, συναντώνται ευρέως, επιζητώνται δε πλέον από την πλειοψηφία των ανθρώπων.

Στις επόμενες σελίδες της Εισαγωγής συνεχίζουμε τη συζήτηση για τη μουσική προσπαθώντας να τεκμηριώσουμε τη σημαντικότητα του ρόλου της και να θέσουμε τα θεμέλια των ιδεών και μεθοδολογιών που ακολουθούμε.

## 1.2 Μουσική

Η μουσική αποτελεί το επίκεντρο των ερευνών διάφορων ακαδημαϊκών τομέων. Για παράδειγμα, οι νευροεπιστήμες ασχολούνται με την επίδραση της μουσικής στον ανθρώπινο εγκέφαλο, ενώ η ψυχολογία αντιμετωπίζει τη μουσική ως γνωσιακό αντικείμενο και διερευνά τις επιπτώσεις της στο ανθρώπινο σώμα. Οι ανθρωπολόγοι και οι εθνομουσικολόγοι μάς δίνουν την ευρύτερη εικόνα της θέσης που κατέχει η μουσική στον ανθρώπινο πολιτισμό, ενώ ο χαρακτήρας και η δομή της μουσικής διερευνώνται από τους μουσικολόγους. Θεωρητικοί μουσικοί και ιστορικοί ασχολούνται με την ανάλυση και την εξιστόρηση της συμπεριφοράς των προγόνων μας σε σχέση με τη μουσική, ενώ ταυτοχρόνως μας δείχνουν πώς η στάση τους έχει επηρεάσει τις δικές μας συνήθειες ακρόασης. Ακόμα και εμπειρογνώμονες βιολόγοι/ορνιθολόγοι αναζητούν να βρουν το κοινό σημείο της μελωδίας και του τραγουδιού στα διάφορα είδη, συγκρίνοντας το τραγούδι άλλων θηλαστικών και πουλιών με το ανθρώπινο, προσπαθώντας με αυτόν τον τρόπο να δώσουν απαντήσεις για την εξελικτική πορεία του ανθρώπου.

[1] Ομήρου *Οδύσσεια*. Οργανισμός εκδόσεων διδακτικών βιβλίων (μετάφραση: Ζήσιμου Σιδέρη), 1992, σελ. 231.

Αξίζει να αναφέρουμε πως η μουσική έπαιζε ιδιαίτερα σημαντικό ρόλο ήδη από την αρχαιότητα· και συγκεκριμένα στη ζωή και την εκπαίδευση των αρχαίων Αθηναίων. Ο Πυθαγόρας ήταν ο πρώτος που έθεσε τις βάσεις της επιστήμης της Μουσικής με την επιστημονικά θεμελιωμένη μουσική θεωρία του. Οι Πυθαγόρειοι εμπνεόμενοι από την «*ιερά τετρακτύ*»[2] ανακάλυψαν την έννοια της σειράς των αρμονικών, χρησιμοποιώντας τους αριθμούς 1, 2, 3 και 4 και χωρίζοντας την χορδή του κανόνα με τους λόγους 1/2, 2/3, 3/4 με αποτέλεσμα τη μαθηματική έκφραση της οκτάβας, της καθαρής πέμπτης και της καθαρής τετάρτης δηλαδή των βασικών διαστημάτων της μουσικής [45]. Εν συνεχεία, η μουσική εξετάστηκε εξονυχιστικά τόσο από τον Πλάτωνα όσο και από μεταγενέστερους του φιλοσόφους, οι οποίοι από την αρχαιότητα θέτουν τα μεγάλα ερωτήματα και προσπαθούν να λύσουν το μυστήριο γύρω από την τέχνη της μουσικής.

Οι ερωτήσεις που έχουν τεθεί από ερευνητές, φιλοσόφους και ιστορικούς είναι πολλές, και κυμαίνονται από τις πιο απλές ως τις πλέον εκκεντρικές και δυσνόητες. Από την άλλη, οι απαντήσεις που έχουν δοθεί είναι ελάχιστες. Αναφέρουμε ενδεικτικά μερικά από αυτά τα ερωτήματα, δείχνοντας έτσι πως η αναζήτηση με θέμα τη μουσική είναι ανεξάντλητη.

Τι είναι η μουσική;

Ποιος ο ρόλος της μουσικής στην ιστορία της ανθρωπότητας, από την αρχαιότητα ως την εποχή μας, αλλά και σε κάθε γνωστό πολιτισμό;

Ποια η αξία της μουσικής και κατά πόσο μπορεί να συγκριθεί με άλλες μορφές τέχνης, όπως η ποίηση, η πεζογραφία ή μια επιστημονική ανακάλυψη;

Τι είδους συναισθήματα μπορεί να προκαλέσει η μουσική και ποια η βαρύτητα μιας έντονης συναισθηματικής εμπειρίας;

Για ποιο λόγο μουσικές συνθέσεις μεγάλης καλλιτεχνικής αξίας αποτυγχάνουν να επηρεάσουν τους ακροατές;

Υπάρχουν όρια ως προς το τι μπορούν μουσικοί και συνθέτες να μεταφέρουν μέσω της μουσικής;

Μπορεί η μουσική να αντιπροσωπεύσει μη μουσικά φαινόμενα;

Πώς έχει επηρεάσει η ανάπτυξη της τεχνολογίας καταγραφής την κατανόησή μας για τη μουσική αλλά και την απόλαυση που αντλούμε από αυτή;

Στη συνέχεια του κεφαλαίου αυτού θα προσπαθήσουμε να καλύψουμε συνοπτικά κάποιους από τους προαναφερθέντες τομείς αλλά και να δούμε ποιες είναι οι απαντήσεις, εάν υπάρχουν, σε κάποιες από τις ερωτήσεις αυτές, ώστε να κατανοήσουμε τη σπουδαιότητα του θέματος.

[2]Η *Τετρακτύς* ήταν η μουσική-αριθμητική τάξη που διέπει το Σύμπαν, σημαίνει «τετράδα» και αποτελείται από τους αριθμούς 1–10, ενώ το άθροισμα των αριθμών 1, 2, 3, 4 ισούται με 10 που είναι ο ιερός αριθμός των Δελφών. Για τους Πυθαγόρειους πάλι ο αριθμός 10 αποτελεί τον τέλειο αριθμό και η Τετρακτύς το ιερό σύμβολο τους, τόσο σημαντικό ώστε ο όρκος τους να είναι «ορκίζομαι σε αυτόν που παρέδωσε στην ψυχή μας την *τετρακτύν*» [45].

### 1.2.1 Η Μουσική και ο Ρόλος της

Η μουσική μάς περιβάλλει στην καθημερινότητά μας και όχι μόνο σε σημαντικές περιστάσεις. Με την ανακάλυψη των φορητών CD- και mp3-players είναι δυνατό να απολαμβάνουμε την αγαπημένη μας μουσική οπουδήποτε και οποιαδήποτε στιγμή της μέρας. Την ακούμε στα μέσα συγκοινωνίας, κάνοντας ψώνια, αλλά και στις κινηματογραφικές ταινίες και στην τηλεόραση, ενώ επιπλέον χρησιμοποιείται για λόγους μάρκετινγκ και θεραπείας. Χαρακτηριστικοί είναι οι λόγοι που αναφέρονται από τον Huron [64] για τη χρησιμότητα της μουσικής στον τομέα της διαφήμισης, οι οποίοι και είναι για διασκέδαση, για δημιουργία δομής και συνέχειας, για να ανακαλέσει αναμνήσεις και να στοχεύσει σε συγκεκριμένες κατηγορίες ανθρώπων. Επιπλέον χρησιμοποιείται για να μεταφέρει κάποιο μήνυμα μέσω των στίχων του τραγουδιού (*lyrical language*), κάτι το οποίο έχει τις ρίζες του στην αρχαία Ελλάδα – η ποίηση (συναισθηματικά φορτισμένη ομιλία) θεωρούνταν να έχει μεγαλύτερη δύναμη από την απλή ομιλία – ενώ τέλος έχει την ικανότητα να επιφέρει κύρος στο διαφημιζόμενο προϊόν.

Κατά συνέπεια, ο τρόπος αξιοποίησης και η αξία αυτή της μουσικής οδηγεί τον κάθε ενδιαφερόμενο, για τους δικούς του λόγους, να αναζητά διαφορετικά γνωρίσματα. Για παράδειγμα ο σκηνοθέτης προσπαθεί να βρει μουσική για την ανάδυση συγκεκριμένων συναισθημάτων, ο φυσιοθεραπευτής χρησιμοποιεί μουσική για να παρακινήσει τον ασθενή του, ο οδηγός προτιμά κάποιο άκουσμα που θα τον κρατήσει σε εγρήγορση, ο δάσκαλος μουσικής κάτι που θα κεντρίσει το ενδιαφέρον των μαθητών του ενώ ο απλός ακροατής επιδιώκει στη μουσική του ακρόαση κάτι καινούριο, ενδιαφέρον και ευχάριστο [64].

Πολλοί υποστήριξαν πως η μουσική παίζει σημαντικό ρόλο στην εξελικτική πορεία του ανθρώπου. Μεταξύ αυτών ο Κάρολος Δαρβίνος, ο οποίος θεωρούσε την ύπαρξη της μουσικής καθώς και τη μουσική δημιουργία εξελικτική συμπεριφορά του ανθρώπου χωρίς εμφανή προσαρμοστική αξία, ενώ ο Pinker ισχυρίζεται πως μας έχει βοηθήσει να επιβιώσουμε, καθώς έχουμε έμφυτη την προδιάθεση να είμαστε τόσο δημιουργοί όσο και λάτρεις της (όπως αναφέρεται στο [7]). Επιπλέον θεωρείται πως κάθε μουσική μορφή έχει γεννηθεί κάτω από συγκεκριμένες κοινωνικές και ιστορικές συνθήκες. Γι' αυτό και από την αρχαιότητα έως σήμερα, η μουσική δημιουργία, και μαζί της οι αισθητικές προτιμήσεις και τα κριτήρια αξιολόγησης ενός μουσικού έργου, υπόκεινται σε μια τεράστια ποικιλία μορφών αλλά και συχνά απρόβλεπτων μετατροπών [175]. Οι ισχυρισμοί αυτοί έχουν όντως νόημα αν σκεφτούμε την απλότητα της αρχέγονης μουσικής, την ύπαρξη των ελάχιστων μουσικών οργάνων, καθώς και των λίγων και απλών μουσικών ήχων. Η εξελικτική πορεία του ανθρώπινου πολιτισμού και κατά συνέπεια και του ανθρώπου, μαζί με τα προβλήματα αλλά και τα πιο πολύπλοκα συναισθήματά του, δημιούργησε πιθανώς την ανάγκη εύρεσης νέων τρόπων έκφρασης, και ταυτοχρόνως την εξέλιξη της μουσικής έκφρασης και τη

διαφορετικότητα της μορφής και της δομής της.

Αλλά τι είναι αυτό που όλοι λατρεύουμε τόσο; Πλήθος ορισμών έχει δοθεί στην πάροδο των ετών. Ένας από αυτούς, ο οποίος προσπαθεί να αναγνωρίσει την πολιτισμική και ιστορική ποικιλομορφία της μουσικής, προτείνεται από τον μουσικολόγο I. Cross (όπως αναφέρεται στο [7]):

> *Ως μουσικές μπορούν να οριστούν οι ανθρώπινες δραστηριότητες, ατομικές και κοινωνικές, που αφορούν τη δημιουργία και την αντίληψη του ήχου και δεν έχουν καμία προφανή και άμεση αποτελεσματικότητα ή σταθερή συναινετική αναφορά*[3].

Επίσης έχει χρησιμοποιηθεί ο όρος «*οργανωμένος ήχος*» που θεωρείται αρκετά εύστοχος [7]. Παρ' όλα αυτά, ο συγκεκριμένος ορισμός δεν έχει κανένα νόημα για μερικούς πολιτισμούς, εξαιρεί ορισμένα είδη μουσικής ενώ περιλαμβάνει ήχους που γενικά δεν θεωρούνται μουσική. Η φράση επινοήθηκε από τον Γάλλο συνθέτη της avant-garde μουσικής Edgar Varése[4], και χρησιμοποιήθηκε και από τον καθ. φιλοσοφίας Levinson (όπως αναφέρεται στο [13]), που θεωρεί τον ήχο οργανωμένο από τον άνθρωπο, και άρα αποκλείει το τραγούδι πουλιών και θηλαστικών, σκοπός της οποίας οργάνωσης είναι ο εμπλουτισμός και η εντατικοποίηση της μουσικής εμπειρίας και ακρόασης.

Έχει αποτελέσει μυστήριο και είναι αντικείμενο έρευνας έως και σήμερα, πώς ο συνδυασμός των τόσο βασικών δομικών στοιχείων του ήχου, όπως η συχνότητα και το πλάτος, γίνεται αντιληπτός από τον άνθρωπο και μορφοποιείται από απλό ήχο σε μουσική, η οποία μπορεί να αλλάξει τη διάθεσή του και να δημιουργήσει έντονες συναισθηματικές αντιδράσεις [51, 72, 148, 163, 183].

Η κατανόηση ενός μουσικού έργου, όσον αφορά τη σημασία και το περιεχόμενο, πάντα δυσκόλευε τον άνθρωπο πιθανώς λόγω αυτής της ιδιόμορφης φύσης του ηχητικού μέσου. Όπως αναφέρει ο Τερζάκης [175] μας φαίνεται εύκολο να μιλήσουμε για περιεχόμενο στην περίπτωση ενός ποιητικού κειμένου ή ενός πίνακα ζωγραφικής, όμως ο αυστηρά αφηρημένος χαρακτήρας του ήχου φαίνεται να αποτρέπει οποιαδήποτε μη τυπολογική ερμηνεία.

Κατά τη διάρκεια της μουσικής ακρόασης το μυαλό του ανθρώπου προβαίνει σε διάφορες αυτόματες και ακούσιες διεργασίες, όπως το φιλτράρισμα, η ταξινόμηση, η πρόβλεψη, οι οποίες οδηγούν σε φυσιολογικές αντιδράσεις του σώματος [149, 150]. Θεωρείται πως καμία άλλη ανθρώπινη δραστηριότητα δεν χρησιμοποιεί ούτε ενοποιεί

[3] Musics can be defined as those temporally patterned human activities, individual and social, that involve the production and perception of sound and have no evident and immediate efficacy or fixed consensual reference.

[4] Ο Varése δημιούργησε μουσική στις αρχές του εικοστού αιώνα, την οποία πολλοί σύγχρονοί του συνθέτες δεν τη δέχονταν ως «μουσική». Χρησιμοποιούσε δε την περιγραφή αυτή για να ξεχωρίσει τις τολμηρές ηχητικές του εξερευνήσεις από τη συμβατική μουσική της εποχής.

τόσα μέρη του μυαλού ταυτοχρόνως, γι΄ αυτό και η μουσική αναφέρεται ως γυμναστική του μυαλού [7]. Ίσως και να είναι τελικά όλες αυτές οι ακούσιες διεργασίες που μας οδηγούν στην πιο προσεκτική ανάλυση της δομής της μουσικής όπου και πετυχαίνουμε την κατανόηση και βλέπουμε εν τέλη την αλληλοσυσχέτιση των μερών της. Από την άλλη, ακόμα και το γεγονός πως μπορεί να είμαστε πολύ καλοί γνώστες μιας συγκεκριμένης μουσικής σύνθεσης και της δομής της και όντως να περιμένουμε να ακούσουμε κάποιο συγκεκριμένο γεγονός «έκπληξης», αυτό δεν μας αποτρέπει από το να την ευχαριστηθούμε εξίσου (*persistence of illusion*). Και αυτό λόγω του ότι η μουσική πολύ απλά μας απορροφάει, κάτι ακόμα που συνηγορεί στο ότι η μουσική ακρόαση αποτελεί μία διαδικασία αυτόματη και ακούσια [77].

Η γνωσιακή επιστήμη (*science of music cognition*) και ψυχολογία της μουσικής (*music psychology*) είναι πλέον οι επιστήμες που μελετούν τη μουσική βάσει των απλούστερων δομικών της στοιχείων, όπως το τονικό ύψος και ο ρυθμός, η αντίληψη και η οργάνωση των οποίων φαίνεται να αποτελεί ουσιαστικό μέρος της ικανότητας ενός ακροατή στο να μετατρέψει τον ήχο σε μουσική, ανεξαρτήτως της λειτουργίας που εξυπηρετεί.

Πολλοί είναι πάντως αυτοί που διαφωνούν κάθετα με αυτού του είδους τις μελέτες και τους ορισμούς, καθώς ασπάζονται μια πιο «ρομαντική αίσθηση», ότι δηλαδή η μουσική είναι προϊόν *υπερφυσικής έμπνευσης* και συνεπώς δεν μπορεί να περιγραφεί βάσει της «απλής» ακουστικής και της φυσικής του ήχου [7].

### 1.2.2 Από την Αρχαιότητα έως Σήμερα

Η ιστορία της μουσικής είναι μακραίωνη[5]. Η ίδια η λέξη θεωρείται εμπνευσμένη από τις *Μούσες* [7], που μιλούσαν με το στόμα των αοιδών και τους υπαγόρευαν τα τραγούδια. Οι διαφορετικοί πολιτισμοί κατά την αρχαιότητα και τον μεσαίωνα ερμηνεύουν τη μουσική με πολύ διαφορετικούς τρόπους. Κάποιες από τις τεκμηριωμένες κοινωνικές λειτουργίες της μουσικής είναι, μεταξύ άλλων, η πρόκληση ευχαρίστησης και η έκφραση συναισθημάτων, η συνοδεία του χορού, η επικύρωση τελετουργικών και μεγάλων γεγονότων της ζωής και η σύνδεσή της με τη θρησκεία και το υπερφυσικό, η προώθηση της κοινωνικής σταθερότητας, ενώ τέλος, θεωρείται ότι έχει κυρίως ηθικό και όχι αισθητικό χαρακτήρα.

Οι περισσότεροι προ-εγγράμματοι λαοί φαίνεται να αντιλαμβάνονται τη μελωδία και τους στίχους ενός τραγουδιού ως μια ενότητα, ενώ σε κάποιους πολιτισμούς η μουσική είναι άρρηκτα συνδεδεμένη με τον χορό. Στην αρχαία Ελλάδα η μουσική ήταν κυρίως φωνητική με τη συνοδεία της λύρας ή της κιθάρας. Σύμφωνα με τον Πλάτωνα η αποκλειστικά ενόργανη μουσική, χωρίς τη συνοδεία χορού ή τραγουδιού, ήταν «τραχιά

[5]Όλες οι ιδέες που παρατίθενται παρακάτω αναφέρονται στα βιβλία των J. Bicknell, *Why Music Moves Us*, Palgrave MacMillan, 2009 [13] και P. Ball,*The Music Instinct*, Oxford Univ. Press, 2010 [7]

και κακόγουστη», ενώ σύμφωνα με τον Αριστοτέλη η ποίηση χωρίς τη συνοδεία της λύρας (γνωστή ως «λυρική» όταν συνοδεύεται από τη λύρα, εξού και η προέλευση της αγγλικής λέξης *lyrics* για τους στίχους των τραγουδιών), δεν έχει συγκεκριμένο όνομα.

Σύμφωνα με τον Πλάτωνα και τον Αριστοτέλη η μουσική εκφράζει συναισθήματα μέσω της μίμησης των ανθρώπινων εκφράσεων και του λόγου, ο Αριστοτέλης συμπλήρωσε στην *Πολιτική* πως τα ανθρώπινα συναισθήματα εν συνεχεία επηρεάζονται από την αναπαράσταση αυτή της μουσικής [77]. Κατά τον 16ο αιώνα, οι αριστοκράτες της εποχής, οι αυτό-ονομαζόμενοι *Camerata*, ισχυρίστηκαν πως το συναίσθημα που εκφράζεται από τη μουσική αποδίδεται στην αναπαράσταση της μελωδίας· συγκεκριμένα θεωρούσαν πως για να κατορθώσει να προκαλέσει κάποια συναισθηματική αντίδραση, τότε η μελωδία πρέπει να είναι γραμμένη σύμφωνα με τον τρόπο έκφρασης του συναισθήματος στην ανθρώπινη ομιλία [77].

Γενικά, οι περισσότεροι αρχαίοι Έλληνες φιλόσοφοι έχουν αναφερθεί στις ιδιότητες της μουσικής. Για τον Πλάτωνα και τον Αριστοτέλη, η μουσική αποτελεί εργαλείο που θα μπορούσε να προωθήσει είτε την κοινωνική αρμονία (*harmony*) είτε, αν δεν χρησιμοποιηθεί σωστά, τη διχόνοια (*discord*) (λέξεις που σήμερα είναι και μουσικοί όροι). Ο Πυθαγόρας πάλι διατυπώνοντας τη θεωρία της «*αρμονίας των σφαιρών*», η οποία αναφέρεται στη σχέση μεταξύ των δομών της μουσικής και εκείνων του φυσικού κόσμου, προσπάθησε να συνδέσει την κοσμική αρμονία με τη μουσική αρμονία και να εξηγήσει τη θέση και την κίνηση των πλανητών χρησιμοποιώντας μουσικούς όρους.

Επίσης συνέδεαν τη μουσική με τα συναισθήματα, το ήθος αλλά και τη δράση (*action*). (Η μουσική θα ενίσχυε τη φυσικά καλή διάθεση της Κλυταιμνήστρας για να εξασφαλίσει την πίστη της στον σύζυγό της.) Ο Σωκράτης θεωρούσε πως η μουσική έχει πολύ σοβαρή επίδραση στα συναισθήματα και στον χαρακτήρα, και γι' αυτόν ακριβώς το λόγο θα μπορούσε να απαγορεύσει ορισμένους μουσικούς τρόπους από την ιδανική πόλη[6], ενώ στην *Πολιτεία* αναφέρει πως ο Δάμων, λαμπρός Αθηναίος μουσικοδιδάσκαλος και πολιτικός σύμβουλος του Περικλή, υποστήριζε πως οι αλλαγές στη μουσική συμβαδίζουν με τις πολιτικές αλλαγές [175]. Ο Αριστοτέλης από την άλλη, συζητάει στην *Πολιτική* για τον ρόλο της μουσικής εκπαίδευσης και της μουσικής σύνθεσης στη ζωή του «ελεύθερου ανθρώπου»[7], ενώ ήδη από την εποχή του Πυθαγόρα η αλληλοσυσχέτιση ψυχής-μουσικής φαινόταν να θεμελιώνει το ρόλο της μουσικής ως πρωταρχικού εκπαιδευτικού και

[6]Διάφοροι μύθοι για τους αρχαίους ελληνικούς τρόπους (*modes*) λένε πως ο Φρυγικός μπορεί να οδηγήσει κάποιον στην κατάσταση του θυμού, σε αντίθεση με τον Δώριο που επιφέρει την ηρεμία. Θεωρούνταν πως οι διαφορετικοί μουσικοί τρόποι είχαν μία συγκεκριμένη ποιότητα και έδιναν ένα συγκεκριμένο «ήθος» ή διάθεση. Η Φρυγία ήταν η πατρίδα του Διονύσου, και θρησκευτικές πρακτικές στη λατρεία του περιλάμβαναν τη δημιουργία εκστατικής κατάστασης μέσω της μουσικής και του χορού.

[7]Θεωρείται ότι η μουσική στην Κίνα δεν έχει εξελιχθεί, αλλά έχει παραμείνει ρυθμική και κρουστή, εξαιτίας των συνθηκών πνευματικής ανελευθερίας. Και αυτό ενώ στην αρχαία Κίνα η μουσική θεωρούνταν δώρο κάποιας ανώτερης δύναμης.

ηθικοπλαστικού μέσου [175].

Ο Πλάτωνας αναφέρει επίσης πως ο ήχος είναι η κίνηση (*percussion*) του αέρα που μεταδίδεται στην ψυχή μέσω της ακοής και του εγκεφάλου. Η κίνηση αυτή ξεκινάει από το κεφάλι και καταλήγει στο ήπαρ. Σύμφωνα με τους αρχαίους, το ήπαρ ήταν η έδρα της ζωτικότητας και των συναισθημάτων. Πίστευαν επίσης ότι η τέχνη, συμπεριλαμβανομένης της μουσικής, είναι μιμητική και μπορεί και αντιπροσωπεύει μη μουσικά φαινόμενα ή ακόμη και μη ακουστικά αντικείμενα και γεγονότα, καθώς επίσης πως μπορεί να μιμηθεί τις καταστάσεις της ψυχής. Ως εκ τούτου, ο ήχος επηρεάζει άμεσα τη σωματική και την ψυχική ευεξία. Λέγεται ότι ο Θαλής χρησιμοποίησε τη μουσική για να αντιμετωπίσει τη «μάστιγα» του άγχους που υφίσταντο οι Σπαρτιάτες, κάτι που σύμφωνα με τον Πλούταρχο κατέστη εφικτό με τη χρήση της αρμονίας, μαζί με τις μαγικές θεραπευτικές δυνάμεις που δόθηκαν στη μουσική του Ορφέα. Αλλά και για τους αρχαίους Αιγύπτιους η μουσική ήταν η φυσική της ψυχής (*physics for soul*), οι Εβραίοι τη χρησιμοποιούσαν για τη θεραπεία του σώματος και της ψυχής, ενώ σε καμία περίπτωση δεν πρέπει να ξεχάσουμε τον «μαγικό-μιμητικό» ρόλο της μουσικής των ινδιάνων, με χρήση μεταξύ άλλων στη θεραπεία, τον εξαγνισμό και την τελετουργία.

Τόσο ο Πυθαγόρας όσο και ο Πλάτωνας θεώρησαν πως η μουσική μπορεί να επιβάλλει την πειθαρχία καθώς και να αποτρέψει ή να παροτρύνει ανθρώπινες πράξεις και συμπεριφορές και άρα να χρησιμοποιηθεί ως «όργανο ελέγχου». Εντούτοις αργότερα ο Πλάτωνας στην *Πολιτεία* άρχισε να θέτει όρια εφόσον πίστευε πως κάποια είδη μουσικής ή μουσικών τρόπων αποτελούν κίνδυνο για τα πολιτικά ήθη καθώς επίσης διαισθανόταν μια βαθιά αναντιστοιχία ανάμεσα στη μουσική που έπαιζαν οι σύγχρονοι του και στο ιεραρχημένο σύμπαν του *Λόγου* που ο ίδιος πάσχιζε να ορθώσει [175].

Οι απόψεις των αρχαίων Ελλήνων επηρέασαν αργότερα τους χριστιανούς και τους ισλαμιστές στοχαστές. Συγγράμματα του Αγίου Αυγουστίνου για τη μουσική αποτελούν σε πολλές περιπτώσεις υπόδειγμα της χριστιανικής παράδοσης. Η μουσική περιβάλλεται από βαθιά αμφιθυμία, τόσο ως πηγή ομορφιάς και πνευματική ανύψωσης, όσο και ως επικίνδυνος πειρασμός και απομάκρυνση από τον Θεό. Από τη μια πλευρά, προσφέρει αισθησιακή απόλαυση που μπορεί να δελεάσει το μυαλό να απομακρυνθεί από τον Θεό άλλα από την άλλη, η ομορφιά της μπορεί να χρησιμοποιηθεί για να τονίσει τα ιερά λόγια και έτσι να οδηγήσει τους ακροατές πιο κοντά στον Θεό. Στην εποχή του Αγίου Αυγουστίνου οι θεατρικές παραστάσεις και η μουσική συνδέονταν στενά με την παγανιστική λατρεία, κάτι το οποίο θεωρείται πως δημιούργησε τη μεταγενέστερη υποψία του για τη μουσική. Όπως αναφέρεται στον Τερζάκη [175] από την εποχή που εδραιώθηκε ο Παπισμός υπήρξε η ανάγκη για κάθαρση της λατρευτικής μουσικής από τα επικίνδυνα ακούσματα της Μέσης Ανατολής, της Παλαιστίνης, της Φοινίκης, της Συρίας και των νησιών του Ανατολικού Αιγαίου. Γι΄ αυτό και τον 6ο αιώνα ο Πάπας Γρηγόριος ο

Μέγας πραγματοποίησε την πρώτη ουσιαστική νομοθεσία, σύμφωνα με την *πλατωνική απαγόρευση* με αποτέλεσμα την καθιέρωση της εκκλησιαστικής μουσικής γνωστής και ως Γρηγοριανό Μέλος.

Ο Γερμανός φιλόσοφος A. Schopenhauer θεωρούσε τη μουσική ως την πιο αξιοσημείωτη και σημαντική μεταξύ των καλών τεχνών λόγω του ότι έχει την ικανότητα να αντικατοπτρίσει τη βούληση του σύμπαντος και γι' αυτό και πολλές φορές την ονόμαζε «πιστό αντίγραφο» (*direct copy*) της βούλησης [77]. Ο σύγχρονος φιλόσοφος R. Scruton (στο [7]) υποστηρίζει ότι η μουσική ακόμα και σήμερα έχει την ικανότητα να προσφέρει ηθική εκπαίδευση: «*Μέσω της μελωδίας, της αρμονίας και του ρυθμού, εισερχόμαστε σε έναν κόσμο συνύπαρξης με τους άλλους, έναν κόσμο γεμάτο συναίσθημα, έναν κόσμο δομημένο και πειθαρχημένο αλλά ελεύθερο*». Γι' αυτό και θεωρεί πως η μουσική έχει τη δύναμη να δομήσει τον χαρακτήρα του ανθρώπου[8].

Στη νεότερη δυτική παράδοση η αντιμετώπιση της μουσικής ως φορέα πνευματικής γνώσης συμβαδίζει με την αντίληψη ότι ο συνθέτης είναι δημιουργική ιδιοφυΐα και ο δεξιοτέχνης ερμηνευτής θεϊκά (ή διαβολικά) εμπνευσμένος. Ο βιολιστής Paganini και ο θρυλικός μπλουζ κιθαρίστας R. Johnson είναι δύο χαρακτηριστικά παραδείγματα μουσικών που φημολογούνταν ότι όφειλαν την καταπληκτική τεχνική τους σε συμφωνία με τον διάβολο.

Υπάρχει φυσικά και η άποψη, αρκετά δημοφιλής παλαιότερα σε μερίδα μαρξιστών ιστορικών, η οποία θέλει το τραγούδι και τη μουσική να είναι απόρροια των εργασιακών δραστηριοτήτων [175]. Με άλλα λόγια, η ομαδική φύση της εργασίας και η ανάγκη για συντονισμό των εργατών οδήγησε στη δημιουργία ρυθμικών και επαναλαμβανόμενων μοτίβων για την εμψύχωση και όχι μόνο του ανθρώπινου δυναμικού. Τέτοιου είδους μουσική τη βρίσκουμε στα τραγούδια των «μαύρων» του Νότου, ενώ αντίστοιχα υπάρχει και η μουσική των λευκών αγροτών και μεταναστών της Αμερικής (*Hillbilly*), το καθαρά πολιτικό τραγούδι των εργατικών συνδικάτων (*Woblly*), το οποίο μετεξελίσσεται στο «τραγούδι της διαμαρτυρίας» [175]. Παρόμοιες τάσεις σχετικά με τον ρόλο και τις συνθήκες δημιουργίας των διαφορετικών ειδών μουσικής βρίσκουμε σε ολόκληρη την υφήλιο.

Η μελέτη της μουσικής άλλων πολιτισμών επισημαίνει πως η συναισθηματική αντίδραση κάθε ανθρώπου στη μουσική βρίσκεται λίγο πολύ σε συνάφεια με τους κανόνες και τις προσδοκίες του πολιτισμού του. Για παράδειγμα, η μουσική του Μπαλί είναι χρηστική (*utilitarian*) και δεν προσφέρει καμιά συγκίνηση, ενώ σε ορισμένα μέρη της Αφρικής θεωρείται πως μουσική χωρίς σταθερό ρυθμό, με σκοπό τον χορό, δεν είναι μουσική αλλά μορφή θρήνου (στο [7]).

[8]Through melody, harmony, and rhythm, we enter a world where others exist besides the self, a world that is full of feeling but also ordered, disciplined but free. That is why music is a character-forming force.

### 1.2.3 Επίδραση της Μουσικής στον Ανθρώπινο Εγκέφαλο

Ο ήχος, άρα και η μουσική, έχει συνεισφέρει στην εξέλιξη και στην κοινωνικοποίηση του ανθρώπου. Εκτός όμως από αυτήν τη λειτουργία της μουσικής, είναι λογικό να διερωτηθεί κανείς ποιες διεργασίες γίνονται στον ανθρώπινο εγκέφαλο. Πλήθος μελετών έχουν καταδείξει τις επιπτώσεις της μουσικής στον ανθρώπινο οργανισμό [74, 95, 149], καθώς και τις φυσιολογικές αντιδράσεις του, προφανώς διαφορετικές από τις καθαρά γνωσιακές. Για παράδειγμα, μπορεί να επηρεάσει το ανοσοποιητικό σύστημα, αυξάνοντας τα επίπεδα των πρωτεϊνών τα οποία συγκρούονται με τις διάφορες μικροβιακές μολύνσεις. Τόσο η εκτέλεση όσο και η ακρόαση της μουσικής μπορεί να ρυθμίσει την παραγωγή ορμονών που επηρεάζουν τη διάθεση, όπως για παράδειγμα η κορτιζόλη, κάτι το οποίο δείχνει ότι υπάρχει σωστή βιοχημική βάση για τη χρήση της μουσικής στη θεραπεία [7]. Επιπλέον μπορεί να επιφέρει δυσκολία στην αναπνοή, αύξηση του καρδιακού ρυθμού, τρόμο, ανατριχίλα, πόνο στο στήθος ή στο στομάχι και (σπάνια) απώλεια συνείδησης ή αλλαγή της θερμοκρασίας του σώματος [13, 74]. Για τους λόγους αυτούς η μουσική έχει χρησιμοποιηθεί εκτενώς στη θεραπεία και στην αποκατάσταση σωματικών και νοητικών βλαβών (για ανασκόπηση πολλών περιπτώσεων βλ. [145]).

Οι νέες τεχνολογίες, όπως η μαγνητική τομογραφία (MRI), επιτρέπουν πλέον στους νευροεπιστήμονες να δουν την ακριβή λειτουργία του εγκεφάλου κατά τη μουσική επεξεργασία και να διερευνήσουν αν υπάρχει επικάλυψη με άλλες νοητικές λειτουργίες [78, 127]. Οι περισσότερες νοητικές διεργασίες, όπως η όραση ή η ομιλία, έχουν αρκετά καλά εντοπισμένα κέντρα ενεργοποίησης στον ανθρώπινο εγκέφαλο. Η μουσική, από την άλλη, θεωρείται πως μπορεί να ενεργοποιήσει το σύνολό του, για παράδειγμα το κέντρο που είναι υπεύθυνο για την κίνηση, το κέντρο του συναισθήματος, τα κέντρα της ομιλίας τα οποία επεξεργάζονται τη σύνταξη και τη σημασιολογία. Σε αντίθεση με την ομιλία, η επεξεργασία της μουσικής δεν εντοπίζεται σε μία ή σε λίγες συγκεκριμένες περιοχές, αλλά «αφορά ολόκληρο τον εγκέφαλο» (*whole brain phenomenon*) [7].

Αυτός είναι ακόμα ένας λόγος που συνηγορεί στη σπουδαιότητα της μουσικής, καθώς κανένα άλλο ερέθισμα δεν μπορεί να συνδυάσει τόσο πολλές διαφορετικές διεργασίες του εγκεφάλου και να τις αναγκάσει να συνομιλούν μεταξύ τους και να αλληλοσυμπληρώνονται – το αριστερό με το δεξί ημισφαίριο, τη λογική με το συναίσθημα. Η αντίληψη του τονικού ύψους, για παράδειγμα, φαίνεται ότι εντοπίζεται ως επί το πλείστον (αλλά όχι αποκλειστικά) στο δεξί ημισφαίριο, παρ' όλα αυτά το αριστερό ημισφαίριο παίζει βασικό ρόλο στις πολύπλοκες γλωσσολογικές διεργασίες για την κατανόηση των στίχων των τραγουδιών [194]. Επίσης, ενώ οι διεργασίες του αριστερού ημισφαιρίου φαίνεται να κυριαρχούν στα θετικά συναισθήματα, το δεξί ημισφαίριο εμπλέκεται στα αρνητικά [126]. Ως εκ τούτου η πλήρης εικόνα των διεργασιών του εγκεφάλου σε σχέση με τη μουσική είναι περίπλοκη, με αποτέλεσμα και τα δύο

ημισφαίρια να λειτουργούν κατά την επεξεργασία του μουσικού σήματος.

### 1.2.4 Η Δομή της Μουσικής και η Σχέση της με την Ομιλία

Η μουσική όπως και η ομιλία αποτελούν επιτεύγματα του ανθρώπου, μέρος της εξελικτικής πορείας του. Η συσχέτιση των δύο θεωρείται ιδιαίτερα περίπλοκη· θα μπορούσε κάποιος να πει πως ενώ σχετίζονται άμεσα διαφέρουν εξίσου πολύ. Χρησιμοποιούν αλληλουχίες από διακριτά δομικά στοιχεία, η ομιλία αποτελείται από φωνήματα και η μουσική από νότες, τα οποία από μόνα τους δεν έχουν κάποιο ιδιαίτερο νόημα. Τα βασικά αυτά στοιχεία οργανώνονται σε δομημένες ακολουθίες (για τη δημιουργία λέξεων και μελωδιών), και έπειτα σε ακόμα πιο πολύπλοκες ιεραρχικές ακολουθίες για τη δημιουργία προτάσεων και τραγουδιών [85], που διαμορφώνονται κατάλληλα για τη μεταφορά μηνυμάτων και συναισθημάτων [21].

Σε αντίθεση με την ομιλία, στη μουσική λείπει το επίπεδο της λέξης μιας και ο συνδυασμός των δομικών της στοιχείων οδηγεί κατευθείαν στο επίπεδο της «πρότασης», δηλαδή της μελωδικής φράσης. Ο ανθρωπολόγος Lèvi-Strauss [86] πάει ένα βήμα παρακάτω σε αυτή τη σύγκριση μεταξύ ομιλίας και μουσικής και αναφέρεται στο μύθο, όπου σε αυτή τη περίπτωση η βασική του μονάδα είναι οι λέξεις οι οποίες συνδυάζονται για τη δημιουργία προτάσεων. Μας υποδεικνύει πως ενώ η ομιλία έχει τρία διακριτά επίπεδα για τη δημιουργία του νοήματος, η μουσική και ο μύθος έχουν μόνο δύο. Ο συγγραφέας διαπιστώνει πως για να μπορέσουμε να κατανοήσουμε τη σχέση της ομιλίας, του μύθου (γραπτού κειμένου) και της μουσικής πρέπει να θεωρήσουμε την ομιλία ως αφετηρία. Τότε θα αντιληφθούμε ότι η μουσική και ο μύθος είναι «σαν δυο αδέλφια καρπός της ομιλίας» αλλά με διαφορετική εξελικτική πορεία, αφού στη μουσική δίνεται έμφαση στη διάσταση του ήχου ενώ στο μύθο δίνεται έμφαση στη νόηση και το νόημα (διαστάσεις ενσωματωμένες και προερχόμενες από την ομιλία).

Παράλληλα, η ύπαρξη θεωριών, που υποστηρίζουν την παρουσία κοινών προγόνων καθώς και την παράλληλη εξέλιξή τους [20], έχει οδηγήσει τα τελευταία χρόνια σε μια σειρά μελετών που επιχειρηματολογεί υπέρ της επικάλυψης των διαδικασιών του εγκεφάλου κατά την επεξεργασία της μουσικής και της γλωσσολογικής δομής (π.χ. συγχορδίες και λέξεις) [122, 124, 162]. Η επικάλυψη των γνωσιακών μηχανισμών που επεξεργάζονται τη συντακτική δομή στη μουσική και στη γλώσσα μερικές φορές λαμβάνεται άρα ως αποδεικτικό στοιχείο εναντίον της εξειδίκευσης κάθε τομέα. Όλες αυτές οι ενδείξεις για τον παρεμφερή τρόπο δημιουργίας της μουσικής και της ομιλίας, την αλληλεξάρτηση και την κοινή πορεία αποτελούν τον κύριο λόγο της παράλληλης μελέτης τους σε σχέση με τους μηχανισμούς του εγκεφάλου.

Ωστόσο, η παρουσία των επικαλυπτόμενων διεργασιών δεν αποτελεί απόδειξη της μη ύπαρξης επιπρόσθετων εξειδικευμένων μηχανισμών [21]. Μελέτες σε ασθενείς με

εγκεφαλική βλάβη υποδηλώνουν τουλάχιστον σε κάποιο βαθμό, ανεξαρτησία μεταξύ των δύο περιοχών όπου λαμβάνει χώρα η γλωσσική και η μουσική επεξεργασία [127, 128]. Η πιο βασική διαφορά της ομιλίας από τη μουσική είναι ο τρόπος παραγωγής τους. Η ομιλία παράγεται από ένα όργανο, τη «φωνή», ενώ η μουσική, στη πιο συνηθισμένη της μορφή, από πολλά διαφορετικά όργανα, συμπεριλαμβανομένης της φωνής. Η ομιλία όντως περιλαμβάνει μεγάλη ποικιλία πολύπλοκων συνιστωσών, όπως περιοδικές και μη περιοδικές συνιστώσες, θόρυβο, διαμορφώσεις συχνότητας και πλάτους κ.ά., αντικατοπτρίζοντας έτσι τις δυνατότητες άρθρωσης του ανθρώπινου φωνητικού συστήματος. Από την άλλη, ο χαρακτηρισμός των θεμελιωδών ακουστικών χαρακτηριστικών της μουσικής θεωρείται εξαιρετικά πιο δύσκολος, καθώς ακουστικά είναι πολύ πιο ποικιλόμορφη. Κάποια από αυτά τα χαρακτηριστικά είναι η ένταση, ο ρυθμός, η διάρκεια, το μέσο παραγωγής με τα δύο βασικότερα το τονικό ύψος ή ακουστική συχνότητα (*pitch*) και τη χροιά. Το τονικό ύψος είναι το πιο σύνηθες χαρακτηριστικό για τη δημιουργία οργανωμένων μουσικών συστημάτων και αποτελεί τη βάση για τον σχηματισμό των διαφορετικών ηχητικών κατηγοριών (όπως τα διαστήματα και οι συγχορδίες). Το ηχόχρωμα από την άλλη είναι εξίσου σημαντικό μιας και αποτελεί το χαρακτηριστικό που προσδίδει το συγκεκριμένο άκουσμα στα διαφορετικά μουσικά όργανα· και φυσικά αποτελεί το βασικό στοιχείο στη φωνή (π.χ. φωνήεντα και σύμφωνα).

Με τη βοήθεια όλων αυτών των χαρακτηριστικών της μουσικής από εκεί και στο εξής οι δυνατότητες παρουσιάζονται σχεδόν άπειρες. Ο κατάλληλος συνδυασμός των χαρακτηριστικών της οδηγεί σε μεγαλύτερα δομικά στοιχεία όπως το μοτίβο, η φράση και η περίοδος για να καταλήξει στη διαμόρφωση της ολοκληρωμένης μουσικής σύνθεσης όπως η φούγκα, το μαδριγάλι, το κοράλ και το ροντό έως τη rock, blues, pop ή metal μουσική κ.ά.

Ολοκληρώνοντας αξίζει να αναφέρουμε, πως ο άνθρωπος γεννιέται σε έναν κόσμο όπου προϋπάρχουν τα δύο διαφορετικά ηχητικά συστήματα, η ομιλία και η μουσική. Τα βρέφη εξελίσσονται σε ενήλικες με έμφυτη γνώση της μητρικής τους γλώσσας και ικανότητα τέρψης από τη μουσική κουλτούρα των γύρω τους. Ωστόσο η δεξιότητα αυτή που εξελίσσει ο άνθρωπος έχει ως συνέπεια τη δημιουργία δυσκολιών που αφορούν την ακρόαση και την παραγωγή διαφορετικών ήχων (με άλλα λόγια, δυσκολία στο να αναπαράγει φωνήματα που δεν συμπεριλαμβάνονται στη μητρική του γλώσσα αλλά και να ευχαριστηθεί μουσική διαφορετικών πολιτισμών). Αυτό, σύμφωνα με τον Patel [123], συμβαίνει για τον απλό λόγο πως το εγγενές ηχητικό περιβάλλον αποτυπώνεται στο μυαλό του ανθρώπου συμβάλλοντας στην προσαρμοστική του ικανότητα σε αυτό, αλλά μπορεί να οδηγήσει σε λάθη και δυσκολίες όταν ακούμε τη γλώσσα ή τη μουσική ενός άλλου πολιτισμού, μιας και όπως αναφέρει «ακούμε με προφορά».

## 1.3 Εντέλει γιατί ακούμε μουσική;

*"Why the keening sounds from Mississippi should strike notes of thrill*
*and terror and wonder in hearts in the suburbs of London,*
*I don't know. It can only be because it goes beyond colour, blood*
*– it goes to the bone. Maybe that's it.*
*If you look closely at the marrow, there's a little bit of blue in there."* [9]

**Keith Richards**

Γιατί τελικά ακούμε μουσική; Οι λόγοι είναι πολλοί και διάφοροι - αισθητικοί, συναισθηματικοί, διανοητικοί, πρακτικοί, ακόμα και ηθικοί. Στο βιβλίο της η Bicknell [13] διερωτάται «Γιατί ο Οδυσσέας κλαίει όταν ακούει το τραγούδι του Δημόδοκου;» και περιγράφει τη σκηνή όπου ο Οδυσσέας ξεσπάει σε κλάματα ακούγοντας ένα τραγούδι του βάρδου. Η συγγραφέας καταλήγει στο συμπέρασμα πως η έντονη αντίδραση του Οδυσσέα οφείλεται στο γεγονός πως η μουσική έχει σημασία, αγγίζει το σώμα αλλά και το μυαλό του, τον συνδέει με το παρελθόν και του θυμίζει σημαντικά γεγονότα της ζωής του.

Στο κεφάλαιο αυτό, είδαμε πως η μουσική έχει συνεισφέρει στην εξέλιξή μας, η μουσική εμπειρία αποτελεί κοινωνική μας δραστηριότητα από τη βρεφική ηλικία, ενώ για ολόκληρη τη ζωή μας αποτελεί μέρος της κοινωνικής μας φύσης. Επιπλέον έχει βαθιές ρίζες στην ανατροφή αλλά και στην κοινωνικοποίησή μας, τόσο γνωσιακή όσο και συναισθηματική. Τη μουσική την ακούμε, τη δημιουργούμε, τη συζητάμε, τη θυμόμαστε, πολλές φορές τη βλέπουμε κιόλας. Σε αυτή τη μελέτη προσπαθούμε να κάνουμε ένα βήμα παρακάτω και να την αναλύσουμε, να την αποδομήσουμε και να τη ξαναδομήσουμε.

[9] Anthony DeCurtis, *Keith Richards: The Rolling Stone Blues Today*, Archives of RollingStone Magazine, May 28, 1998.

# Κεφάλαιο 2

# Επισκόπηση Ερευνητικών Περιοχών

Τα τελευταία χρόνια έχει σημειωθεί ραγδαία αύξηση των ψηφιακών δεδομένων τόσο στο Διαδίκτυο όσο και στους προσωπικούς υπολογιστές, τα οποία δημιουργούνται από τους ίδιους τους χρήστες. Η αύξηση αυτή οφείλεται στην ψηφιοποίηση των ακουστικών και οπτικών μέσων (μουσική και βίντεο) αλλά και στην πληθώρα των ψηφιακών συσκευών που διατίθενται στην αγορά. Το μέγεθος του όγκου δεδομένων φαίνεται επίσης από τη δημιουργία ιστοτόπων όπως το youtube, το facebook, ο Last.fm, το StereoMood κ.ά. Αν και το πλήθος των δεδομένων προσφέρεται προς ευχαρίστηση του χρήστη, δεν του εξασφαλίζει απαραίτητα μια ευχάριστη εμπειρία, ώστε να μπορεί να εξερευνά και να μοιράζεται όλο το ψηφιακό περιεχόμενο που έχει στη διάθεση του. Συνεπώς, ο μεγάλος αριθμός των δεδομένων δημιουργεί την ανάγκη επεξεργασίας του περιεχομένου τους, την ανάλυση, την αναγνώριση και την κατηγοριοποίηση, τη μεταγραφή, τη δημιουργία ευρετηρίων ηχητικών και πολυμεσικών βάσεων δεδομένων, αλλά και την ανίχνευση των σημαντικών τμημάτων τους για τη δημιουργία μικρότερων περιγραφικών αποσπασμάτων και περιλήψεων.

Συγκεκριμένα, χρήσιμες μουσικές εφαρμογές είναι η κατηγοριοποίηση μουσικών οργάνων [8, 12], η κατηγοριοποίηση της μουσικής βάσει του είδους [9, 11, 62, 120, 179] ή των εκφραστικών ιδιοτήτων της [23], η μουσική μεταγραφή [136], η εύρεση της μουσικής σε ηχητικές ροές δεδομένων, η ανίχνευση σημαντικών γεγονότων ή σημείων έκπληξης καθώς και η δημιουργία μουσικών περιλήψεων (*music thumbnails*) [94, 129, 200]. Οι παραπάνω εφαρμογές απαιτούν λύσεις σε προβλήματα επεξεργασίας της πληροφορίας (*information processing*), και συγκεκριμένα στην ανάπτυξη αποτελεσματικών ψηφιακών μεθόδων επεξεργασίας του μουσικού σήματος για την ανάλυση της δομής του. Επιπλέον, βασικός στόχος είναι η παραγωγή περιγραφικών αναπαραστάσεων στο επίπεδο του σήματος, που θα αποτελέσουν το βασικό εργαλείο για την ανάπτυξη των μεθόδων αυτών.

## 2.1 Μουσικά Όργανα

Η αυτόματη κατηγοριοποίηση, λόγω της πολυπλοκότητας και της ποικιλομορφίας της μουσικής, αποτελεί πρόκληση. Το θέμα της αναγνώρισης μουσικών οργάνων, παρουσιάζει ιδιαίτερο ενδιαφέρον και αποτελεί ερευνητικό πεδίο της τελευταίας δεκαπενταετίας. Η πλειοψηφία των μελετών χειρίζεται το πρόβλημα χρησιμοποιώντας ως πηγή ήχου μεμονωμένες νότες [1, 35, 36], ενώ λιγότερες μελέτες προσεγγίζουν το θέμα χρησιμοποιώντας μουσικές φράσεις από πραγματικές ηχογραφήσεις [19, 38, 92]. Ένα από τα πλεονεκτήματα της ανάλυσης μεμονωμένων μουσικών τόνων είναι η απλοποίηση του προβλήματος όσον αφορά την επεξεργασία του σήματος και την εξαγωγή χαρακτηριστικών. Αποτέλεσμα αυτού είναι η δυνατότητα διερεύνησης νέων, πιο περίπλοκων και με περισσότερες δυνατότητες αναπαραστάσεων των μουσικών σημάτων. Επιπρόσθετα, υπάρχουν αρκετές βάσεις δεδομένων μουσικών ήχων [53, 117, 181], οι οποίες έχουν δημιουργηθεί με τρόπο συστηματικό και άρα μπορούν να χρησιμοποιηθούν για τη διεξαγωγή έγκυρων μελετών.

Στη διδακτορική αυτή διατριβή, ασχολούμαστε με το θέμα της αναγνώρισης μεμονωμένων μουσικών τόνων στα πλαίσια των μικροδομών των ήχων αυτών, εξάγοντας πιο πολύπλοκα χαρακτηριστικά βασισμένα στη θεωρία των διαμορφώσεων (*modulations*) και των φράκταλ (*fractal*). Όπως θα δείξουμε σε επόμενα κεφάλαια, χαρακτηριστικά που βασίζονται στις μεθοδολογίες αυτές εμπεριέχουν πολύτιμη πληροφορία για τις εφαρμογές ανάλυσης και κατηγοριοποίησης που εξετάζουμε. Στη συνέχεια ορίζουμε το ηχόχρωμα, βασική έννοια για τη διάκριση των μουσικών οργάνων, και εξετάζουμε κάποια από τα χαρακτηριστικά των μουσικών σημάτων, με στόχο την κατανόηση και τη μελέτη της δομής τους με τον καλύτερο δυνατό τρόπο.

### 2.1.1 Ηχόχρωμα

Οι άνθρωποι έχουν διαρκώς την τάση να ταξινομούν τον κόσμο γύρω τους, ανάγκη η οποία πιθανώς οφείλεται στο γεγονός πως ο ανθρώπινος νους είναι μέρος του κόσμου (σύμπαντος), ο οποίος έχει συγκεκριμένη δομή και δεν είναι χαώδης [86]. Ο ήχος στην περίπτωση αυτή φυσικά και δεν αποτελεί εξαίρεση. Προσπαθούμε να συλλάβουμε κάθε μεμονωμένο ήχο, να τον συνδέσουμε με τα χαρακτηριστικά του και να τον κατηγοριοποιήσουμε σύμφωνα με διάφορες πτυχές, όπως φυσικός ή τεχνητός, μεταβαλλόμενος ή σταθερός, ή σύμφωνα με το μέσο παραγωγής του. Το τελευταίο, που είναι ίσως το πιο σημαντικό ισχύει και για τα μουσικά όργανα, τα οποία ταξινομούνται σε διάφορες οικογένειες, ανάλογα με τον τρόπο κατασκευής τους (σχήμα, υλικό) και τις φυσικές τους ιδιότητες. Οι τέσσερις κύριες κατηγορίες ή οικογένειες μουσικών οργάνων είναι: τα έγχορδα (π.χ., βιολί, κοντραμπάσο), τα ξύλινα πνευστά (π.χ., κλαρινέτο,

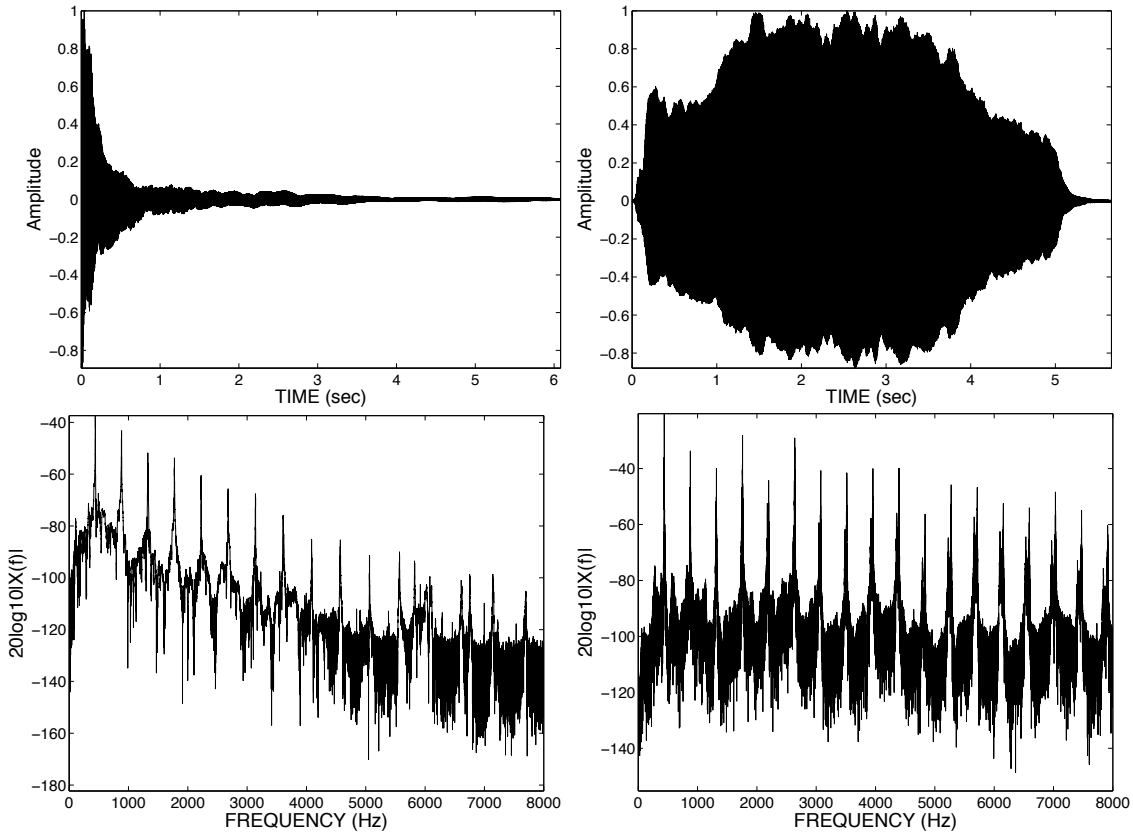
φαγκότο, σαξόφωνο, όμποε), τα χάλκινα πνευστά (π.χ., κόρνο, τούμπα, τρομπέτα, τρομπόνι) και τα κρουστά (π.χ., πιάνο). Το κύριο χαρακτηριστικό που ξεχωρίζει τα μουσικά όργανα μεταξύ τους είναι το ηχόχρωμα (*timbre*) ή αλλιώς χροιά[1]. Το ηχόχρωμα αποτελεί αναμφισβήτητα το πιο προσωπικό χαρακτηριστικό της μουσικής και είναι το κλειδί για τις προτιμήσεις μας ειδικά όσον αφορά το τραγούδι. Γνωστοί μουσικοί όπως η Billie Holiday, ο Frank Sinatra, ο Bob Dylan, ή ο Tom Waits έχουν άμεσα αναγνωρίσιμη φωνητική χροιά, η οποία είναι χαρακτηριστική της φωνής τους. Το ίδιο ισχύει και για τα μουσικά όργανα. Το πραγματικό μυστήριο του ηχοχρώματος είναι πως παρά το γεγονός ότι πρόκειται για κάτι το τόσο αόριστο και ασαφές, το ανθρώπινο μυαλό καταφέρνει να εναρμονιστεί απόλυτα με τις αποχρώσεις του [7].

Ο προσδιορισμός του ηχοχρώματος μέσω της κυματομορφής αποτελεί μια από τις κύριες σχέσεις των χαρακτηριστικών του ήχου. Η σχέση αυτή παρουσιάζει ιδιαίτερες δυσκολίες στην περιγραφή της (σε αντίθεση, για παράδειγμα, με την ένταση (*loudness*) και το τονικό ύψος (*pitch*), καθώς το ηχόχρωμα αλλά και η κυματομορφή είναι δύο ιδιαίτερα πολύπλοκες ποσότητες. Όλοι οι σύνθετοι ήχοι, όπως οι ήχοι των μουσικών οργάνων, αποτελούν ένα συνδυασμό διαφορετικών συχνοτήτων ακέραιων πολλαπλάσιων της θεμελιώδους συχνότητας $f_0$ (π.χ., $f_0$, $2f_0$, $3f_0$, $4f_0$ και ούτω καθεξής). Αυτή η ιδιότητα των μουσικών ήχων αναφέρεται ως «αρμονικότητα» (*harmonicity*) και οι ξεχωριστές συχνότητες ως «αρμονικές» (*harmonics*). Στην Εικόνα 2.1 βλέπουμε τη νότα A4 για το πιάνο και το βιολί στον χρόνο και τη συχνότητα. Στις εικόνες της δεύτερης σειράς φαίνονται καθαρά οι αρμονικές της θεμελιώδους συχνότητας $f_0$.

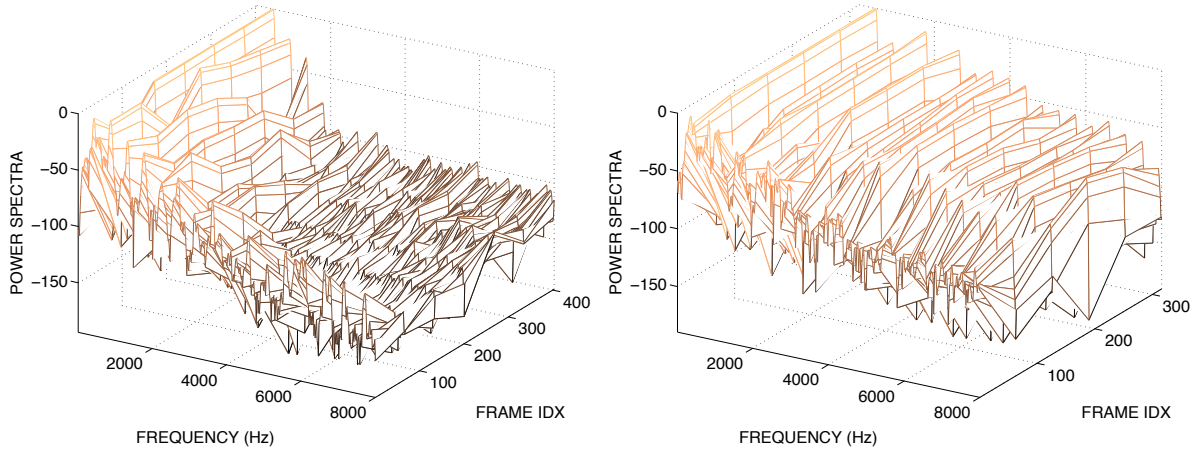
Το ηχόχρωμα, σύμφωνα με τον ASA (American Standards Association) [2], είναι το χαρακτηριστικό που ξεχωρίζει δύο ήχους της ίδιας τονικότητας, έντασης και διάρκειας, και άρα σχετίζεται με την κατηγοριοποίηση των περιβαλλοντικών πηγών ήχου[2]. Σύμφωνα με ένα λιγότερο αυστηρό ορισμό, το ηχόχρωμα προσδιορίζεται από τον αριθμό και τη σχετική θέση και ένταση (*shaping*) των αρμονικών του οργάνου, δηλαδή την κατανομή του πλάτους τους [134], ως συνέπεια των συντονισμών του οργάνου. Ο Fletcher [46] έδειξε ότι αυτή η αναλογία δεν είναι τόσο απλή, καθώς το ηχόχρωμα εξαρτάται εξίσου από τη θεμελιώδη συχνότητα και από την ένταση του ήχου. Εν κατακλείδι, το ηχόχρωμα εξαρτάται από τις απόλυτες συχνότητες αλλά και από τα σχετικά πλάτη των αρμονικών ενός τόνου τα οποία ποικίλλουν ανάλογα με το μουσικό όργανο. Η ανομοιομορφία αυτή των αρμονικών χαρακτηρίζει τον ήχο των οργάνων θαμπό ή γλυκό, όταν υπάρχουν δυνατές χαμηλότερες αρμονικές, ή οξύ και διαπεραστικό, όταν υπάρχουν δυνατές υψηλότερες αρμονικές. Στο Σχήμα 2.2 βλέπουμε τη νότα A4 για το πιάνο και το βιολί σε τρισδιάστατη απεικόνιση στον χρόνο και στη συχνότητα, όπου και φαίνεται η εξέλιξη των αρμονικών,

[1] Στην αγγλική γλώσσα το ηχόχρωμα ή timbre αναφέρεται και ως tone color ή tone quality.

[2] *Timbre is the quality of sound which distinguishes two sounds of the same pitch, loudness and duration and is thus associated with the identification of environmental sound sources.*



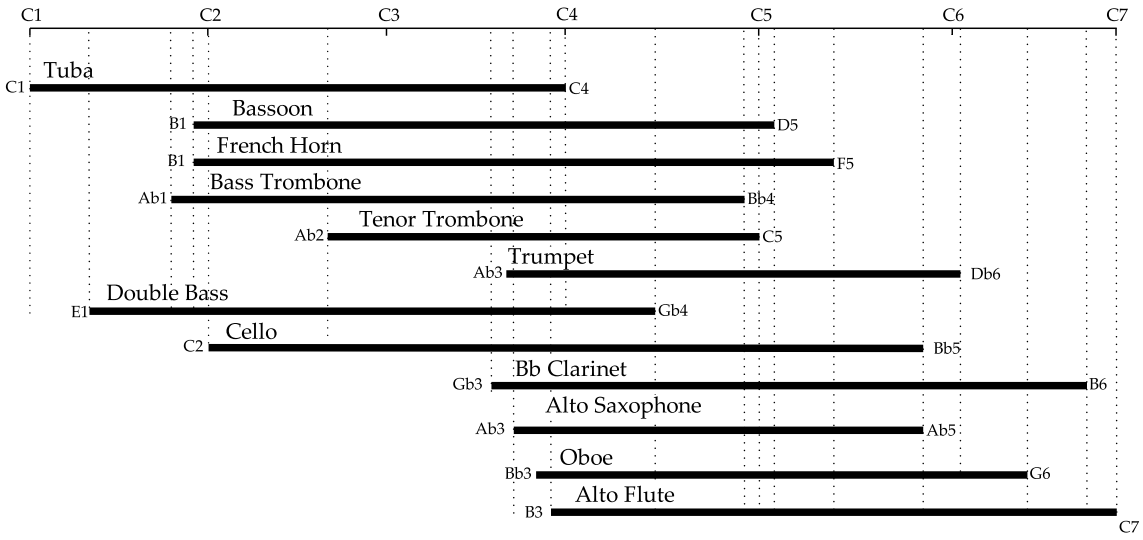

*Σχήμα 2.1:* Η κυματομορφή της νότας A4 για το πιάνο (αριστερά) και το βιολί (δεξιά) στον χρόνο (σειρά 1), και το φάσμα τους (σειρά 2).




*Σχήμα 2.2:* Τρισδιάστατη απεικόνιση στον χρόνο και τη συχνότητα για τη νότα A4 για το πιάνο (αριστερά) και το βιολί (δεξιά), όπου φαίνεται η εξέλιξη των αρμονικών, δηλαδή η θέση και το πλάτος τους στο χρόνο.

δηλαδή τόσο η θέση όσο και το πλάτος τους στο χρόνο.

Ο Bregman [18] αναφέρει για το ηχόχρωμα: «Αν λάβουμε υπόψη μας κάθε στιγμή της κάθε συνιστώσας της συχνότητας, θα συνειδητοποιήσουμε πως οι ήχοι διαφέρουν ακουστικά μεταξύ τους με πάρα πολλούς τρόπους». Και διερωτάται «μήπως το ανθρώπινο




*Σχήμα 2.3:* Οι συχνότητες των μουσικών οργάνων που αφορούν την συγκεκριμένη ανάλυση και η επικάλυψη του εύρους τους.

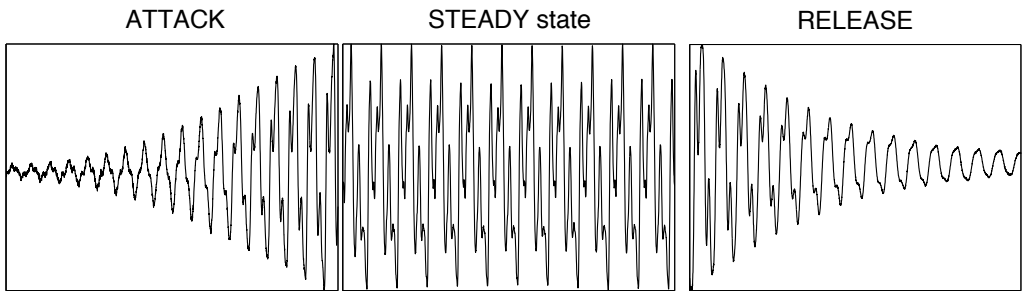
ακουστικό σύστημα αντεπεξέρχεται σε αυτή την πολυπλοκότητα χρησιμοποιώντας μόνο ένα μικρό αριθμό διαστάσεων;»[3] Και η αλήθεια είναι πως το ανθρώπινο ακουστικό σύστημα ερμηνεύει το ηχόχρωμα με το δικό του υποκειμενικό τρόπο. Όταν αναπαράγεται ένας ήχος στα 200 Hz (με αρμονικές στα 400 Hz, στα 800 Hz και ούτω καθεξής), το αυτί αντιλαμβάνεται τη συχνότητα των 200 Hz ως το τονικό ύψος, ενώ τις διαφορές στα πλάτη των αρμονικών τις συνδέει με το ηχόχρωμα. Άρα αντιλαμβάνεται τον συνδυασμό των στατικών και δυναμικών χαρακτηριστικών του ήχου (έτσι όπως διαμορφώνονται από το χαρακτηριστικό envelope κάθε οργάνου) ως ένα ενιαίο αντιληπτικό φαινόμενο.

Μερικά από τα χαρακτηριστικά του ήχου των μουσικών οργάνων, τα οποία και αναλύονται στα επόμενα κεφάλαια, αναφέρονται εν συντομία στη συνέχεια, με βάση τις περιγραφές του Olson [116]. Το **φλάουτο**, σε αντίθεση με τα περισσότερα μουσικά όργανα, έχει θεμελιώδη συχνότητα που φέρει ένα σημαντικό μέρος της παραγόμενης ακουστικής ενέργειας. Σε χαμηλές συχνότητες είναι πιο πλούσιο σε αρμονικές, ενώ σε υψηλές συχνότητες οι αρμονικές είναι σχεδόν ανύπαρκτες προσδίδοντας στον ήχο καθαρότητα. Στο Σχήμα 2.3 φαίνεται το εύρος των συχνοτήτων των οργάνων που περιγράφονται. Το γεγονός ότι η θεμελιώδης συχνότητα περιέχει αυτή την ποσότητα ενέργειας οδηγεί στον χαρακτηριστικό ήχο του φλάουτου, το οποίο θεωρείται ως το πιο λαμπρό και πιο αγνό από όλα τα μουσικά όργανα. Στο **κλαρινέτο**, επίσης, το μεγαλύτερο μέρος της ενέργειας βρίσκεται στη θεμελιώδη συχνότητα, γεγονός που καθιστά τον ήχο του διαυγή και λαμπερό. Στις χαμηλότερες συχνότητες παράγει δυνατούς ήχους, ενώ οι άρτιες αρμονικές είναι σχεδόν ανύπαρκτες λόγω του κυλινδρικού σωλήνα που είναι

[3]If you take each moment of each frequency component into account, you realize that sounds can differ from one another acoustically in an astonishingly large number of ways. Does the auditory system deal with this complexity by collapsing the differences into a small number of dimensions?

κλειστός στο ένα άκρο. Το **όμποε**, έχει το μεγαλύτερο μέρος της ενέργειάς του στη θεμελιώδη συχνότητα και τις αρμονικές με συχνότητα 500–1500 Hz. Ο ήχος του είναι φωτεινός και ορισμένες φορές θυμίζει το φλάουτο όσον αφορά την καθαρότητα. Οι χαμηλότεροι τόνοι είναι πλούσιοι σε αρμονικές, και αυτό έχει ως συνέπεια ήχο οξύ και συχνά διαπεραστικό. Το **σαξόφωνο**, από την άλλη, έχει αρμονικές με πολλή ενέργεια, ενώ η δομή τους δεν παρουσιάζει κάποιο συγκεκριμένο μοτίβο. Ο ήχος του μοιάζει τόσο με τα χάλκινα όσο και με τα ξύλινα πνευστά και χαρακτηρίζεται γεμάτος και πλούσιος σε συνδυασμό με τη δυνατή του ένταση. Η **τρομπέτα** επίσης έχει πλούσιο αρμονικό περιεχόμενο σε όλο το εύρος της, αν και η ανάπτυξη του πλάτους στις χαμηλότερες αρμονικές είναι ταχύτερη σε σχέση με τις υψηλότερες αρμονικές. Παρατηρείται επίσης αύξηση του αριθμού των αρμονικών όσο αυξάνεται η ένταση κατά τη χρονική εξέλιξη της νότας [123] και χαρακτηρίζεται από την καθαρότητα και τη φωτεινότητα του ήχου της. Παρ' όλα αυτά ο ήχος της τρομπέτας μπορεί να είναι εξίσου διαπεραστικός αλλά και μαλακός και πλούσιος. Αντίθετα, στο **φαγκότο** η θεμελιώδης συχνότητα και οι χαμηλότερες αρμονικές είναι χαμηλές ως προς την ένταση σε χαμηλές συχνότητες, ενώ η **τούμπα** παράγει δυνατούς ήχους στην περιοχή των χαμηλών συχνοτήτων. Το **κόρνο** παίζει σε ένα υψηλότερο τμήμα των αρμονικών του σε σύγκριση με τα περισσότερα χάλκινα πνευστά, ενώ το κωνικό του σχήμα θεωρείται υπεύθυνο για τον χαρακτηριστικό ήχο του ο οποίος συχνά περιγράφεται ως «γλυκός». Τέλος, το αρμονικό περιεχόμενο των ήχων του **μπάσου** είναι πολύ πλούσιο στις χαμηλότερες συχνότητες.

Αν και είναι αρκετά εύκολο για τους ανθρώπους, ειδικά τους εκπαιδευμένους μουσικούς, να αναγνωρίζουν τα διάφορα όργανα, αυτό δεν ισχύει όταν ακούγεται μόνο η σταθερή μέση κατάσταση της νότας. Η δυσκολία στη διαφοροποίηση της χροιάς των οργάνων έγκειται επίσης στο γεγονός ότι είναι ένα πολυδιάστατο φαινόμενο, με παρεπόμενο να μην είναι εφικτό να συμβολιστεί από μονοδιάστατες κλίμακες οι οποίες θα μπορούσαν να χρησιμοποιηθούν για σύγκριση ή κατάταξη [134]. Στη δεκαετία του 1970 ο μουσικολόγος Ian Gray [55] πρότεινε τρεις βασικές «διαστάσεις» της χροιάς των οργάνων, τρία χαρακτηριστικά (διαφορετικά της συχνότητας και του πλάτους) που κάνουν τον έναν ήχο διαφορετικό από τον άλλο. Τα χαρακτηριστικά αυτά είναι η φωτεινότητα (η ένταση των υψηλών αρμονικών), το attack (η σταδιακή αύξηση των αρμονικών ώσπου να φτάσει ο ήχος στη σταθερή του κατάσταση) και η κατανομή των αρμονικών κατά τη χρονική διάρκεια μιας νότας (η αυξομείωση του πλάτους τους). Άρα η αναγνώριση των μουσικών οργάνων βασίζεται στο άκουσμα όλων των μεταβατικών σταδίων μιας νότας, εκτός της σταθερής της κατάστασης, που σημαίνει την αρχή (attack) και το τέλος του (release) [58], δεδομένου ότι τα μεταβατικά αυτά στάδια περιλαμβάνουν κάποιες συνιστώσες θορύβου που επηρεάζουν την αντίληψη της χροιάς. Για παράδειγμα, το φλάουτο, με τη σχετικά απλή αρμονική δομή του, προκειμένου να αποκτήσει τον ξεχωριστό του ήχο, θα πρέπει να προηγηθεί




*Σχήμα 2.4:* Αρχή (*attack*), μεσαία σταθερή κατάσταση (*steady state*) και τέλος (*release*) της κυματομορφής της νότας A3 για το B♭ κλαρινέτο.

από ένα μικρό ήχο «puff» ή θόρυβο. Αυτό είναι ένα χαρακτηριστικό στοιχείο της χροιάς του, το οποίο δεν μπορεί να επιτευχθεί από έναν συνθετικό ήχο [113], και θα εξαφανιζόταν, αν ακουγόταν μόνο η σταθερή κατάσταση του ήχου. Το ίδιο ισχύει και για την τρομπέτα, ενώ ομοίως ζωτικής σημασίας ακουστικό χαρακτηριστικό είναι το «ξύσιμο» του δοξαριού στη χορδή του βιολιού, ή το τρίξιμο που κάνει το γλωσσίδι του κλαρινέτου [58]. Οι Iverson *et al.* [66] σύγκριναν τη συνεισφορά του attack και της σταθερής κατάστασης της νότας και κατέληξαν στο συμπέρασμα ότι είναι συγκρίσιμη, υποδεικνύοντας ότι εξέχοντα χαρακτηριστικά για τη δημιουργία ολοκληρωμένων μουσικών τόνων παρουσιάζονται και στις δύο καταστάσεις. Ωστόσο, αναφέρουν πως η απουσία του attack μπορεί να επηρεάσει αρνητικά την κατάταξη ενός οργάνου σε πνευστό, έγχορδο ή κρουστό.

Η διάρκεια των μεταβατικών αυτών καταστάσεων του ήχου ποικίλλει όχι μόνο μεταξύ των οργάνων αλλά και μεταξύ των τόνων στις υψηλότερες και στις χαμηλότερες οκτάβες. Όπως αναφέρει ο Hall [58], χαρακτηριστική διάρκεια του attack των τόνων μπορεί να είναι τα 20 ms ή λιγότερο για το όμποε, 30–40 ms για το κλαρινέτο και την τρομπέτα, 70–90 ms για το φλάουτο και το βιολί. Επιπλέον, οι νότες πάνω από τη μεσαία Ντο (C) (που ορίζεται και ως C4 στα περίπου 261 Hz) έχουν περίοδο 2–4 ms, με αποτέλεσμα να χρειάζονται αρκετές δεκάδες περίοδοι δονήσεων έως ότου επέλθει η σταθερή κατάσταση. Ωστόσο, στο [47] η διάρκεια του attack μιας νότας αναφέρεται ως $50 \pm 20$ ms, ανεξάρτητα από τη νότα ή το όργανο. Τα εν λόγω στοιχεία σχετικά με τις διαφορές των μεταβατικών καταστάσεων των τόνων μας οδηγούν στην υπόθεση ότι η συνολική διάρκεια μιας νότας προσφέρει ζωτικής σημασίας ενδείξεις για την ταυτοποίηση της. Το Σχήμα 2.4 δείχνει το attack, τη σταθερή κατάσταση και το release για τη νότα A3 του B♭ κλαρινέτου.

Στην ανάλυση που ακολουθούμε προσπαθούμε να λάβουμε υπόψη μας τις διαφορές των μεταβατικών αυτών καταστάσεων και να αναλύσουμε τους μουσικούς ήχους βάσει αυτών.

## 2.2 Είδη Μουσικης

### 2.2.1 Στυλ και είδος

Ο όρος *«μουσικό είδος»* αποτελεί τον πιο δημοφιλή και ευρέως διαδεδομένο τρόπο περιγραφής ενός μουσικού κομματιού τόσο μεταξύ των χρηστών όσο και στη μουσική βιομηχανία [5]. Αποτελεί τη βασική μέθοδο, η οποία για χρόνια τώρα χρησιμοποιείται, οργάνωσης βάσεων δεδομένων, συστηματοποίησης μουσικών βιβλιοθηκών και μουσικών καταστημάτων αλλά και περιγραφής της μουσικής για τη διευκόλυνση των χρηστών κατά την εύρεση ενός μουσικού κομματιού ή παραπλήσιων καλλιτεχνών και δίσκων. Μολονότι δεν θεωρείται δόκιμος ως ορισμός αφού τα όρια των διαφορετικών ειδών είναι ασαφή [147], αποτελεί παρά ταύτα την καλύτερη δυνατή επιλογή περιγραφικού προσδιορισμού των χαρακτηριστικών ενός μουσικού κομματιού καθώς διευκολύνει την εύρεση ομοιοτήτων και διαφοροποιήσεων μεταξύ των διαφορετικών ειδών.

Η τακτική του ανθρώπου κατά την προσπάθεια εύρεσης νέας ή άγνωστης μουσικής είναι τελείως υποκειμενική. Για παράδειγμα αναζητά μουσική με παρεμφερή γνωρίσματα και βασίζεται σε χαρακτηριστικά όπως τη μελωδία, την αρμονία, το ρυθμό κ.ά. [52], με συγκεκριμένο συναισθηματικό περιεχόμενο, ή μουσική συγκεκριμένου στυλ και *«υφής»* [65]. Άλλωστε, αν κοιτάξουμε τις δικές μας δισκοθήκες θα δούμε πως κάποιοι ταξινομούν τα μουσικά τους αρχεία βάσει της χρονολογίας, του καλλιτέχνη, της χώρας προέλευσης και κυρίως με βάση το είδος.

Ένα παράδειγμα που αποδεικνύει τη διαφοροποίηση στη μεθοδολογία που θα μπορούσε να χρησιμοποιήσει κάποιος για την περιγραφή ενός μουσικού κομματιού, όπως για παράδειγμα το *Yesterday* των Beatles είναι το ακόλουθο [5]. Από τη μία μπορούμε να το περιγράψουμε ως *«Brit-Pop»*, λόγω του ότι το συγκεκριμένο συγκρότημα άνθισε στην Βρετανία τη δεκαετία του ′60 και έβαλε τις βάσεις για το συγκεκριμένο μουσικό είδος (*intentional concept*), ενώ από την άλλη θα μπορούσαμε να το χαρακτηρίσουμε ως ένα γλυκό «*pop*» κομμάτι λόγω της ταχύτητας του, της ύπαρξης των έγχορδων και της μελαγχολικής φωνής (*extensional concept*). Εξάλλου σε έρευνα του 2000 [118] η οποία πραγματοποιήθηκε σε τρεις μεγάλες διαδικτυακές βάσεις δεδομένων, μεταπώλησης και διανομής μουσικής (allmusic.com, amazon.com, mp3.com) οι οποίες αποτελούνται από 531, 719 και 430 μουσικά είδη αντίστοιχα, βρέθηκε πως οι κοινές λέξεις περιγραφής είναι μόνο 70. Μουσικά είδη όπως η *rock* και η *pop* δεν ορίζονται με τον ίδιο τρόπο ενώ ίδιες κατηγορίες δεν περιλαμβάνουν τα ίδια μουσικά κομμάτια.

Φυσικά ο άνθρωπος, σύμφωνα με την έρευνα των Gjerdingen και Perrott [52], στην προσπάθεια του να χαρακτηρίσει ένα μουσικό κομμάτι επηρεάζεται από πολλούς διαφορετικούς παράγοντες: (α) την ύπαρξη συγκεκριμένων ακουστικών γνωρισμάτων (*distinctive features*) τα οποία συνδέονται με συγκεκριμένα είδη (π.χ. κατά την δεκαετία

του 1990 η new-country μουσική συνδεόταν άμεσα με την ύπαρξη του ήχου του βιολιού), (β) την περίοδο που πλάθεται ο χαρακτήρας του (*the plasticity period*) και άρα τα εφηβικά χρόνια κατά τα οποία διαμορφώνονται οι μουσικές του προτιμήσεις. (γ) Το φαινόμενο «*Fisheye-Lens*» που υπαγορεύει πως αναλόγως με τις προτιμήσεις του, ο άνθρωπος είναι γνώστης των συγκεκριμένων ειδών και υποκατηγοριών τους αλλά όχι άλλων, (δ) η ηλικία, το φύλο, η μόρφωση, η καταγωγή κ.ά. (*demographic bias*). Χαρακτηριστικό παράδειγμα αποτελεί το είδος «Rhythm & Blues» το οποίο για μικρότερους ηλικιακά ανθρώπους αποτελεί ξεχωριστό είδος, ενώ στην πραγματικότητα ήταν ένας μουσικός όρος για τον χαρακτηρισμό της μουσικής των Αμερικάνων της Αφρικής και άρα κάτι τελείως διαφορετικό από αυτό που ίσως οι περισσότεροι θεωρούμε. (ε) Συνήθως κατηγορίες με συνοπτικά ονόματα όπως rock, jazz, country αποτελούν βασικές κατηγορίες της μουσικής, ενώ άλλες όπως cool jazz, new traditional country αποτελούν υποκατηγορίες. Τέλος, (ζ) οι μουσικόφιλοι κυρίως έχουν την τάση να περιγράφουν τη μουσική χρησιμοποιώντας περισσότερους από έναν όρους (π.χ. «*Country* με στοιχεία από *blues-rock* και μπάσο με αίσθηση *rockabilly*»), ενώ επίσης αναπτύσσουν πλούσιες μουσικές αναπαραστάσεις σε αντίθεση με τους μη-γνώστες οι οποίοι βασίζουν την απόφαση τους για τη διάκριση των διαφορετικών ειδών σε 1–2 παραδείγματα.

Γενικά αντιλαμβανόμαστε πως ο κάθε άνθρωπος, σύμφωνα με τις προτιμήσεις του αλλά και τις «μουσικές του γνώσεις», περιγράφει τα μουσικά δεδομένα με διαφορετικό τρόπο. Αυτό έχει ως συνέπεια την ύπαρξη πολλών διαφορετικών ειδών αποτελούμενων συνήθως από εξίσου πολλές υποκατηγορίες, οι οποίες έχουν δημιουργηθεί κατά το πέρας των χρόνων τόσο εξαιτίας των διαφορετικών περιγραφών όσο και λόγω της συνεχούς εξέλιξης της μουσικής. Καταλήγουμε λοιπόν πως η αναγνώριση των ειδών της μουσικής αποτελεί ένα αρκετά σύνθετο και δύσκολο πρόβλημα εκ φύσεως.

Έως τώρα, αναδείξαμε την πολυπλοκότητα της ταυτοποίησης του είδους ενός μουσικού κομματιού. Αν θεωρήσουμε όμως πως οι περισσότεροι, με την πάροδο των ετών και της μουσικής ακρόασης, αναπτύσσουμε την ικανότητα να μπορούμε να αναγνωρίζουμε το είδος του, τότε ποια είναι η χρονική διάρκεια που χρειαζόμαστε για να αντεπεξέλθουμε στη συγκεκριμένη εργασία. Σε έρευνα [52] που πραγματοποιήθηκε με στόχο την εύρεση αυτής της διάρκειας, βρέθηκε πως ο άνθρωπος χρειάζεται μόνο 250 ms για την αναγνώριση της κλασικής μουσικής (σε ποσοστό επιτυχίας 70%), παρ' όλο που στη μικρή αυτή χρονική διάρκεια το μουσικό κομμάτι εξελίσσεται ελάχιστα (π.χ. μία αρμονία, 1–2 νότες στη μελωδία και το μπάσο). Για άλλα είδη όπως για παράδειγμα το blues το ποσοστό αυτό μειώνεται αισθητά, ενώ τα πειραματικά τους αποτελέσματα έδειξαν πως το ποσοστό αναγνώρισης δέκα διαφορετικών ειδών είναι 72% όταν ακούμε τρία δευτερόλεπτα και 60% για μισό δευτερόλεπτο. Η επιτυχία της διάκρισης των διαφορετικών μουσικών κατηγοριών με το άκουσμα τόσο μικρής χρονικής διάρκειας θεωρούν πως είναι εφικτή

γιατί το ηχόχρωμα περικλείει τόσο τη φασματική όσο και τη χρονική εναλλαγή του ακουστικού σήματος, κάτι το οποίο είναι ενδεικτικό της διαφορετικότητας των ειδών. Με άλλα λόγια ο άνθρωπος χρησιμοποιεί και αντιλαμβάνεται ακούσια την πολύ άμεση πληροφορία της μουσικής σχετική με το ηχόχρωμα, την υφή και τα όργανα που παίζουν, χωρίς να διερωτηθεί αν για παράδειγμα το μουσικό κομμάτι είναι γρήγορο ή αργό ή αν περιλαμβάνει το άκουσμα κάποιου συγκεκριμένου οργάνου.

Σύμφωνα με πειραματικές αξιολογήσεις η παρουσία του στυλ ή αλλιώς είδους βρίσκεται και στον κάθετο άξονα της μουσικής [52]. Ο κάθετος άξονας μας δίνει πληροφορίες για την αρμονική δομή της μουσικής και άρα των μουσικών τόνων που συμμετέχουν στην σύνθεση της αρμονίας αυτής, ενώ ο οριζόντιος άξονας από την άλλη μας δίνει πληροφορίες για την έναρξη των διάφορων μουσικών οργάνων και άρα για το ρυθμό, την ταχύτητα του μουσικού κομματιού αλλά και την εξέλιξη της μελωδίας.

Στη διατριβή αυτή, μελετάμε το θέμα της αναγνώρισης των ειδών μουσικής στα πλαίσια των μικροδομών και των μακροδομών των ήχων αυτών, εξάγοντας αναπαραστάσεις του σήματος βασισμένες στη θεωρία διαμορφώσεων. Όπως θα δείξουμε σε επόμενα κεφάλαια, τα χαρακτηριστικά που εξετάζουμε εμπεριέχουν πολύτιμη πληροφορία για την εφαρμογή που εξετάζουμε.

## 2.3 Εξαγωγή Χαρακτηριστικών και Μέθοδοι Αναγνώρισης

### 2.3.1 Εξαγωγή Χαρακτηριστικών

#### Γενικά Χαρακτηριστικά

Το πρώτο βήμα σε ένα σύστημα αναγνώρισης μουσικής και δημιουργίας αναπαραστάσεων του μουσικού σήματος, βάσει οποιουδήποτε πλαισίου κατηγοριοποίησης, είναι η εξαγωγή χαρακτηριστικών από το μουσικό σήμα. Κατά καιρούς έχουν προταθεί διάφορα σύνολα χαρακτηριστικών, που προσπαθούν να ανιχνεύσουν τη χρονική μικροδομή των ήχων, και άρα υπολογίζονται σε μικρά ηχητικά τμήματα λίγων χιλιοδευτερολέπτων (ms), τη χρονική μακροδομή και άρα υπολογίζονται σε μεγαλύτερα ηχητικά τμήματα ή βασίζονται σε μετρήσεις των αναπαραστάσεων των μικροδομών, όπως ο μέσος όρος, η διακύμανση κ.ά. Τα χαρακτηριστικά αυτά μπορούν να μετρηθούν (α) απευθείας από την κυματομορφή, (β) ύστερα από κάποιο μετασχηματισμό του σήματος (π.χ. FFT (fast Fourier transform), μετασχηματισμό κυματιδίων (wavelets)), (γ) βάσει κάποιου μοντέλου, όπως το AM-FM (Amplitude-Frequency Modulation) ή το sinusoidal μοντέλο ενώ τέλος (δ) δύναται να εξαχθούν με τρόπο ώστε να μιμούνται τις

λειτουργίες του ανθρώπινου ακουστικού συστήματος.

Συγκεκριμένα, για το θέμα της αναγνώρισης ειδών μουσικής οι μεθοδολογίες που έχουν προταθεί προσεγγίζουν το θέμα είτε βάσει κανόνων (*prescriptive approaches*) όπου και εξάγονται χαρακτηριστικά χαμηλού επιπέδου, είτε βάσει μετρικών ομοιότητας των μουσικών κομματιών (*emergent approaches*) [5]. Τα πιο συνηθισμένα χαρακτηριστικά που χρησιμοποιούνται μπορεί να βασίζονται στην αντίληψη και σε διάφορους ψυχοακουστικούς μετασχηματισμούς (π.χ., Mel Frequency Cepstral Coefficients (MFCC), sharpness, loudness) [90], στη γραμμική πρόβλεψη, στην ενέργεια (π.χ., συνολική ενέργεια, αρμονική ενέργεια ή ενέργεια του θορύβου), στον χρόνο (π.χ., zero-crossing rate, log-attack time, temporal decrease κ.ά.), στο φάσμα και σε στατιστικές μετρήσεις του (π.χ., centroid, spectral flux, spectral rolloff, bandwidth, harmonic energy skewness) [1, 19, 35, 115], στην αρμονικότητα του σήματος (π.χ., βασική συχνότητα, inharmonicity, odd to even ratio κ.ά.), καθώς επίσης και στις παραγώγους των διαφόρων χαρακτηριστικών. Επιπρόσθετα, σε διάφορες μελέτες χρησιμοποιούνται χαρακτηριστικά που βασίζονται στο πρότυπο MPEG-7 (π.χ., harmonic centroid, deviation, spread, variation κ.ά.), καθώς και στη δομή της μουσικής (*music-related features*) (όπως ο ρυθμός [4, 179], το τονικό ύψος, modality, articulation, dynamics, brightness [49] κ.ά.). Τα τελευταία είναι πολύ σημαντικά ειδικά σε εφαρμογές αναγνώρισης βάσει του είδους ή των εκφραστικών ιδιοτήτων [48], όπου ο συνδυασμός των δομικών στοιχείων της μουσικής παίζει σημαντικό ρόλο για την επίτευξη της διαφορετικότητας των μουσικών συνθέσεων. Μεταξύ άλλων, ο μετασχηματισμός κυματιδίων (wavelet transform) έχει χρησιμοποιηθεί για την εξαγωγή χαρακτηριστικών και την ανάλυση ακουστικών σημάτων [44, 180], για την αυτόματη κατηγοριοποίηση [81], τη δημιουργία ευρετηρίων [171] αλλά και σε εφαρμογές αναγνώρισης μουσικών φράσεων [82, 130]. Τέλος, αρκετοί αλγόριθμοι πλέον χρησιμοποιούν βραχέος χρόνου φασματικά χαρακτηριστικά τα οποία συνδυάζονται ή ομαδοποιούνται για τη δημιουργία στατιστικών μετρήσεων που αφορούν μεγαλύτερες χρονικές διάρκειες [147, 155]. Για επισκόπηση και περαιτέρω λεπτομέρειες σχετικά με τα διάφορα χαρακτηριστικά που έχουν χρησιμοποιηθεί σε εφαρμογές κατηγοριοποίησης μουσικών σημάτων βλ. [5, 60, 125, 147].

**Mel Frequency Cepstral Coefficients (MFCCs)**

Σε πολλές μελέτες όπου διάφορα χαρακτηριστικά (μεταξύ άλλων τα MFCC και τα φασματικά χαρακτηριστικά κ.ά.) συνδυάζονται για τη βελτίωση των συστημάτων αναγνώρισης και συγκρίνονται έχει καταδειχθεί πως τα MFCC υπερτερούν [30, 35, 115, 135, 159]. Για το λόγο αυτό, στην έρευνά μας συγκρίνουμε την απόδοση των προτεινόμενων χαρακτηριστικών με αυτή των MFCC.

Τα MFCC (*Mel Frequency Cepstral Coefficients*) τα οποία προτάθηκαν από τους Davis

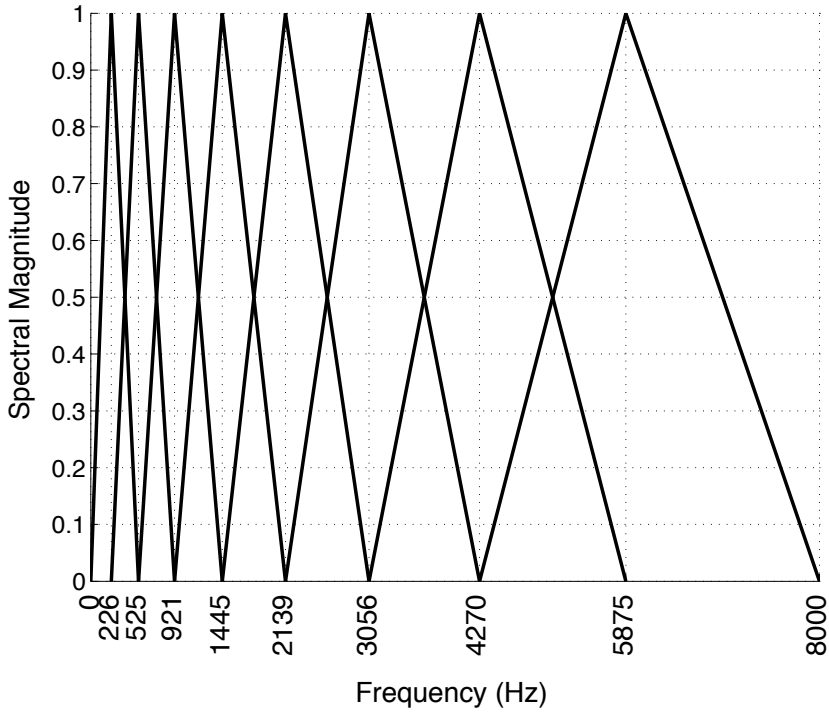
και Mermelstein [29] συγκαταλέγονται στα πιο διαδεδομένα διανύσματα χαρακτηριστικών τόσο σε εφαρμογές που αφορούν την επεξεργασία φωνής όσο και της μουσικής λόγω της καλής απόδοσής τους και της χαμηλής υπολογιστικής τους πολυπλοκότητας. Τα MFCC προσομοιώνουν το ανθρώπινο ακουστικό σύστημα και απεικονίζουν την περιβάλλουσα του φάσματος (*spectral envelope*) ενός ηχητικού σήματος. Ως περιβάλλουσα ορίζεται το σχήμα του φάσματος ισχύος ενός μικρού τμήματος του ακουστικού σήματος. Η φασματική περιβάλλουσα έχει την ικανότητα να αναπαραστήσει σημαντικές αντιληπτικές πληροφορίες του ήχου, ενώ σημαντικό χαρακτηριστικό της αποτελεί το γεγονός πως ήχους με παραπλήσια φασματική περιβάλλουσα τους αντιλαμβανόμαστε ως παρόμοιους.

Τα MFCC στηρίζονται στην ικανότητα της ομομορφικής ανάλυσης να διαχωρίζει το σήμα πηγής από τη διέγερση του ηχητικού σωλήνα [142]. Για τον υπολογισμό τους ακολουθούνται τρία βήματα. Αρχικά, υπολογίζεται η λογαριθμική ενέργεια των σημάτων, τα οποία έχουν φιλτραριστεί από μία τριγωνική συστοιχία φίλτρων σε κλίμακα mel (στο Σχήμα 2.5 φαίνεται συστοιχία 8 τριγωνικών φίλτρων βασισμένη στην Mel κλίμακα). Στη συνέχεια, η λογαριθμική ενέργεια χρησιμοποιείται για τον υπολογισμό του cepstrum. Τέλος, χρησιμοποιούμε τον διακριτό μετασχηματισμό συνημιτόνου (*Discrete Cosine Transform, DCT*) για την αποσυσχέτιση και τη μείωση της διάστασης των συντελεστών του λογαριθμικού Mel φάσματος. Ο υπολογισμός των MFCC για 20 τριγωνικά φίλτρα σύμφωνα με το [29] ορίζεται ως:

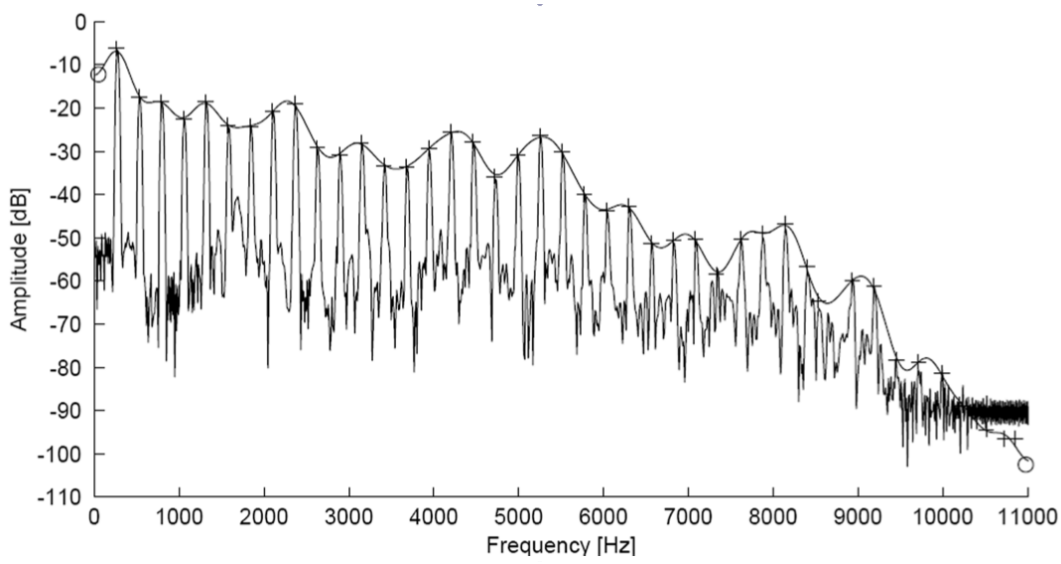
$$\text{MFCC}_i = \sum_{k=1}^{20} E_k \cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{20}\right],\ i=0,...,M \tag{2.1}$$

όπου $M$ ο αριθμός των συντελεστών του cepstrum και $E_k$ η λογαριθμική ενέργεια στην έξοδο του $k$-οστού φίλτρου, όπου $k=1,2,...,20$. Η μείωση της διάστασης των συντελεστών του DCT έχει ως αποτέλεσμα τη συμπαγή και εξομαλυμένη προσέγγιση της φασματικής περιβάλλουσας. Το Σχήμα 2.6 παρουσιάζει τη φασματική περιβάλλουσα μαζί με το φάσμα ενός ακουστικού παραθύρου νότας από βιολί.

Γενικά η επεξεργασία με συστοιχίες φίλτρων συναντάται πάρα πολύ συχνά σε μεθόδους ανάλυσης και εξαγωγής χαρακτηριστικών και βασίζεται στο γεγονός πως το ακουστικό σύστημα του ανθρώπου επεξεργάζεται την εισερχόμενη πληροφορία φιλτράροντάς την σε διαφορετικές επικαλυπτόμενες συχνοτικές περιοχές. Σε αυτή την περίπτωση σημαντικό ρόλο αποτελεί η επιλογή τόσο του είδους των φίλτρων, όσο και το εύρος και η κεντρική συχνότητα που τοποθετούνται τα φίλτρα αυτά. Για τον υπολογισμό των MFCC χρησιμοποιείται η κλίμακα Mel, ενώ παράδειγμα άλλης κλίμακας είναι η Bark. Φίλτρα τα οποία συναντώνται συχνά εκτός των τριγωνικών είναι τα Gabor, Gammatone κ.ά.




*Σχήμα 2.5:* Συστοιχία 8 τριγωνικών φίλτρων βασισμένη στη Mel κλίμακα.




*Σχήμα 2.6:* Φασματική Περιβάλλουσα ηχητικού σήματος για το βιολί [154].

### Φράκταλς

Ιδέες της φράκταλ θεωρίας[4] έχουν επίσης χρησιμοποιηθεί για την ανάλυση της δομής της μουσικής. Από διάφορες μελέτες έχουν προκύψει στοιχεία πως οι μουσικοί ήχοι και

[4]Ο όρος «φράκταλ», που προέρχεται από τη λατινική λέξη *fractus*, η οποία σημαίνει «σπασμένο» (*broken*), επινοήθηκε από τον Mandelbrot για να περιγράψει αντικείμενα που είναι ασυνήθιστα (ή «κατακερματισμένα») και άρα δεν μπορούν να περιγραφούν με τη βοήθεια της κλασικής γεωμετρίας [101]. Ο Mandelbrot ορίζει ένα σύνολο $F$ ως φράκταλ όταν έχει φράκταλ διάσταση που υπερβαίνει την τοπολογική του διάσταση. Ένα από τα πιο σημαντικά χαρακτηριστικά των φράκταλ είναι ότι έχουν παρόμοια δομή σε

η δομή των διαφορετικών ειδών μουσικής εμφανίζει φράκταλ πτυχές καθώς και ιδιότητες αυτοομοιότητας. Οι Voss και Clark [184, 185] διερεύνησαν τις $1/f^{\beta}$ πτυχές στη μουσική και στην ομιλία υπολογίζοντας το φάσμα ισχύος αργά μεταβαλλόμενων ποσοτήτων, όπως η ένταση και η συχνότητα. Στο [14] διερευνήθηκε η φράκταλ διάσταση (καθώς και η φράκταλ διάσταση σε πολλαπλές κλίμακες) διαφορετικών ειδών μουσικής, όπου και προτάθηκε πως η χρήση μετρήσεων της φράκταλ διάστασης μπορεί να ωφελήσει στη διάκριση των μουσικών ειδών. Οι Su και Wu [169, 170] εφάρμοσαν τον συντελεστή Hurst σε αλληλουχίες μουσικών τόνων και σε μουσικές φράσεις και σημείωσαν ότι η μουσική παρουσιάζει παρόμοιες φράκταλ ιδιότητες με την κίνηση Brown *(fractional Brownian motion)*. Ιδιότητες αυτοομοιότητας, όσον αφορά την ακουστική συχνότητα των σημάτων, παρατηρήθηκαν στο [63], όπου και μελετήθηκαν διάφορες πτυχές της φράκταλ γεωμετρίας. Η δομή και η πολυπλοκότητα των μουσικών ήχων βάσει μετρήσεων της φράκταλ διάστασης ή του $1/f$ θορύβου διερευνήθηκε επίσης στα [6, 28, 57, 112, 143, 144].

Δεδομένων των ενδείξεων για την ύπαρξη φράκταλ ιδιοτήτων στη μουσική, όπως η κίνηση Brown, η χρήση της φράκταλ διάστασης για την ταξινόμηση των διαφορετικών ειδών της μουσικής, καθώς και στοιχείων ύπαρξης αυτοομοιότητας στους μουσικούς τόνους, σκοπεύουμε να διερευνήσουμε περαιτέρω αν η φράκταλ ανάλυση σε πολλαπλές κλίμακες θα μπορούσε να μας δώσει πληροφορίες σχετικά με τη δομή των μουσικών σημάτων, λαμβάνοντας υπόψη ότι τέτοιες μέθοδοι έχουν ήδη χρησιμοποιηθεί με επιτυχία σε εφαρμογές αναγνώρισης φωνής [105, 133].

**AM-FM Διαμορφώσεις**

Ψυχοφυσιολογικές έρευνες έχουν αποδείξει ότι η ανθρώπινη ακοή βασίζεται σε μεγάλο βαθμό στις διαμορφώσεις πλάτους και συχνότητας. Το ανθρώπινο αυτί μέσω της διαδικασίας μεταγωγής των διαμορφώσεων συχνότητας σε διαμορφώσεις πλάτους (*FM to AM transduction*) μπορεί να αντιλαμβάνεται τις διαμορφώσεις συχνότητας [140, 176]. Η χρονική μικροδομή των μουσικών σημάτων συνίσταται σε στιγμιαίες διαμορφώσεις πλάτους και συχνότητας των κύριων συντονισμών της και χαρακτηρίζει τις κυματομορφές των ήχων αυτών. Μεγάλες διαμορφώσεις, όπως το βιμπράτο (*vibrato*) και το τρέμολο (*tremolo*) γίνονται εύκολα αντιληπτές, ενώ μικρότερες όχι, παρ' όλα αυτά συντελούν στη δημιουργία «φυσικών» ήχων [186], με ιδιαίτερη σημασία στη σύνθεση μουσικής [58]. Επιπλέον, οι διαμορφώσεις μπορούν να χρησιμοποιηθούν και στην ανάλυση μεσο- και μακροδομών για την περιγραφή των μουσικών φαινομένων και των σχέσεων των βασικών δομικών μονάδων τους.

Οι διαμορφώσεις έχουν μελετηθεί για την ανάλυση και τη σύνθεση των ήχων μουσικών οργάνων [173], προκειμένου να καθοριστούν οι παράμετροι σύνθεσης για

πολλαπλές κλίμακες.

τη μοντελοποίησή τους. Επίσης έχουν χρησιμοποιηθεί σε εφαρμογές αναγνώρισης, και συγκεκριμένα για τη διάκριση φωνής και μουσικής ή φωνής από μη φωνητικά σήματα [75,80,106,114,152,172], καθώς και στην κατηγοριοποίηση ειδών μουσικής [68]. Στα [38, 89] διαμορφώσεις πλάτους εξάγονται ως χαρακτηριστικά για την αναγνώριση μουσικών οργάνων και ειδών μουσικής, ώστε να περιγράψουν το φαινόμενο του τρέμολο, το οποίο μετριέται στο εύρος συχνοτήτων μεταξύ 4–8 Hz και την «ταχύτητα» του μουσικού σήματος στο εύρος 10-40 Hz. Παρόμοιες ιδέες βασισμένες στο μοντέλο ημιτονοειδών (*sinusoidal model*) [111] έχουν χρησιμοποιηθεί για τη μοντελοποίηση ήχων [153] και για τον διαχωρισμό του ήχου σε διαφορετικές πηγές (*source separation*) [22].

Ενδείξεις για την ύπαρξη μη-γραμμικών φαινομένων κατά την παραγωγή φωνής, όπως αυτά των διαμορφώσεων [104], έχουν οδηγήσει στην επεξεργασία και ανάλυση των σημάτων φωνής με τη βοήθεια μοντέλων διαμόρφωσης (AM-FM) με σκοπό την ανίχνευσή τους. Επιπλέον, έχει αναπτυχθεί ένας *μη-γραμμικός αλγόριθμος διαχωρισμού ενέργειας* (*Energy Separation Algorithm, ESA*) για την αποδιαμόρφωση των συντονισμών της φωνής σε συνιστώσες πλάτους και συχνότητας [104], με πολυζωνικές επεκτάσεις σε υπέρθεση πολλαπλών AM-FM σημάτων [17, 59, 137]. Η μοντελοποίηση αυτή έχει χρησιμοποιηθεί σε εφαρμογές αυτόματης αναγνώρισης φωνής [33] και σύνθεσης [138], ενώ επίσης έχει αποδειχθεί χρήσιμη στην αναγνώριση και στην ανίχνευση φωνής σε συνθήκες θορύβου [33, 40]. Στηριζόμενοι μεταξύ άλλων σε αυτές τις ενδείξεις για την ύπαρξη διαμορφώσεων σε σήματα μουσικής μελετάμε το μοντέλο διαμόρφωσης AM-FM σε σήματα μουσικών οργάνων και διαφορετικών ειδών μουσικής με σκοπό την κατηγοριοποίησή τους.

### 2.3.2 Μέθοδοι Αναγνώρισης

Διάφορες τεχνικές αναγνώρισης προτύπων έχουν εφαρμοστεί για τη μοντελοποίηση των μουσικών ήχων, οι οποίες πολλές φορές δεν είναι κατ' ανάγκη αποτελεσματικές στη μοντελοποίηση της χρονικής εξέλιξης των ήχων. Για παράδειγμα, τα Γκαουσιανά μοντέλα (*Gaussian mixture models, GMM*) είναι σε θέση να παραμετροποιήσουν την κατανομή των παρατηρήσεων, μολονότι δεν μπορούν να μοντελοποιήσουν τη δυναμική εξέλιξη των χαρακτηριστικών της μουσικής, όπως, για παράδειγμα, τα κρυφά Μαρκοβιανά μοντέλα (*hidden Markov models, HMM*). Επιπλέον διαχωρισμός των μεθόδων εκμάθησης των μοντέλων που χρησιμοποιούνται είναι: μέθοδοι με επίβλεψη (*supervised*), οι οποίες χωρίζονται στους στατικούς ταξινομητές, όπως τα Γκαουσιανά μοντέλα που προαναφέραμε, οι Μηχανές Διανυσμάτων υποστήριξης (*Support Vector Machines, SVM*), οι Κ-κοντινότεροι γείτονες (*K-nearest neighbor, KNN*), το Vector Quantization, η Linear Discriminant Analysis (LDA), τα Τεχνητά Νευρωνικά Δίκτυα (*Artificial Neaural Networks, ANNs*), και σε δυναμικούς όπως τα κρυφά Μαρκοβιανά μοντέλα (HMM), μέθοδοι χωρίς εκμάθηση των δεδομένων (*unsupervised*) όπως η τεχνική της ομαδοποίησης των

χαρακτηριστικών (*clustering*), π.χ. K-means, και τέλος διάφορα μετρικά ομοιότητας.

Οι ερευνητικές μελέτες που συγκρίνουν τους διαφορετικούς αλγόριθμους σε συστήματα κατηγοριοποίησης μουσικών οργάνων είναι λίγες [39, 91, 109]. Επιπλέον οι διάφορες μελέτες χρησιμοποιούν διαφορετικές βάσεις δεδομένων και ταυτοχρόνως αξιολογούν διαφορετικό αριθμό μουσικών οργάνων κάτι που καθιστά ανέφικτη τη σύγκρισή τους. Γι' αυτό το λόγο, οι ερευνητικές εργασίες οι οποίες βασίζονται σε ακριβή σύγκριση με προηγούμενες είναι σχεδόν ανύπαρκτες.

Κυριότεροι εκπρόσωποι αλγορίθμων της Αναγνώρισης Προτύπων στο θέμα της κατηγοριοποίησης μουσικών οργάνων είτε από μεμονωμένες νότες είτε από μουσικές φράσεις μεμονωμένων οργάνων είναι οι ακόλουθοι: ο αλγόριθμος KNN χρησιμοποιήθηκε σε διάφορες πρώιμες ερευνητικές εργασίες [1, 35, 110]. Οι ερευνητές του [35], σε επόμενη δουλειά τους [36] χρησιμοποιούν HMM και δείχνουν πως η αναγνώριση μουσικών οργάνων βελτιώνεται. Στο [91] συγκρίνονται τρεις διαφορετικές μέθοδοι οι οποίες είναι οι: Multidimensional Gauss, KNN και Learning Vector Quantization (LVQ) όπου καλύτερη απόδοση έχουν τα KNN. Τέλος τα SVMs χρησιμοποιούνται στα [37, 39, 71, 109], τα GMMs στα [19, 37, 39, 109] και τα HMMs στα [3, 34].

Αντίστοιχα στην κατηγοριοποίηση ειδών μουσικής δημοφιλείς αλγόριθμοι για τη μοντελοποίηση των μουσικών σημάτων είναι οι εξής: KNN [146, 179], GMM [119, 179], HMM [36, 132, 146, 157, 166] και SVM [56, 87, 88, 90, 100, 146, 190], ενώ στις εργασίες [131, 132] HMM και μεταβλητής διάρκειας HMM χρησιμοποιήθηκαν για την ταξινόμηση μουσικών μοτίβων.

Η μέθοδος Non-Negative Matrix Factorization (NMF) έχει χρησιμοποιηθεί ευρέως σε θέματα αναγνώρισης μουσικής, όπως η μεταγραφή [107, 165], η κατηγοριοποίηση οργάνων [12], η κατηγοριοποίηση ειδών μουσικής [11, 62] καθώς και στην αναγνώριση ακουστικών γεγονότων [27]. Συγκεκριμένα, στο [62] εφαρμόζεται η μέθοδος NMF για τη δημιουργία αναπαραστάσεων του ηχοχρώματος των μουσικών ήχων οι οποίες και μοντελοποιούνται στη συνέχεια με Γκαουσιανές. Η μοντελοποίηση αυτή βελτιώνει τα ποσοστά αναγνώρισης σε σχέση με τα MFCC, συγκεκριμένα επιφέρει επιτυχία 73.9% στη βάση GTZAN [179], ενώ οι συγγραφείς θεωρούν πως η συμπίεση του χώρου των χαρακτηριστικών λόγω της τεχνικής που χρησιμοποιείται μειώνει τον χρόνο εκπαίδευσης των Γκαουσιανών μοντέλων καθώς επίσης και τον θόρυβο των δεδομένων. Στο [121] εφαρμόζεται η ιδέα των διαμορφώσεων (*joint-acoustic and modulation frequency*) λαμβάνοντας υπόψη τις αργές μεταβολές της διαμόρφωσης των μουσικών σημάτων. Η ιδέα αυτή βασίζεται σε εργασία των Sukittannon και Atlas με επιτυχία αναγνώρισης 91% μεταξύ δέκα διαφορετικών ειδών.

Στο [108] χρησιμοποιείται η τεχνική Optimum-Path Forest, η οποία και μοντελοποιεί το πρόβλημα της αναγνώρισης ως ένα γράφο στον χώρο των χαρακτηριστικών,

χρησιμοποιώντας 26 MFCC με επιτυχία αναγνώρισης 98.6%. Στο [10] εφαρμόζονται ιδέες της Όρασης Υπολογιστών για τη μοντελοποίηση των μουσικών σημάτων χρησιμοποιώντας clustering για την ομαδοποιήση των ακουστικών χαρακτηριστικών. Η μοντελοποίηση αυτή επέφερε ποσοστό επιτυχία 86.5% στην βάση GTZAN, ενώ τα χαρακτηριστικά που χρησιμοποιήθηκαν περιελάμβαναν στατιστικές μετρήσεις του φάσματος όπως η μέση τιμή, η διακύμανση, η μέγιστη και η ελάχιστη τιμή κ.ά. Τέλος, στο [84] η χρήση χαρακτηριστικών τα οποία μετράνε τις μακρο-διαμορφώσεις των μουσικών σημάτων (*octave-based modulation spectral contrast*), πέτυχαν ακρίβεια αναγνώρισης 84%.

Για επισκόπηση της σχετικής βιβλιογραφίας σχετικά με την αυτόματη αναγνώριση μουσικών οργάνων, ειδών μουσικής, καθώς και για επισκόπηση συστημάτων στον τομέα του Music Information Retrieval (MIR) βλ. τα [60, 147, 178] αντίστοιχα.

Στην ερευνητική αυτή εργασία για τη μοντελοποίηση των μουσικών σημάτων διαφορετικών μουσικών οργάνων χρησιμοποιούμε Γκαουσιανά και κρυφά Μαρκοβιανά μοντέλα, ενώ για την αναγνώριση των διαφορετικών ειδών μουσικής εφαρμόζουμε κρυφά Μαρκοβιανά μοντέλα και Support Vector Machines (SVM).

### Κρυφά Μαρκοβιανά Μοντέλα (HMM)

Τα κρυφά Μαρκοβιανά μοντέλα (*hidden Markov models - HMM*) [141] αποτελούν μία σημαντική μέθοδο αναγνώρισης προτύπων η οποία εφαρμόζεται με μεγάλη επιτυχία σε ποικίλες εφαρμογές. Τα HMMs είναι ένα στατιστικό εργαλείο που αντιπροσωπεύει μια σειρά στοχαστικών κρυφών (μη παρατηρήσιμων) Μαρκοβιανών διαδικασιών. Επιπλέον είναι δυναμικά μοντέλα, με την έννοια ότι είναι σχεδιασμένα για να καλύπτουν τις μακροδομές των παρατηρήσεων.

Σε ένα HMM καθορίζονται οι ακόλουθες ποσότητες και παράμετροι:

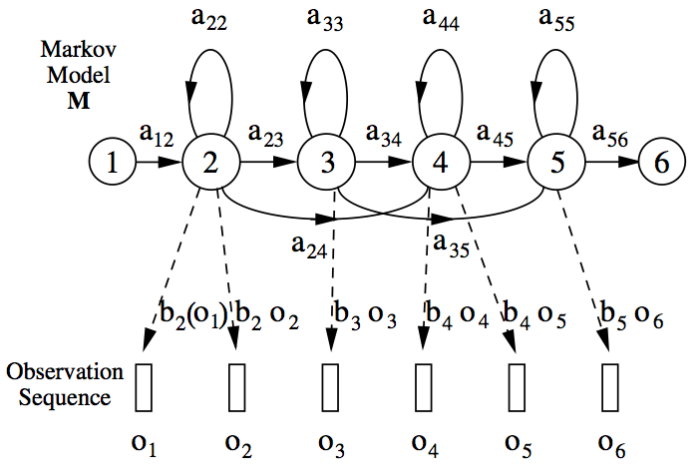
- $Q = \{1, 2, ..., N\}$, το σύνολο των κρυφών καταστάσεων με μέγεθος $N$.
- $O = (o_1, o_2, ..., o_T)$, η ακολουθία των παρατηρήσεων με διάρκεια $T$.
- $A = [a_{ij}]$, $i$=1,2,...,$n$ και $j$=1,2,...,$n$, η μήτρα πιθανοτήτων μετάβασης $a_{ij}$ από την κατάσταση $i$ στην κατάσταση $j$

$$a_{ij} = P[q_{t+1} = j | q_t = i],\ 1 \leq i, j \leq N \tag{2.2}$$

  όπου $q_t \in Q$ συμβολίζει την κατάσταση τη χρονική στιγμή $t$.

- $\pi = (\pi_1, \pi_2, ...\pi_N)$, οι αρχικές πιθανότητες των καταστάσεων:

$$\pi_i = P[q_1 = i],\ 1 \leq i \leq N \tag{2.3}$$




*Σχήμα 2.7:* Παράδειγμα left-to-right κρυφού Μαρκοβιανού μοντέλου με τέσσερις καταστάσεις (χωρίς την αρχική και τελική κατάσταση) [193].

- $B = b_i(\cdot) : i = 1, \ldots, N$, οι κατανομές πιθανότητας των παρατηρήσεων στις διαφορετικές καταστάσεις

$$b_i(o_t) = P[o_t|q_t = i], \ i = 1, \ldots, N \tag{2.4}$$

Το HMM μοντέλο μπορεί να συμβολιστεί συνοπτικά ως $\lambda = (A, B, \pi)$. Οποιαδήποτε εφαρμογή αναγνώρισης με HMM βασίζεται στην εκπαίδευση των παραμέτρων $\lambda = (A, B, \pi)$ και τον υπολογισμό και τη μεγιστοποίηση της πιθανότητας $P(O|\lambda)$. Για την επίλυση του προβλήματος χρησιμοποιείται ο αλγόριθμος Baum-Welch, γνωστός και ως EM (Expectation Maximization), ο οποίος κάνει χρήση του αλγόριθμου forward-backward. Τέλος, για την αποκωδικοποίηση και την εύρεση της πιθανότερης ακολουθίας καταστάσεων χρησιμοποιείται ο αλγόριθμος Viterbi [15].

Η πιθανότητα της κάθε κατάστασης να παράγει μία παρατήρηση συνήθως δίνεται με την χρήση συνδυασμού Γκαουσιανών κατανομών (*Gaussian mixtures*)

$$b_i(o_t) = \sum_{k=1}^{M} w_{jk} \mathcal{N}(o_t|\mu_{ik}, \Sigma_{jk}) \tag{2.5}$$

όπου $M$ ο αριθμός των Γκαουσιανών, $w_{jk}$ τα βάρη των Γκαουσιανών με $\sum_{k=1}^{M} w_{jk} = 1$ και $w_{jk} \geq 0$, για $1 \leq j \leq N, 1 \leq k \leq M$, και $\mathcal{N}(o_t|\mu, \Sigma)$ η πυκνότητα πιθανότητας της Γκαουσιανής με μέση τιμή $\mu$ και πίνακα αυτοσυσχέτισης $\Sigma$.

Τέλος, όσον αφορά την τοπολογία του HMM μοντέλου αυτή μπορεί να αποτελείται μόνο από μεταβάσεις από αριστερά προς τα δεξιά (left-to-right) (βλ. Σχήμα 2.7) είτε να είναι εργοδική (*ergodic*), δηλαδή να μπορεί να μεταβεί από οποιαδήποτε κατάσταση

σε οποιαδήποτε άλλη κατάσταση. Η επιλογή της τοπολογίας εξαρτάται από την κάθε εφαρμογή, παρ' όλα αυτά στην αναγνώριση φωνής χρησιμοποιείται η τοπολογία left-to-right.

**Support Vector Machines (SVM)**

Τα SVMs προτάθηκαν από τον Vapnik [182] το 1995, ως μηχανή μάθησης για την επίλυση προβλημάτων διαχωρισμού δύο τάξεων. Αν και αρχικά σχεδιάστηκαν για ταξινόμηση μόνο δύο τάξεων (*binary classification*), η επιτυχής επέκταση τους σε περισσότερες κατηγορίες τα καθιστά έναν από τους ταχύτερα αναπτυσσόμενους τομείς έρευνας στο χώρο της μηχανικής μάθησης επειδή επιτυγχάνουν ταχύτερη και καλύτερη γενίκευση ακόμα και σε περιπτώσεις όπου ο χώρος των χαρακτηριστικών είναι υψηλών διαστάσεων και τα δεδομένα εκπαίδευσης είναι λίγα. Έχει αποδειχθεί ότι υπερέχουν των κλασικών τεχνικών ελαχιστοποίησης σφάλματος, ενώ εφαρμόζονται σε πληθώρα εφαρμογών όπως η κατηγοριοποίηση κειμένων, η αναγνώριση αντικειμένων, η αναγνώριση ομιλίας και ομιλητών, καθώς και η αναγνώριση και κατηγοριοποίηση μουσικής με μεγάλη επιτυχία.

Τα SVMs προτείνουν σχεδιασμό των δικτύων που βασίζεται στα αποτελέσματα της Στατιστικής Θεωρίας Μάθησης (*Statistical Learning Theory*) και στην αρχή Ελαχιστοποίησης Δομημένου Κινδύνου (*Structural Risk Minimization, SRM*). Σε αντίθεση με υπάρχουσες κλασικές μεθόδους (όπως τα νευρωνικά δίκτυα) επιδιώκουν την ελαχιστοποίηση ενός άνω φράγματος του σφάλματος γενίκευσης μέσω μεγιστοποίησης του περιθωρίου μεταξύ του υπερεπιπέδου διαχωρισμού των δεδομένων. Η βασική ιδέα των SVMs για ένα πρόβλημα δυαδικού διαχωρισμού είναι η εύρεση μιας γραμμικής συνάρτησης η οποία ορίζει το υπερεπίπεδο το οποίο διαχωρίζει τις δύο τάξεις με αποτέλεσμα το μικρότερο σφάλμα γενίκευσης μεταξύ όλων των πιθανών υπερεπιπέδων. Το βέλτιστο υπερεπίπεδο έχει άρα το μέγιστο περιθώριο διαχωρισμού μεταξύ των τάξεων και ορίζεται ως το άθροισμα των αποστάσεων από το υπερεπίπεδο των πλησιέστερων σε αυτό σημείων των δύο τάξεων. Τα πλησιέστερα σημεία στο υπερεπίπεδο αυτό ονομάζονται διανύσματα υποστήριξης (*support vectors*). Ο διαχωρισμός μη-γραμμικών δεδομενών επιτυγχάνεται με τη χρήση πυρήνων (*kernels*), (π.χ. πολυωνυμικός, RBF (*Radial Basis Function*), Sigmoid κ.ά). Τέλος, η κλασική μεθοδολογία των SVMs, όπως ήδη αναφέραμε, λύνει το πρόβλημα της ταξινόμησης δύο τάξεων. Στις περισσότερες όμως εφαρμογές, οι τάξεις είναι περισσότερες. Σε αυτή την περίπτωση χρησιμοποιούνται προεκτάσεις της βασικής θεωρίας οι οποίες λύνουν το πρόβλημα της πολυταξικής ταξινόμησης, (π.χ. one vs. all, one vs. one), για λεπτομέρειες βλ. [15, 182].

**K-means clustering**

Ο K-means [98] είναι ένας από τους απλούστερους αλγόριθμους εκμάθησης χωρίς επίβλεψη που λύνουν το πρόβλημα της ομαδοποίησης (*clustering*) των δεδομένων ή διανυσμάτων χαρακτηριστικών, τα οποία είναι όμοια μεταξύ τους, βάσει κάποιου κριτηρίου ελαχιστοποίησης.
Τα βασικά βήματα του αλγορίθμου είναι τα ακόλουθα:

1) Εκ των προτέρων επιλογή του αριθμού των $K$ κέντρων.

2) Τυχαία δημιουργία $K$ ομάδων και καθορισμός των κέντρων (*centroids*) τους $\mu_k$, $k = 1, \ldots, K$.

3) Ταξινόμηση των δεδομένων στην ομάδα με το κοντινότερο κέντρο.

4) Υπολογισμός νέων $K$ κέντρων ως βαρύκεντρα (*barycenters*) των ομάδων που προέκυψαν από το προηγούμενο βήμα, μετά την αρχική ανάθεση όλων των χαρακτηριστικών σε ομάδες.

5) Ανάθεση των χαρακτηριστικών στα νέα κέντρα, έτσι ώστε το άθροισμα της τετραγωνικής απόστασης μεταξύ των δεδομένων με το κοντινότερο $\mu_k$ κέντρο να είναι ελάχιστο.

6) Επανάληψη των βημάτων 4 και 5 έως ότου ο αλγόριθμος συγκλίνει, δηλαδή τα κέντρα σταματήσουν να αλλάζουν θέση.

Η συνάρτηση ελαχιστοποίησης ορίζεται ως:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \tag{2.6}$$

όπου $||x_n - \mu_k||^2$ η τετραγωνική Ευκλείδεια απόσταση μεταξύ ενός διανύσματος χαρακτηριστικών $x_n$ και του κέντρου $\mu_k$ και $r_{nk} \in \{0, 1\}$ ένας δείκτης, όπου το $k = 1, ..., K$ δείχνει την ομάδα στην οποία έχει ανατεθεί το $x_n$. Όταν το $x_n$ έχει ανατεθεί στην ομάδα $k$, τότε $r_{nk} = 1$ και $r_{nj} = 0$ (για $j \neq k$) [15].

Το πρόβλημα άρα ανάγεται στην εύρεση των παραμέτρων $r_{nk}$ και $\mu_j$ για την ελαχιστοποίηση του $J$. Η επαναληπτική διαδικασία που διενεργείται για την εύρεση των παραμέτρων έως τη σύγκλιση του αλγορίθμου, αντιστοιχούν στα βήματα του αλγορίθμου EM (Expectation-Maximization) [15].

Τέλος, βασική παράμετρο για σωστή ομαδοποίηση αποτελεί η αρχική τυχαία τοποθέτηση των κέντρων, μιας και διαφορετική τοποθέτηση επιφέρει και διαφορετικό αποτέλεσμα. Αν και η διαδικασία της ομαδοποίησης πάντα συγκλίνει, ο αλγόριθμος K-means δεν βρίσκει απαραίτητα τη βέλτιστη παραμετροποίηση, άρα σε πολλές περιπτώσεις

βασική προϋπόθεση είναι η επανάληψη του αλγορίθμου για την εύρεση των καλύτερων αρχικών κέντρων.

### Προεπεξεργασία και Μετασχηματισμοί χαρακτηριστικών

#### *Ανάλυση σε Κύριες Συνιστώσες (PCA)*

Η ανάλυση κύριων συνιστωσών (*Principal Component Analysis, PCA*) [15] είναι ένας ορθογώνιος μετασχηματισμός, που προβάλλει το σύνολο των αρχικών χαρακτηριστικών σε ένα χώρο χαμηλότερης διάστασης, με τέτοιο τρόπο έτσι ώστε το σφάλμα ανακατασκευής να είναι το ελάχιστο δυνατό (το οποίο και υπολογίζεται ως το μέσο τετραγωνικό σφάλμα μεταξύ των διανυσμάτων χαρακτηριστικών στον αρχικό χώρο και στον χώρο προβολής). Επιπλέον, μετασχηματίζει το σύνολο χαρακτηριστικών, το οποίο μπορεί να αποτελείται από συσχετισμένες μεταβλητές, σε ένα σύνολο ασυσχέτιστων μεταβλητών, τις κύριες συνιστώσες. Η μήτρα μετασχηματισμού υπολογίζεται από τα $k$ ιδιοδιανύσματα που αντιστοιχούν στις $k$ μεγαλύτερες ιδιοτιμές της μήτρας αυτοσυσχέτισης των δεδομένων.

Ο μετασχηματισμός αυτός των χαρακτηριστικών σε έναν χώρο χαμηλότερης διάστασης, έχει ως αποτέλεσμα τη μείωση των παραμέτρων που υπολογίζονται κατά την εκπαίδευση του συστήματος κατηγοριοποίησης. Επίσης δημιουργεί σύνολα χαρακτηριστικών τα οποία είναι ασυσχέτιστα μεταξύ τους, κάτι το οποίο αποτελεί βασική παραδοχή για μοντελοποίηση με HMMs, όπου οι πιθανότητες παρατήρησης μοντελοποιούνται με μεγάλη επιτυχία όταν χρησιμοποιούνται Γκαουσιανές διαγώνιας συμμεταβλητότητας (*diagonal covariance matrices*).

#### *Μέθοδοι Επιλογής Χαρακτηριστικών*

Η επιλογή των χαρακτηριστικών αποτελεί ιδιαίτερα σημαντικό βήμα προ-επεξεργασίας στην αναγνώριση προτύπων, τόσο για την κατηγοριοποίηση και την επιτυχία αναγνώρισης όσο και για την υπολογιστική βελτιστοποίηση του συστήματος. Το ποσοστό της επιτυχούς κατηγοριοποίησης ενός συστήματος αναγνώρισης μπορεί να μειωθεί λόγω περιττών και ακατάλληλων χαρακτηριστικών ενώ η πολυπλοκότητα του συστήματος αυξάνεται όσο ο αριθμός των χαρακτηριστικών μεγαλώνει. Από την άλλη πλευρά, η αισθητή μείωση του αριθμού των χαρακτηριστικών μπορεί να οδηγήσει στη μείωση της διακριτικής ικανότητας ενός συνόλου χαρακτηριστικών και ως εκ τούτου στη μείωση της ακρίβειας του συστήματος αναγνώρισης.

Η αυτόματη επιλογή χαρακτηριστικών βασίζεται σε τεχνικές βελτιστοποίησης, λαμβάνοντας υπόψη μια ακολουθία από χαρακτηριστικά και επιλέγοντας ένα κατάλληλο υποσύνολο που οδηγεί στη μεγιστοποίηση κάποιου κριτηρίου. Οι βασικές κατηγορίες αλγορίθμων επιλογής και αναζήτησης χαρακτηριστικών (*feature selection algorithms*) είναι δύο: οι αλγόριθμοι αναζήτησης προς τα εμπρός (*sequential forward selection,*

*SFFS*) και οι αλγόριθμοι αναζήτησης προς τα πίσω (*sequential backward selection, SFBS*). Στην πρώτη περίπτωση (SFFS), ο αλγόριθμος ξεκινά με ένα μηδενικό σύνολο χαρακτηριστικών και, για κάθε βήμα, το καλύτερο χαρακτηριστικό που ικανοποιεί κάποιο κριτήριο περιλαμβάνεται στο τρέχον σύνολο χαρακτηριστικών. Ο αλγόριθμος ταυτοχρόνως ελέγχει τη δυνατότητα βελτίωσης του συγκεκριμένου κριτηρίου αν κάποιο χαρακτηριστικό αποκλειστεί από το υποσύνολο. Κατ' αυτόν τον τρόπο τα χειρότερα χαρακτηριστικά απορρίπτονται από το σύνολο, δηλαδή πραγματοποιείται ένα βήμα προς τα πίσω της διαδοχικής αυτής διαδικασίας (SBS). Ως εκ τούτου, ο αλγόριθμος SFFS εξελίσσεται δυναμικά αυξάνοντας και μειώνοντας τον αριθμό των χαρακτηριστικών έως ότου επιτευχθεί η επιθυμητή διάσταση. Ο αλγόριθμος SFBS λειτουργεί με ανάλογο τρόπο, μόνο που σε αυτή την περίπτωση η διαδικασία αναζήτησης ξεκινά από το αρχικό/συνολικό διάνυσμα χαρακτηριστικών έως ότου επιτευχθεί η επιθυμητή διάσταση.

## 2.4 Ανίχνευση Σημαντικών Γεγονότων (*Salient Events*)

Η ποσότητα των μουσικών/ηχητικών και πολυμεσικών δεδομένων στο Διαδίκτυο αυξάνεται συνεχώς (διαφορετικές ηχογραφήσεις, διαλέξεις και παρουσιάσεις, τηλεοπτικά προγράμματα κ.λπ.). Επισημειώσεις των αρχείων παρ' όλα αυτά δεν υπάρχουν, κάτι που καθιστά δύσκολη όχι μόνο την αναζήτηση των σωστών αρχείων αλλά και τη σάρωση του περιεχομένου τους. Αυτό έχει ως συνέπεια την ανάγκη εύρεσης νέων τεχνικών για την κατηγοριοποίηση, τον εντοπισμό γεγονότων και τη δημιουργία συνόψεων του περιεχόμενου των εν λόγω δεδομένων.

Το πιο ενδιαφέρον ζήτημα στην ανίχνευση γεγονότων είναι ο εντοπισμός των απότομων αλλαγών αντί των επαναλαμβανόμενων κανονικών γεγονότων, γι' αυτό και η κατάτμηση των ηχητικών ροών θεωρείται σε αυτή την περίπτωση ο κύριος στόχος ενός συστήματος. Παρ' όλα αυτά, όσον αφορά τα μουσικά δεδομένα, οι περισσότερες μελέτες στο συγκεκριμένο πεδίο ασχολούνται με την εύρεση των όμοιων τμημάτων ενός μουσικού κομματιού (και άρα του ρεφρέν) [26, 93, 158]. Με τον τρόπο αυτό έχει κανείς τη δυνατότητα να βρει φράσεις που παρουσιάζουν ομοιότητες, να τις ορίσει ως φράσεις «κλειδιά» (*key-phrases*), και κατά συνέπεια να τις επιλέξει ως σημαντικές.

Στη μουσική ο όρος «γεγονός» δύναται να σηματοδοτεί οποιαδήποτε υπομονάδα της δομής του μουσικού κομματιού όπως μια νότα, ένας χρονικός παλμός, μια συγχορδία, ή ένα μελωδικό τμήμα. Οι ερευνητικές μελέτες συχνά εστιάζουν στην ανίχνευση της εμφάνισης κάθε νότας, μέσω της ανίχνευσης δηλαδή του attack [158]. Η κυρία βιβλιογραφία στηρίζεται στην εξαγωγή βασικών ακουστικών χαρακτηριστικών όπως: ενέργεια, zero crossing rate, Linear Prediction Coefficients (LPC), MFCC, chroma χαρακτηριστικά κ.ά., ενώ η εύρεση των σημαντικών γεγονότων προσεγγίζεται

με μεθοδολογίες όπως το clustering και τα HMM [93, 189]. Μεθοδολογίες όπως αυτές έχουν ως σκοπό την ταξινόμηση/κατηγοριοποίηση και κατάτμηση των ηχητικών σημάτων, η οποία έχει ως αποτέλεσμα την ανίχνευση των επαναλήψεων και άρα την εύρεση των σημαντικών φράσεων. Διαφορετικές προσεγγίσεις αποσκοπούν στην ανάλυση της δομής των μουσικών σημάτων, εξάγοντας χαρακτηριστικά όπως ο ρυθμός, η αρμονική δομή κ.ά. [158], ή στην εύρεση ομοιοτήτων με τη χρήση μετρικών ομοιότητας (*similarity matrix*) [26].

Τα μοντέλα που βασίζονται στην «έκπληξη» και στη σημαντικότητα της μουσικής και του ήχου γενικότερα, π.χ. με την έννοια των δυναμικών εναλλαγών [61], για την αυτόματη εξαγωγή μουσικών αποσπασμάτων (*music snippets*) [94] είναι περιορισμένα. Παρ' όλα αυτά, τα μοντέλα προσοχής έχουν χρησιμοποιηθεί εκτενώς για τη δημιουργία σύνοψης από δεδομένα βίντεο, όπου σε κάθε καρέ μιας ακολουθίας βίντεο ορίζεται μια τιμή προσοχής, ανάλογα με την προσοχή του θεατή [97]. Η σημαντικότητα του ήχου σε αυτή την περίπτωση ορίζεται με βάση τα χαρακτηριστικά ενέργειας – θεωρείται πως η ένταση προσελκύει την προσοχή των ανθρώπων. Στο [187] επιχειρείται η δημιουργία περιλήψεων βίντεο, όπου η ανάλυση της ηχητικής ροής προσεγγίζεται μέσω της κατάτμησης και της ταξινόμησης των ηχητικών γεγονότων, προκειμένου να εξαχθούν τα όρια των τμημάτων, χρησιμοποιώντας χαρακτηριστικά, όπως τα MFCC. Τέλος, η μέθοδος K-means εφαρμόζεται για την επιλογή των τμημάτων που συμπεριλαμβάνονται στην περίληψη.

Στην εργασία αυτή προσεγγίζουμε το θέμα της ανίχνευσης σημαντικών ηχητικών γεγονότων με τον υπολογισμό της σημαντικότητας (*saliency detection*), η οποία αποτελεί μετρικό της βαρύτητας του ήχου σε κάθε καρέ (*frame*) του βίντεο.

## 2.5 Βασεις Δεδομένων

Χρησιμοποιήσαμε μία βάση δεδομένων για την πειραματική αξιολόγηση στην αναγνώριση μουσικών οργάνων και δύο βάσεις για την αξιολόγηση της αναγνώρισης των ειδών μουσικής. Κάποιες από αυτές όπως για παράδειγμα η βάση του πανεπιστημίου της *«IOWA»* [181] με ηχητικά σήματα διαφορετικών οργάνων και η *«GTZAN»* [179] με αποσπάσματα διαφορετικών ειδών μουσικής είναι ιδιαίτερα διαδεδομένες. Αυτό τις καθιστά ιδιαίτερα χρήσιμες για τη διενέργεια συγκριτικών αξιολογήσεων με άλλες state-of-the-art μεθόδους της βιβλιογραφίας. Στον Πίνακα 2.1 παρουσιάζεται συνοπτική περίληψη των βάσεων που χρησιμοποιήθηκαν.

*Πίνακας 2.1:* Βάσεις Δεδομένων.

| Βάση | # Μουσικών Οργάνων | # Αρχείων | Ποσοστό (%) Μικρότερης και Μεγαλύτερης σε Πλήθος Αρχείων Κατηγορίας |
|---|---|---|---|
| IOWA | 18 | 3210 | ποικίλλει αναλόγως το μουσικό όργανο & τον τρόπο εκτέλεσης |
| Βάση | # Ειδών Μουσικής | # Αρχείων | |
| GTZAN | 10 | 1000 | 10% / 10% |
| 1517-Artists | 19 | 3180 | 3.96% / 5.88% |

**Βάση Δεδομένων για Αναγνώριση Μουσικών Οργάνων**

**IOWA**

Η βάση δεδομένων του πανεπιστημίου της *«IOWA»* [181] είναι μια ιδιαίτερα γνωστή βάση διαθέσιμη χωρίς κόστος για οποιαδήποτε χρήση[5]. Αποτελείται από ηχογραφήσεις 18 διαφορετικών μουσικών οργάνων, τα δεδομένα των οποίων έχουν συλλεχθεί υπό επαγγελματικές συνθήκες και αποτελούνται από ολόκληρο το εύρος συχνοτήτων, καλύπτοντας τις δυναμικές ενδείξεις piano, mezzoforte και forte. Τα μουσικά αρχεία έχουν συχνότητα δειγματοληψίας $fs$=44.1kHz, 16-bit, Mono, και τύπο αρχείου .wav. Για κάποια από τα μουσικά όργανα, π.χ. τα πνευστά και τα έγχορδα, υπάρχουν ηχογραφήσεις διαφορετικών τεχνικών όπως για παράδειγμα arco και pizzicato, βιμπράτο, και μη βιμπράτο. Στον Πίνακα 2.2 βλέπουμε λεπτομέρειες σχετικά με τα μουσικά όργανα και το πλήθος των αρχείων.

**Βάσεις Δεδομένων για Αναγνώριση Ειδών Μουσικής**

**GTZAN**

Το σύνολο δεδομένων *«GTZAN»*[6] δημιουργήθηκε από τους Tzanetakis et al. και χρησιμοποιήθηκε στο γνωστό άρθρο τους [179] του 2002. Έχει καθιερωθεί ως ένα από τα πιο σημαντικά σύνολα δεδομένων για την εφαρμογή της αναγνώρισης μουσικών ειδών και πλέον αποτελεί μία από τις παλαιότερες και πιο διαδεδομένες βάσεις, η οποία χρησιμοποιείται από την ερευνητική κοινότητα για λόγους σύγκρισης των αποτελεσμάτων διαφορετικών μεθοδολογιών. Αποτελείται από αποσπάσματα 1000 μουσικών κομματιών (30 δευτερόλεπτων) κατανεμημένα ομοιόμορφα σε 10 διαφορετικά μουσικά είδη. Τα αποσπάσματα αυτά έχουν συχνότητα δειγματοληψίας $fs$=22.05kHz, 16-bit, Mono και τύπο αρχείου .au. Τα αρχεία έχουν συλλεχθεί από διάφορες πηγές όπως

[5]http://theremin.music.uiowa.edu/MIS.html
[6]http://marsyas.info/download/data_sets/

*Πίνακας 2.2:* Λεπτομερής λίστα της βάσης IOWA.

| Μουσικά όργανα (18) | #Αρχείων (3210) |
|---|---|
| Alto Flute | 99 |
| Alto Saxophone | 192 |
| Upright Bass arco | 289 |
| Bass Clarinet | 138 |
| Bass Flute | 102 |
| Bassoon | 123 |
| Bass Trombone | 131 |
| B♭ Clarinet | 139 |
| Cello | 676 |
| E♭ Clarinet | 115 |
| Flute | 227 |
| French Horn | 97 |
| Oboe | 104 |
| Piano | 260 |
| Soprano Saxophone | 96 |
| Tenor Trombone | 99 |
| Trumpet | 212 |
| Tuba | 111 |

CD, ραδιόφωνο και ηχογραφήσεις υπολογιστή, ώστε να αντιπροσωπεύουν διαφορετικές συνθήκες εγγραφής ενώ η βάση δεν περιλαμβάνει καμία πληροφορία σχετική με τα μουσικά κομμάτια, όπως για παράδειγμα όνομα καλλιτέχνη, τίτλο ή άλμπουμ. Σημειώνουμε, πως η βάση έχει δεχθεί κριτική όσον αφορά την εγκυρότητα καθώς και την ποιότητα των μουσικών κομματιών, παρ' όλα αυτά εξακολουθεί να χρησιμοποιείται στην ερευνητική κοινότητα μιας και θεωρείται πως όλα τα συστήματα αναγνώρισης τα οποία αξιολογούνται με τη βάση αυτή αντιμετωπίζουν ακριβώς τα ίδια προβλήματα. Κάποια από τα προβλήματα που αναφέρονται αφορούν την ακεραιότητα των κατηγοριών, όπως για παράδειγμα την ύπαρξη των ίδιων ακριβώς κομματιών σε μία κατηγορία, την ύπαρξη του ίδιου καλλιτέχνη και άλμπουμ σε κάποια συγκεκριμένη κατηγορία, τη λανθασμένη κατηγοριοποίηση αλλά και την ποιότητά τους [168]. Στον Πίνακα 2.3 βλέπουμε λεπτομέρειες σχετικά με τα είδη και το πλήθος των αρχείων.

### 1517-Artists

Η βάση δεδομένων «*Artists*»[7] [156] (βλ. Πίνακα 2.4 για λεπτομέρειες) περιλαμβάνει αποσπάσματα 3180 γνωστών τραγουδιών (30 δευτερολέπτων), 1517 διαφορετικών καλλιτεχνών, τα οποία έχουν ταξινομηθεί σε 19 διαφορετικά είδη. Η ποιότητα της βάσης, σύμφωνα με τον δημιουργό της, έχει διασφαλιστεί επιλέγοντας τα 190 πιο δημοφιλή τραγούδια του κάθε είδους, σε σχέση με τον συνολικό αριθμό ακροάσεων, για το κάθε

[7] http://www.seyerlehner.info/index.php?p=1_3_Download

*Πίνακας 2.3:* Λεπτομερής λίστα της βάσης GTZAN.

| Μουσικά Είδη (10) | # Αρχείων (1000) |
|---|---|
| blues | 100 |
| classical | 100 |
| country | 100 |
| disco | 100 |
| hiphop | 100 |
| jazz | 100 |
| metal | 100 |
| pop | 100 |
| reggae | 100 |
| rock | 100 |

*Πίνακας 2.4:* Λεπτομερής λίστα της βάσης Artists.

| Μουσικά Είδη (19) | # Αρχείων (3115) |
|---|---|
| blues | 186 |
| country | 103 |
| hiphop | 87 |
| jazz | 103 |
| new age | 82 |
| reggae | 83 |
| classical | 46 |
| folk | 98 |
| latin | 86 |
| rock & pop | 117 |
| alternative & punk | 116 |
| electronic & dance | 92 |
| soul & r&b | 113 |
| world | 76 |
| religious | 71 |
| children's | 74 |
| easy listening & vocals | 98 |
| comedy & spoken word | 68 |
| soundtracks & more | 72 |

είδος. Τα αποσπάσματα αυτά έχουν συχνότητα δειγματοληψίας $fs$=44.1kHz, Stereo και τύπο αρχείου .mp3

## 2.6 Κίνητρα και Ερευνητικές Συνεισφορές

Σε αυτή την ενότητα παρουσιάζουμε συνοπτικά τα κίνητρα που μας ώθησαν στις συγκεκριμένες επιλογές για τη διεκπεραίωση της διδακτορικής αυτής έρευνας, καθώς και τις κύριες ερευνητικές συνεισφορές οι οποίες περιλαμβάνουν τις θεματικές περιοχές [41, 167, 197–200].

**Κίνητρα**

Τα κίνητρα για την ερευνητική αυτή μελέτη είναι πολλά. Βασίζονται τόσο σε ιδέες που διατυπώθηκαν από στοχαστές παλαιότερων χρόνων όσο και σε ενδείξεις μελετών νεότερων ερευνητών.

Λαμβάνοντας υπόψη τους ισχυρισμούς του Πλάτωνα και του Αριστοτέλη πως η μουσική είναι μιμητική ως προς τη φύση, τα ανθρώπινα συναισθήματα ή ακόμα και τις ιδιότητες των αντικειμένων, τις αποδείξεις του Mandelbrot [101] πως η φύση περιέχει δομές (π.χ., βουνά, ακτές, δομές των φυτών) οι οποίες θα μπορούσαν να περιγραφούν με τη θεωρία των φράκταλ, καθώς και περαιτέρω ενδείξεις πως η μουσική θα μπορούσε να ακολουθήσει παρόμοια μοντελοποίηση (βλ. Εν. 2.3.1)· εξετάζουμε τη δομή των μουσικών ήχων χρησιμοποιώντας ιδέες από τη θεωρία των φράκταλ, βασισμένοι επίσης στην έρευνα της ομάδας μας όπου παρόμοιες ιδέες έχουν διερευνηθεί σε σήματα φωνής και εφαρμογές ανάλυσης, κατηγοριοποίησης και αυτόματης αναγνώρισης [31, 79, 105, 133].

Επιπλέον, στηριζόμενοι στις ενδείξεις για την ύπαρξη μη-γραμμικών φαινομένων κατά την παραγωγή φωνής [104], καθώς και στις ενδείξεις για την ύπαρξη μικρο-διαμορφώσεων στα σήματα μουσικών σημάτων [18], χρησιμοποιούμε και επεκτείνουμε το AM-FM μοντέλο για την ανίχνευση των μικροδομών και των χαρακτηριστικών τους. Σημαντικές είναι επίσης οι ενδείξεις της χρησιμότητας του μοντέλου αυτού στην αναγνώριση και ανίχνευση φωνής [33, 40] καθώς και σε εφαρμογές αναγνώρισης και κατηγοριοποίησης της ακουστικής πληροφορίας των ηχητικών σημάτων [18, 24].

Τέλος, οι ενδείξεις για την παράλληλη εξέλιξη και τη σχέση της φωνής με τη μουσική, καθώς και η επικάλυψη των γνωσιακών μηχανισμών κατά την επεξεργασία των δομών τους, μας ωθούν στη διερεύνηση, επέκταση και προσαρμογή μεθοδολογιών και μοντέλων που με επιτυχία έχουν εφαρμοστεί στην επεξεργασία σημάτων φωνής, αλλά μέχρι τώρα όχι στην επεξεργασία μουσικής.

**Συνεισφορές**

Οι κύριες ερευνητικές μας συνεισφορές μπορούν να συνοψισθούν ως εξής:

1. *Επεξεργασία των μουσικών σημάτων διαφορετικών μουσικών οργάνων στο πλαίσιο της φράκταλ γεωμετρίας.* Ανάλυση της πολυπλοκότητας και της ανομοιογένειας των σημάτων μουσικής με μετρήσεις βασισμένες στη φράκταλ διάσταση. Προσδιορίζουμε το προφίλ των διαφορετικών μουσικών οργάνων με τη χρήση του αλγορίθμου *μορφολογικής κάλυψης* που βασίζεται στη διάσταση Minkowski σε πολλαπλές κλίμακες, και διερευνούμε τη φράκταλ διάσταση για τις διαφορετικές καταστάσεις των σημάτων, π.χ., attack και σταθερή κατάσταση. Τέλος, εξάγουμε εύρωστες και χρήσιμες αναπαραστάσεις τις οποίες αξιολογούμε σε πειράματα

κατηγοριοποίησης μουσικών οργάνων.

2. *Ανάλυση των μουσικών σημάτων διαφορετικών μουσικών οργάνων με μη-γραμμικά μοντέλα διαμορφώσεων* για την εξαγωγή χαρακτηριστικών και τη μοντελοποίηση των μικροδομών των ήχων των μουσικών οργάνων. Διερευνούμε και επεκτείνουμε τον αλγόριθμο ESA ενώ επιπλέον εφαρμόζουμε τον Επαναληπτικό-ESA όπου και εξετάζουμε τις δυνατότητές του για καλύτερες εκτιμήσεις των χαρακτηριστικών καθώς και πιο εύστοχο προσδιορισμό των δομών και χαρακτηριστικών των μουσικών σημάτων.

3. *Πειραματική αξιολόγηση των προτεινόμενων μεθόδων και χαρακτηριστικών για την εφαρμογή αναγνώρισης μουσικών οργάνων* με τη χρήση Γκαουσιανών (GMM) και κρυφών Μαρκοβιανών μοντέλων (HMM), σε σχέση με βασικά χαρακτηριστικά, όπως τα MFCC, τα οποία επιλέχθηκαν για την καλή τους απόδοση αλλά και την αποδοχή τους σε παρόμοιες μελέτες. Η αξιολόγηση αυτή καταδεικνύει τη δυναμική των προτεινόμενων μεθόδων στην συγκεκριμένη εφαρμογή.

4. *Ανάλυση των μουσικών σημάτων διαφορετικών ειδών μουσικής με μη-γραμμικά μοντέλα διαμορφώσεων* για την εξαγωγή χαρακτηριστικών και τη μοντελοποίηση των μικροδομών και των μακροδομών τους. Βάσει της αξιολόγησης του αλγόριθμου ESA και της γνώσης που αποκτήθηκε από την ανάλυση μουσικών σημάτων διαφορετικών μουσικών οργάνων, εξετάζουμε και επεκτείνουμε τον αλγόριθμο για τη δημιουργία αναπαραστάσεων ικανών στην ταξινόμηση των διαφορετικών ειδών μουσικής. Προτείνουμε τη δημιουργία συστοιχίας φίλτρων εστιασμένων στα μουσικά σήματα, για την ανάπτυξη πιο εύρωστων χαρακτηριστικών. Παράλληλα, διερευνούμε εναλλακτικές μορφές αναπαραστάσεων των χαρακτηριστικών βασισμένες στη μακροδομή των σημάτων μουσικής. Αξιολογούμε τον συνδυασμό των AM-FM χαρακτηριστικών με μετρήσεις βασισμένες στη φράκταλ διάσταση σε σχέση με τα MFCC, με χρήση των κρυφών Μαρκοβιανών μοντέλων.

   Τα αποτελέσματα παρουσιάζονται ιδιαίτερα ενθαρρυντικά καταδεικνύοντας πως τα προτεινόμενα χαρακτηριστικά δύνανται να περιγράψουν σημαντικά φαινόμενα των μουσικών σημάτων όπως για παράδειγμα τις μικρο-μεταβολές των δομών τους ενώ αναπαραστάσεις που βασίζονται στις μακροδομές των σημάτων επιφέρουν μείωση της πολυπλοκότητας του συστήματος κατηγοριοποίησης. Τέλος, η εισαγωγή ιδεών όπως η «μουσική» συστοιχία φίλτρων επιδεικνύει ιδιαίτερη διακριτική ικανότητα στην κατηγοριοποίηση των μουσικών σημάτων.

5. *Δημιουργία αναπαραστάσεων Bag-of-Words για την ανάλυση των μουσικών σημάτων διαφορετικών ειδών μουσικής.* Διατυπώνουμε τη διαδικασία εξαγωγής

χαρακτηριστικών μέσω του μοντέλου Bag-of-Words καθιστώντας δυνατή την εισαγωγή εναλλακτικών αναπαραστάσεων των μουσικών σημάτων, που αναπαριστούν πιθανές φράσεις ή μοτίβα του μουσικού κομματιού. Ακολουθώντας την προσέγγιση αυτή επιτυγχάνεται η δημιουργία ενός *«μουσικού λεξικού»*, το οποίο χρησιμοποιείται για τη διάκριση των γνωρισμάτων των διαφορετικών ειδών της μουσικής. Τα χαρακτηριστικά που τελικά εξάγονται αξιολογούνται με τη χρήση των Support Vector Machines (SVMs) με αξιόλογα αποτελέσματα. Με τη μέθοδο αυτή αντιμετωπίζονται διάφορα προβλήματα πολυπλοκότητας κατά την κατηγοριοποίηση, λόγω των νέων συμπαγών αναπαραστάσεων.

6. *Μελέτη των μη-γραμμικών μοντέλων διαμορφώσεων για την ανίχνευση σημαντικών ακουστικών γεγονότων.* Βασιζόμενοι σε προηγούμενη ερευνητική εργασία μας [42] εξετάζουμε την καταλληλότητα των AM-FM χαρακτηριστικών, προτείνουμε υπολογιστικές μεθόδους σύμμειξής τους για τη δημιουργία αναπαραστάσεων σημαντικότητας (*saliency*), οι οποίες και αποτελούν το κριτήριο επιλογής αντιληπτικά σημαντικών ηχητικών γεγονότων για τη δημιουργία συνοπτικών ηχητικών αποσπασμάτων. Παράλληλα, επεκτείνουμε και βελτιώνουμε τον αλγόριθμο δημιουργίας συνόψεων με αποτέλεσμα αντιληπτικά ποιοτικές περιλήψεις. Ελέγχουμε τους προτεινόμενους αλγόριθμους για εξαγωγή γεγονότων καθώς και τις τελικές περιλήψεις με διεξοδικές ποσοτικές αξιολογήσεις.

   Τέλος, επεκτείνουμε και αξιολογούμε τις προτεινόμενες μεθόδους και ιδέες μας και σε πολυμεσικά δεδομένα. Οι κατευθύνσεις με τις οποίες ασχολούμαστε αφορούν διεξοδικές ποσοτικές αξιολογήσεις του αλγόριθμου δημιουργίας περιλήψεων, έλεγχο της αποδοτικότητας των αυτόματων περιλήψεων και διερεύνηση καινούριων χαρακτηριστικών καθώς και μεθόδων σύμμειξής τους.

7. *Ανάπτυξη συστηματικής βάσης δεδομένων από ταινίες και ταξιδιωτικά ντοκιμαντέρ*, η οποία επισημειώνεται με τη μονοτροπική και την πολυτροπική σημαντικότητα του βίντεο καθώς και με τη σημασιολογική πληροφορία και αποτελεί σημαντικό εργαλείο για τις αξιολογήσεις των αλγορίθμων ανίχνευσης ηχητικών ή/και πολυμεσικών γεγονότων και των ηχητικών και πολυμεσικών περιλήψεων από βίντεο.

**Δομή της εργασίας**

Στις επόμενες ενότητες παρουσιάζονται τα αποτελέσματα της έως τώρα έρευνάς μας. Στο Κεφάλαιο 3 παρουσιάζουμε την ανάλυση των μουσικών σημάτων διαφορετικών μουσικών οργάνων, βάσει της φράκταλ διάστασης σε πολλαπλές κλίμακες και τα πειραματικά αποτελέσματα κατηγοριοποίησης. Στη συνέχεια εφαρμόζουμε το AM-FM μοντέλο διαμορφώσεων, Κεφάλαιο 4, και παρουσιάζουμε πειραματικά αποτελέσματα

αρχικά σε δύο σετ πειραμάτων για την αναγνώριση των οργάνων και στη συνέχεια για την εφαρμογή της αναγνώρισης των διαφορετικών ειδών της μουσικής. Επιπλέον, εξετάζουμε και αξιολογούμε τη μέθοδο Bag-of-Words χρησιμοποιώντας ως αναπαραστάσεις τα χαρακτηριστικά που έχουν προκύψει από την προηγούμενη ανάλυση. Στο Κεφάλαιο 5 περιγράφουμε την εργασία μας για την ανίχνευση σημαντικών ακουστικών γεγονότων, και αναλύουμε περαιτέρω το είδος των εξαχθέντων γεγονότων. Επιπλέον, περιγράφουμε τη διαδικασία δημιουργίας της βάσης δεδομένων ταινιών και τη συνεισφορά μας στο θέμα της ανίχνευσης πολυμεσικών γεγονότων. Στο Κεφάλαιο 6 συνοψίζουμε τις ερευνητικές μας συνεισφορές και προτείνουμε τις μελλοντικές κατευθύνσεις της συγκεκριμένης έρευνας.

# Κεφάλαιο 3

# Ανάλυση και Μοντελοποίηση Μουσικών Σημάτων σε Πολλαπλές Κλίμακες με Φράκταλ Μεθόδους

Σε αυτό το κεφάλαιο διερευνούμε μη-γραμμικές μεθόδους, εμπνευσμένες από τη θεωρία των Φράκταλ για την ανάλυση της δομής των μουσικών σημάτων σε πολλαπλές κλίμακες. Σε πολλές εφαρμογές η κατηγοριοποίηση των οργάνων στο επίπεδο της οικογένειας θα μπορούσε να είναι επαρκής. Ωστόσο, στην προσέγγισή μας επικεντρωνόμαστε στη διάκριση μεταξύ των διαφορετικών οργάνων, υποδεικνύοντας ομοιότητες που παρατηρούνται στις διάφορες οικογένειες. Προτείνουμε τη βραχέος χρόνου φράκταλ διάσταση σε πολλαπλές κλίμακες (MFD) για την περιγραφή των μουσικών σημάτων, θεωρώντας πως θα φανεί χρήσιμη για την ποσοτικοποίηση της πολυπλοκότητας των διαφορετικών καταστάσεων της κυματομορφής τους, αναλύοντας όχι μόνο τη σταθερή κατάσταση των ήχων αλλά και το attack τους.

## 3.1 Φράκταλ Διάσταση σε Πολλαπλές Κλίμακες (MFD)

Τα περισσότερα χαρακτηριστικά που εξάγονται από σήματα μουσικής με σκοπό την κατηγοριοποίηση είναι εμπνευσμένα από αντίστοιχες μελέτες στη φωνή, το ίδιο και η φράκταλ διάσταση σε πολλαπλές κλίμακες (*Multiscale Fractal Dimension, MFD*) που χρησιμοποιείται σε αυτή την έρευνα. Έχει διατυπωθεί πως οι ήχοι της φωνής σε κάποιες χρονικές κλίμακες παρουσιάζουν ποσότητες τύρβης (*turbulence*). Ο Mandelbrot [101] ισχυρίστηκε πως οι δομές της τυρβώδους ροής αυτού του είδους μπορούν να μοντελοποιηθούν με τη θεωρία των φράκταλ.

Ιδέες σαν και αυτήν προέτρεψαν τον Maragos [102] να χρησιμοποιήσει τη βραχέος χρόνου φράκταλ διάσταση ως χαρακτηριστικό για την ποσοτικοποίηση της τυρβώδους

ροής των ήχων της φωνής, με τη χρήση ενός αποτελεσματικού αλγορίθμου [102, 103] ο οποίος βασίζεται στη Minkowski-Bouligand διάσταση $D$ [43, 103]. Η διάσταση αυτή υπολογίζει το μήκος κάποιου συνόλου $F$ (πιθανώς κατακερματισμένου) σε πολλαπλές κλίμακες, με τη δημιουργία ενός «καλύμματος Minkowski» (*Minkowski cover*). Με άλλα λόγια, το σύνολο $F$ καλύπτεται με δίσκους ακτίνας $s$, οι οποίοι τοποθετούνται ως προς το κέντρο τους σε κάθε σημείο του $F$. Ο αλγόριθμος αναφέρεται και ως *μορφολογική μέθοδος κάλυψης* (*morphological covering method*) και τα βήματα που ακολουθούνται είναι τα εξής:

*Βήμα 1.* Δημιουργία του καλύμματος Minkowski, με τη χρήση *δισδιάστατων* μορφολογικών τελεστών – δηλαδή, του μορφολογικού τελεστή της διαστολής $\oplus$ (γνωστό και ως άθροισμα Minkowski). Η διαστολή του συνόλου $F$ (που παριστάνει τον γράφο του σήματος) πραγματοποιείται με ένα κυρτό, συμμετρικό επίπεδο σύνολο $B$ αυξανόμενης ακτίνας $sB = \{sb : b \in B\}$, όπου $s \geq 0$ η παράμετρος της κλίμακας:

$$F \oplus sB = \{z + sb \in \mathbb{R}^2 : z \in F, b \in B\}. \tag{3.1}$$

Στη συνέχεια, υπολογίζουμε το εμβαδόν $A_B(s) = \text{area}(F \oplus sB)$ του διεσταλμένου γράφου σε πολλαπλές κλίμακες. Τέλος, το ακόλουθο όριο της καλυμμένης περιοχής σε log-log κλίμακα υπολογίζει τη φράκταλ διάσταση:

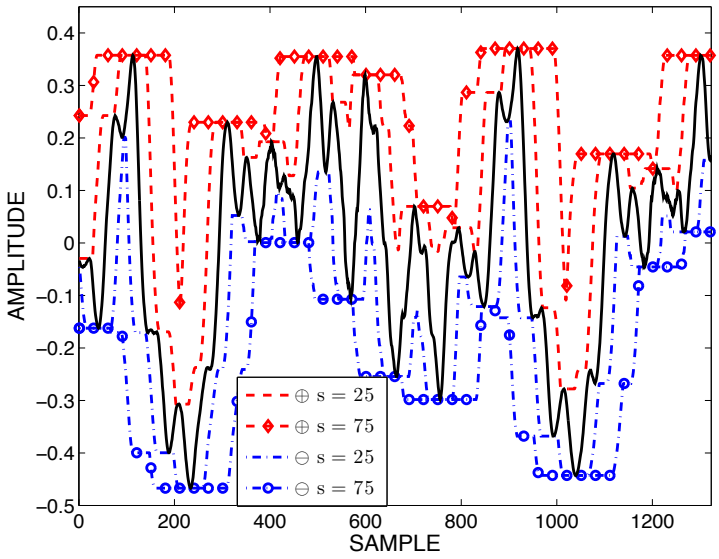
$$D = \lim_{s \to 0} \frac{\log[A_B(s)/s^2]}{\log(1/s)}. \tag{3.2}$$

Ιδανικά το σύνολο $B$ είναι ένας μοναδιαίος δίσκος. Ωστόσο, η διάσταση $D$ παραμένει αμετάβλητη όσο το $B$ είναι συμπαγές, κυρτό και συμμετρικό [103]. Στην περίπτωση διακριτών σημάτων επιλέγουμε ως $B$ μία προσέγγιση του δίσκου, και άρα ένα μοναδιαίας ακτίνας, κυρτό και συμμετρικό υποσύνολο του $\mathbb{Z}^2$.

*Βήμα 2.* Έχει καταδειχθεί [102, 103] πως το παραπάνω όριο για τον υπολογισμό της διάστασης $D$ δεν θα αλλάξει, αν προσεγγίσουμε το $A_B(s)$ με το εμβαδόν της διαφοράς μεταξύ του μορφολογικού τελεστή διαστολής (*dilation*) $\oplus$ και διάβρωσης (*erosion*) $\ominus$ του $N$-δειγμάτων διακριτού σήματος $F[n]$ από μία κοίλη και συμμετρική συνάρτηση $G_s[n]$:

$$A_B(s) = \sum_{n=0}^{N-1} ((F \oplus G_s) - (F \ominus G_s))[n], \tag{3.3}$$

για $s = 1, \ldots, s_{max} \leq \frac{N}{2}$. Η προσέγγιση αυτή μειώνει σημαντικά την πολυπλοκότητα, καθώς οι δισδιάστατοι τελεστές αντικαθίστανται από μονοδιάστατους, οι οποίοι είναι ισοδύναμοι με απλές μη-γραμμικές συνελίξεις. Περαιτέρω μείωση της πολυπλοκότητας [$O(N^2)$ σε $O(N)$] επιτυγχάνεται αν εκτελέσουμε την παραπάνω διαδικασία αναδρομικά (*scale-recursive*):



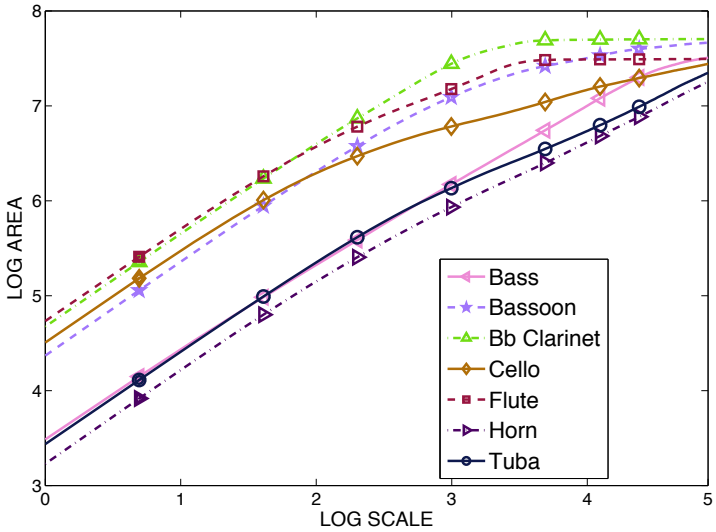

*Σχήμα 3.1:* Σταθερή κατάσταση ηχητικού σήματος του κοντραμπάσου (μαύρη ενιαία γραμμή) και τα dilation $\oplus$ και erosion $\ominus$ για τις κλίμακες $s$ = 25, 75.

$$
\begin{aligned}
F \oplus G[n] &= \max_{-1 \le k \le 1} \{F[n+k] + G[k]\},\ s = 1 \\
F \ominus G[n] &= \min_{-1 \le k \le 1} \{F[n+k] - G[k]\},\ s = 1 \\
F \oplus G_{s+1} &= (F \oplus G_s) \oplus G,\ 2 \le s \le s_{max} \\
F \ominus G_{s+1} &= (F \ominus G_s) \ominus G,\ 2 \le s \le s_{max}.
\end{aligned}
\tag{3.4}
$$

όπου $G = G_1$ και $s = 1, \ldots, s_{max} \le \frac{N}{2}$.

Οι διαστολές και οι διαβρώσεις του σήματος, έτσι όπως υπολογίζονται στην περίπτωση αυτή, έχουν παρόμοια υπολογιστική δομή (αλλά ταχύτερη) με τη γραμμική συνέλιξη και τη συσχέτιση, αντίστοιχα [103]. Δημιουργούν δε, την περιοχή του εμβαδού ως ένα κάλυμμα που είτε καλύπτει, είτε αποκολλάται από τον γράφο του σήματος σε διάφορες κλίμακες. Το Σχήμα 3.1 παρουσιάζει μια ειδική περίπτωση όπου το $B$ είναι ένα οριζόντιο συμμετρικό τμήμα τριών δειγμάτων με μηδενικό ύψος, πράγμα που σημαίνει ότι η συνάρτηση $G[n]$ ισούται με μηδέν για $n = -1, 0, 1$ και $-\infty$ αλλού. Η ειδική αυτή περίπτωση κάνει τον αλγόριθμο ακόμα πιο γρήγορο και αποδοτικό, επειδή οι αντίστοιχες διαστολές και διαβρώσεις ταυτίζονται με το τοπικό μέγιστο (*max*) και ελάχιστο (*min*) εντός ενός μετακινούμενου παραθύρου. Επιπλέον, η φράκταλ διάσταση στις διαφορετικές κλίμακες είναι αμετάβλητη σε οποιοδήποτε αφινικό (*affine*) μετασχηματισμό του σήματος.

*Βήμα 3.* Στην πράξη, το $D$ μπορεί να εκτιμηθεί από τη μέτρηση της κλίσης μιας ευθείας η οποία προσεγγίζει τα δεδομένα $\log[A_B(s)]$ έναντι $\log(s)$ με τη μέθοδο των ελαχίστων τετραγώνων,




*Σχήμα 3.2:* $\log[A_B(s)]$ έναντι $\log(s)$ για τα εφτά όργανα που εξετάζουμε και τη νότα C3. Για το B♭ κλαρινέτο και το φλάουτο δείχνουμε την κλίση για τη νότα C5 (και πλαίσια ανάλυσης των 30 ms).

$$\log[A_B(s)] = (2 - D)\log(s) + \text{constant}, \quad \text{as } s \to 0 \tag{3.5}$$

υποθέτοντας ότι το εμβαδόν $A_B(s) \approx s^{2-D}$ όταν η κλίμακα $s \to 0$. Ωστόσο, τα πραγματικά σήματα δεν έχουν την ίδια δομή σε όλες τις κλίμακες, επομένως η τιμή του εκθέτη στη δύναμη $s^{2-D}$ δύναται να ποικίλλει. Άρα υπολογίζουμε την κλίση των δεδομένων $\log[A_B(s)]$ έναντι $\log(s)$ τοπικά, επί μικρών ολισθούμενων παραθύρων κλίμακας $\{s, s+1, ...s+w\}$, όπου $w$ το εύρος του παραθύρου κλίμακας, τα οποία κινούνται κατά μήκος του άξονα των κλιμάκων $s$. Οι διαδοχικές αυτές εκτιμήσεις της διάστασης δημιουργούν ένα προφίλ της τοπικής *φράκταλ διάστασης σε πολλαπλές κλίμακες (Multiscale Fractal Dimension, MFD)* $D[s,t]$ σε κάθε χρονική στιγμή $t$ του βραχέος χρόνου παραθύρου ανάλυσης του σήματος. Η τοπική κλίση της ευθείας αυτής είναι η εκτίμηση του $2-D$ που μας δίνει τη φράκταλ διάσταση. Στα πειράματα που ακολουθούν έχουμε χρησιμοποιήσει $w = 10$. Το Σχήμα 3.2 δείχνει τον γράφο του $\log[A_B(s)]$ έναντι $\log(s)$ για διάφορα μουσικά όργανα, όπου βλέπουμε τη διαφορά της κλίσης για μεγαλύτερες κλίμακες $s$. Επιπλέον, η διάσταση $D$ κυμαίνεται μεταξύ 1 και 2 για τοπολογικά μονοδιάστατα σήματα (δηλαδή για συνεχείς συναρτήσεις μιας μεταβλητής). Όσο μεγαλύτερη είναι η διάσταση $D$ τόσο μεγαλύτερος είναι ο γεωμετρικός κατακερματισμός του γράφου του σήματος. Η διάσταση $D$ στη μικρότερη διακριτή χρονική κλίμακα εκτιμάται ως χαρακτηριστικό βραχέος χρόνου για εφαρμογές κατάτμησης του ηχητικού σήματος και ανίχνευση γεγονότων. Τέλος, η συνάρτηση $D[s,t]$ ονομάζεται επίσης *φρακτόγραμμα* (*fractogram*) και δύναται να παρέχει πληροφορίες για τον βαθμό ανατάραξης που μπορεί να υπάρχει στις πολλαπλές κλίμακες των βραχέος χρόνου ήχων [102, 105].

Ο συγκεκριμένος αλγόριθμος είναι επίσης σημαντικός επειδή έχει γραμμική πολυπλοκότητα – $O(N)$ προσθέσεις – υποθέτοντας ένα $N$-δειγμάτων σήμα, δεδομένου ότι οι απαιτούμενες min-max πράξεις είναι υπολογιστικά ισοδύναμες με προσθέσεις. Συγκριτικά με τα MFCC – $O(N \log N)$ πολλαπλασιασμοί – τα οποία χρησιμοποιούνται καθ' όλη την πειραματική αξιολόγηση, βλέπουμε ότι η χρήση των MFD είναι προτιμότερη, δεδομένης της μικρότερης υπολογιστικής πολυπλοκότητας.

Σε γενικές γραμμές, η βραχέος χρόνου φράκταλ διάσταση στη μικρότερη διακριτή κλίμακα ($s = 1$) μπορεί να βοηθήσει στη διάκριση ήχων διαφορετικών κατηγοριών, ενώ σε υψηλότερες κλίμακες το MFD προφίλ μπορεί να προσφέρει περαιτέρω πληροφορίες που θα ενισχύσουν και θα βελτιώσουν την κατηγοριοποίησή τους. Ωστόσο, από την έρευνα των [105] και [133] έχουν προκύψει στοιχεία που αποδεικνύουν ότι χαρακτηριστικά όπως η MFD διάσταση (σε συνδυασμό με άλλα) μπορούν να επιφέρουν βελτίωση σε συστήματα αναγνώρισης λέξεων για τις βασικές βάσεις δεδομένων φωνής. Στη συγκεκριμένη μελέτη χρησιμοποιούμε τα προφίλ της MFD διάστασης ως εργαλείο για την ανάλυση της δομής των σημάτων μουσικής σε πολλαπλές κλίμακες.

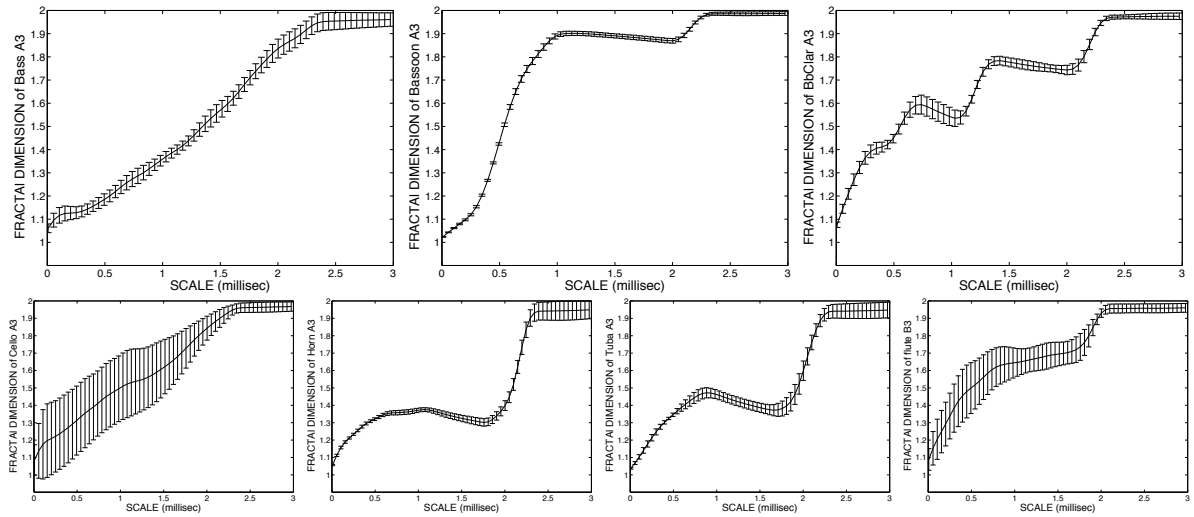
## 3.2 Ανάλυση των MFD σε Σήματα Μουσικών Οργάνων

### 3.2.1 MFD στη Σταθερή Κατάσταση

Η ανάλυσή μας βασίζεται τόσο στη διάκριση των διαφορετικών οργάνων όσο και στη διερεύνηση των διαφορών μεταξύ του attack και της σταθερής κατάσταση των ήχων. Σκοπός μας είναι να δείξουμε ότι η φράκταλ διάσταση του attack διαφοροποιείται αρκετά στα διαφορετικά όργανα, με αποτέλεσμα την ύπαρξη περισσότερων χρήσιμων πληροφοριών σε εφαρμογές αναγνώρισης.

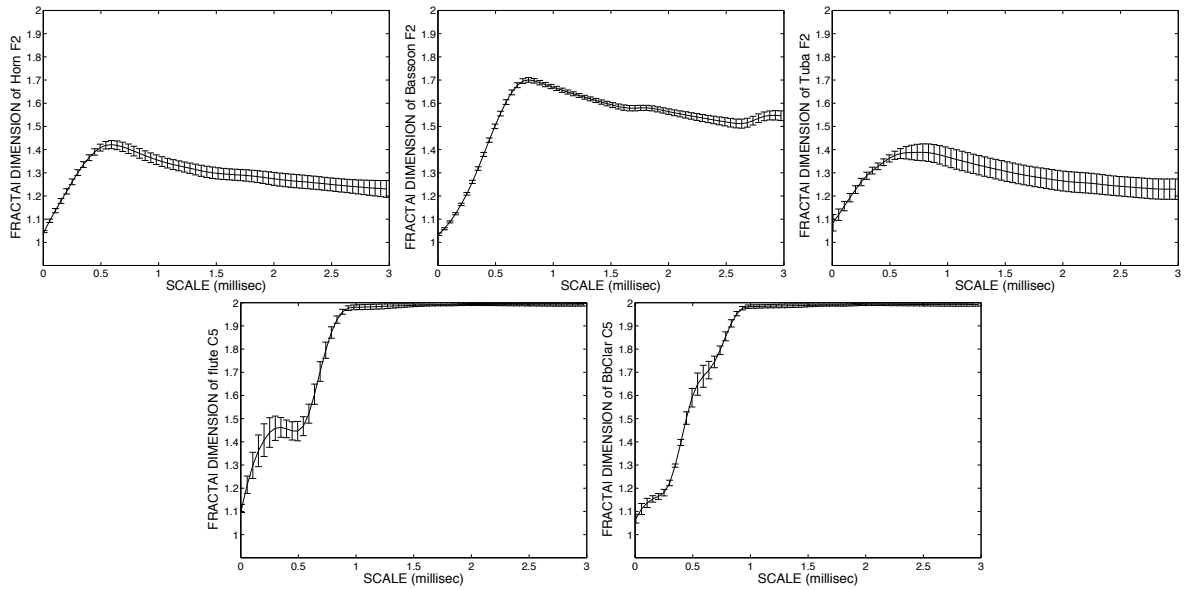
Για την ανάλυση της σταθερής κατάστασης χρησιμοποιήθηκε το σύνολο των ήχων από τα ακόλουθα όργανα: κοντραμπάσο, φαγκότο, B♭ κλαρινέτο, βιολοντσέλο, φλάουτο, γαλλικό κόρνο και τούμπα. Τα βραχέος χρόνου MFD υπολογίστηκαν σε τμήματα των 30 ms με επικάλυψη 15 ms για όλη τη διάρκεια των ήχων. Ωστόσο, για την ανάλυση που ακολουθεί έχουμε επεξεργαστεί μόνο τη σταθερή κατάσταση των σημάτων. Τα σήματα έχουν συχνότητα δειγματοληψίας 44.1 kHz και τα αντίστοιχα MFD[$s$] προφίλ αναλύθηκαν για διακριτές κλίμακες $s = 1, \ldots, 133$, που αντιστοιχούν σε χρονικές κλίμακες $s_t$ από 1/(44.1) έως 3 ms. Παρόμοια αποτελέσματα λήφθηκαν από την ανάλυση με χρονικά παράθυρα των 50 ms.

Το Σχήμα 3.3 δείχνει τη μέση τιμή του MFD και την τυπική απόκλιση (error bars) για τη νότα A3 για όλα τα όργανα εκτός του φλάουτου, που παρουσιάζεται για τη νότα B3. Το MFD προφίλ που παρουσιάζουμε είναι χαρακτηριστικό για τις ακόλουθες οκτάβες




*Σχήμα 3.3:* Μέση τιμή του MFD προφίλ (μεσαία γραμμή) και τυπική απόκλιση (error bars) για τη νότα A3 για τα όργανα μπάσο, φαγκότο, B♭ κλαρινέτο (πρώτη γραμμή) και τσέλο, κόρνο και τούμπα, και τη νότα B3 για το φλάουτο (δεύτερη γραμμή) (για παράθυρα ανάλυσης 30 ms, με 15 ms επικάλυψη).

κάθε οργάνου (βλ. Σχήμα 2.3 για το εύρος των συχνοτήτων των οργάνων και την επικάλυψή τους): κοντραμπάσο για όλο το εύρος, φαγκότο και κόρνο για τις οκτάβες 3–5, B♭ κλαρινέτο και φλάουτο για τις οκτάβες 3–4, και βιολοντσέλο για τις οκτάβες 2–4. Το Σχήμα 3.4 δείχνει τα MFD προφίλ για τις χαμηλότερες οκτάβες του φαγκότο, της τούμπας και του κόρνου (οκτάβες 1–2), όπου παρατηρούνται ορισμένες ομοιότητες: παρουσιάζουν την πρώτη τους κορυφή και υψηλότερη τιμή $D$ σε κλίμακα $s_t = 0.5$, ενώ στη συνέχεια η διάσταση $D$ μειώνεται σε μια ενδιάμεση τιμή. Παρόλα αυτά, εμφανίζουν κάποιες σημαντικές διαφορές· η μέγιστη διάσταση $D$ είναι περίπου στο $1.8$ για το φαγκότο, ενώ η τούμπα και το κόρνο έχουν $D = 1.5$. Επιπλέον, η διάσταση $D$ στην τούμπα παρουσιάζει πιο σημαντικές αποκλίσεις μεταξύ των διαφορετικών πλαισίων ανάλυσης της νότας σε σχέση με το κόρνο. Στις υψηλότερες οκτάβες του B♭ κλαρινέτου και του φλάουτο (οκτάβες 5–6) (βλ. Σχήμα 3.4 δεύτερη σειρά), παρατηρούμε πως τα MFD προφίλ για τις συχνοτικές αυτές περιοχές παρουσιάζουν τη μέγιστη τιμή $D$ περίπου $1.9$ σε μικρές χρονικές κλίμακες $s_t = 0.8$ και τη διατηρούν έως τη μέγιστη κλίμακα. Το κοντραμπάσο και το τσέλο έχουν πιο ομοιόμορφα MFD προφίλ με κάπως αυξημένη απόκλιση του $D$ μεταξύ των διαδοχικών πλαισίων ανάλυσης για τόνους χαμηλότερης συχνότητας. Τέλος, εκτός από τις δύο τελευταίες περιπτώσεις, τα υπόλοιπα μουσικά όργανα παρουσιάζουν συγκεκριμένες διαφορές ανάμεσα στις χαμηλότερες και στις υψηλότερες οκτάβες, με αναλλοίωτα δε χαρακτηριστικά για τις συγκεκριμένες οκτάβες. Στον Πίνακα 3.1 βλέπουμε τη μέση τιμή του MFD των οργάνων για τη σταθερή κατάσταση και για το συνολικό εύρος συχνοτήτων (δυναμική *forte*) σε συγκεκριμένες χρονικές κλίμακες $s_t$ τις οποίες θεωρήσαμε κομβικά σημεία μετά την ανάλυση που προηγήθηκε. Στην παρένθεση φαίνεται η τυπική




*Σχήμα 3.4:* Μέση τιμή του MFD προφίλ και τυπική απόκλιση για τη νότα F2 για τα όργανα κόρνο, φαγκότο και τούμπα (πρώτη γραμμή) και τη νότα C5 για το φλάουτο και το B♭ κλαρινέτο (δεύτερη γραμμή). Τα MFD προφίλ που παρουσιάζονται είναι αντιπροσωπευτικά για τις χαμηλότερες οκτάβες των τριών οργάνων της πρώτης γραμμής, και για τις υψηλότερες οκτάβες των οργάνων της δεύτερης γραμμής (για πλαίσια ανάλυσης 30 ms με 15 ms επικάλυψη).

*Πίνακας 3.1:* Μέση τιμή του MFD και τυπική απόκλιση για διάφορα σημεία της κλίμακας $s_t$ των MFD προφίλ.

| | Μέση τιμή MFD (std) | | | | | |
|---|---|---|---|---|---|---|
| Κλίμακες (ms) | $s_t = 1/44$ (ms) | $s_t = 0.5$ | $s_t = 1$ | $s_t = 1.5$ | $s_t = 2$ | $s_t = 2.5$ |
| Μπάσο | 1.11 (0.050) | 1.21 (0.037) | 1.31 (0.040) | 1.39 (0.040) | 1.52 (0.039) | 1.61 (0.038) |
| Φαγκότο | 1.04 (0.004) | 1.47 (0.006) | 1.75 (0.070) | 1.78 (0.080) | 1.80 (0.090) | 1.83 (0.010) |
| Τσέλο | 1.12 (0.017) | 1.47 (0.066) | 1.63 (0.076) | 1.73 (0.077) | 1.80 (0.067) | 1.85 (0.058) |
| Κλαρινέτο | 1.14 (0.035) | 1.69 (0.033) | 1.84 (0.035) | 1.90 (0.027) | 1.95 (0.021) | 1.96 (0.017) |
| Φλάουτο | 1.13 (0.018) | 1.77 (0.036) | 1.90 (0.037) | 1.95 (0.021) | 1.98 (0.010) | 1.98 (0.010) |
| Κόρνο | 1.06 (0.002) | 1.38 (0.006) | 1.49 (0.009) | 1.54 (0.019) | 1.59 (0.022) | 1.64 (0.024) |
| Τούμπα | 1.10 (0.026) | 1.35 (0.013) | 1.40 (0.120) | 1.36 (0.015) | 1.38 (0.017) | 1.42 (0.022) |

απόκλιση, η οποία δείχνει τη διακύμανση για τις συγκεκριμένες κλίμακες. Για τις μετρήσεις αυτές δεν λάβαμε υπόψη τη μεταβλητότητα των MFD στις διαφορετικές οκτάβες. Τα πιο ομοιογενή με μικρότερη μεταβλητότητα MFD προφίλ παρατηρούνται στο κόρνο, στην τούμπα και στο φαγκότο για μικρότερες κλίμακες, και για στο B♭ κλαρινέτο και το φλάουτο για μεγαλύτερες κλίμακες.

Η έως τώρα ανάλυση της φράκταλ διάστασης σε πολλαπλές κλίμακες στη σταθερή κατάσταση των μουσικών ήχων επιβεβαιώνει τους αρχικούς ισχυρισμούς μας, ότι τα MFD παρέχουν πληροφορίες σχετικά με τη διαφορετικότητα των μουσικών οργάνων. Ακόμη και στις περιπτώσεις οργάνων που ανήκουν στην ίδια οικογένεια ή στο ίδιο εύρος συχνοτήτων και παρουσιάζουν παραπλήσια χαρακτηριστικά, εντοπίζονται συγκεκριμένες διαφοροποιήσεις στη διάσταση $D$, στην κλίμακα $s_t$, ή στη μεταβλητότητα της διάστασης $D$ στις διαφορετικές κλίμακες. Τέλος, διακρίνουμε την εξάρτηση του MFD προφίλ

από την ακουστική συχνότητα των ήχων, κάτι που πραγματευόμαστε περαιτέρω στην Ενότητα 3.2.3.

### 3.2.2 MFD στο Attack

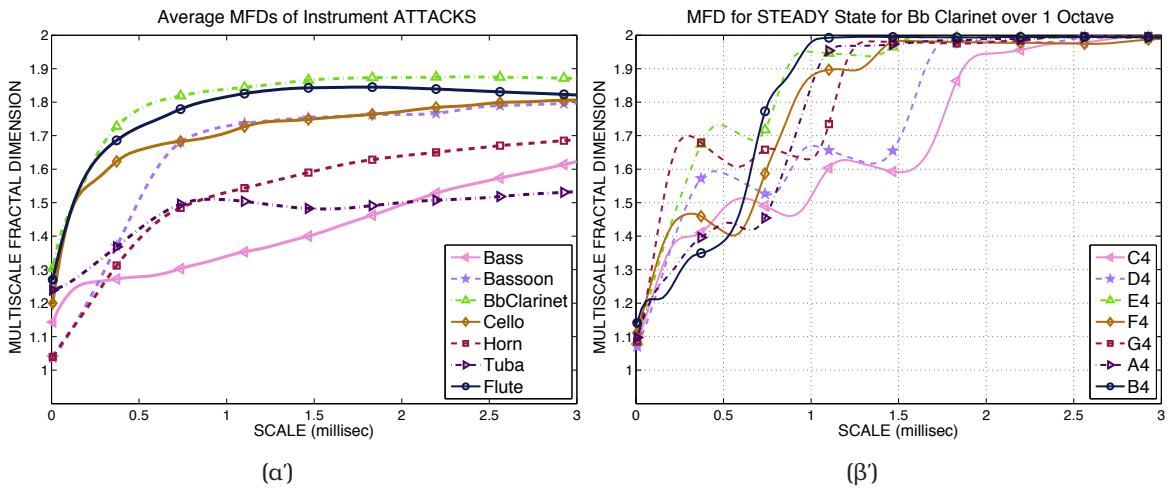
Συνεχίζουμε την ανάλυση των MFD εξετάζοντας το attack των μουσικών ήχων για να διερευνήσουμε πιθανές διαφορές του από τη σταθερή κατάσταση.

Η επεξεργασία που ακολουθούμε είναι παρόμοια με αυτή της σταθερής κατάστασης, λαμβάνοντας υπόψη μας τις ιδιαιτερότητες κάθε οργάνου σε σχέση με τη διάρκεια του attack. Παρατηρούμε πως το MFD προφίλ του attack παρουσιάζει παρόμοιες τάσεις με τη σταθερή κατάσταση των ήχων. Ωστόσο, ορισμένες από τις διαφορές είναι οι εξής: έχει υψηλότερη τιμή $D$ σε μικρότερες κλίμακες $s_t$ και παρουσιάζει μεγαλύτερη ανομοιομορφία και τραχύτητα σε σύγκριση με τη σταθερή κατάσταση. Οι διαφορές αυτές θα μπορούσαν ενδεχομένως να εξηγηθούν από τις συνιστώσες θορύβου που εμφανίζονται στην αρχή των ήχων, για παράδειγμα το «ξύσιμο» του δοξαριού στη χορδή του βιολιού, το τρίξιμο που κάνει το γλωσσίδι του κλαρινέτου κ.ά., όπως συζητήθηκε στην Ενότητα 2.1.1.

Στο Σχήμα 3.5 (α) βλέπουμε τη μέση τιμή του MFD για το attack και για όλο το εύρος συχνοτήτων των οργάνων που εξετάζουμε (δυναμική *forte*). Παρατηρούμε αυξημένη τιμή του $D(s = 1)$ για τη μικρότερη διακριτή κλίμακα καθώς και συγκεκριμένες διαφορές του MFD προφίλ μεταξύ των οργάνων. Εν κατακλείδι, η ανάλυση του attack παρουσίασε διαφορές τόσο μεταξύ attack και σταθερής κατάστασης του ίδιου τόνου καθώς και μεταξύ των διαφορετικών οργάνων. Θεωρούμε τη διαφοροποίηση του MFD προφίλ μεταξύ του attack και της σταθερής κατάστασης σημαντική, καθώς θα μπορούσε να σηματοδοτήσει τη μετάβαση από την αρχική στη σταθερή κατάσταση της νότας. Το Σχήμα 3.6 παρουσιάζει παραδείγματα των δύο καταστάσεων για τη νότα A3 για το τσέλο και για τη νότα F4 για το φλάουτο. Για το τσέλο παρατηρούμε υψηλότερη τιμή $D(s = 1)$ και πιο ανομοιόμορφο προφίλ στο attack, ενώ για το φλάουτο οι δύο καταστάσεις παρουσιάζουν περισσότερες ομοιότητες. Ωστόσο, βλέπουμε πως οι τιμές του MFD στο attack είναι αυξημένες για το συνολικό προφίλ.

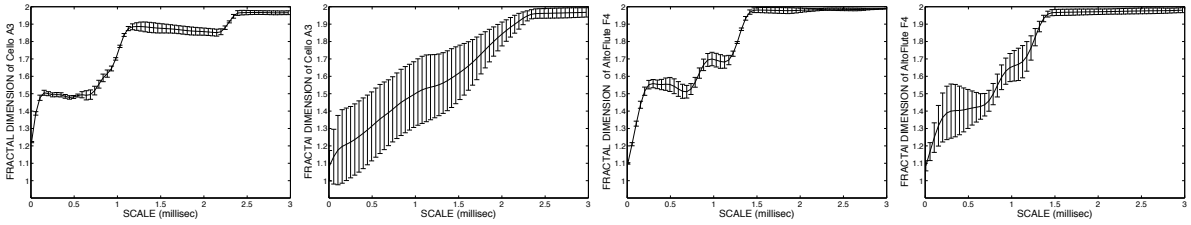
### 3.2.3 Η Μεταβλητότητα του MFD για το Ίδιο Όργανο

Το Σχήμα 3.5 (β) παρουσιάζει τα MFD προφίλ για τις νότες C4–B4 του B♭ κλαρινέτου για μια οκτάβα, με συχνότητες περίπου 260–493Hz, όπου και επιβεβαιώνονται οι προηγούμενες παρατηρήσεις μας, ότι δηλαδή η MFD διάσταση στις διαφορετικές κλίμακες εξαρτάται από την ακουστική συχνότητα του ήχου. Παρατηρούμε πως το προφίλ για ήχους υψηλότερης συχνότητας καταλαμβάνει την πρώτη κορυφή και έχει την υψηλότερη τιμή $D$ για μικρότερες κλίμακες $s_t$, αν και διατηρεί τη μορφή του σε σχέση με




(α′) (β′)

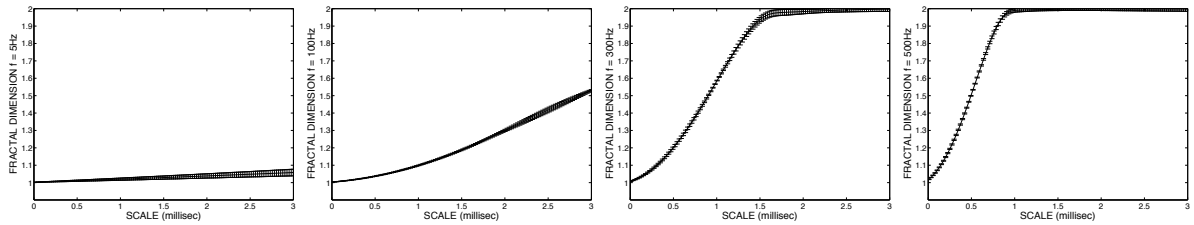
*Σχήμα 3.5:* (α′) Μέση τιμή του MFD προφίλ υπολογισμένη για το attack των εφτά μουσικών οργάνων, και για όλο το εύρος συχνοτήτων (με χρήση παραθύρων ανάλυσης 30 ms). (β′) MFD προφίλ για τη σταθερή κατάσταση τόνων μιας οκτάβας του B♭ κλαρινέτου, για ένα παράθυρο των 30 ms.




*Σχήμα 3.6:* Μέση τιμή του MFD προφίλ και τυπική απόκλιση του attack και της σταθερής κατάστασης, για τη νότα A3 για το τσέλο (πρώτη και δεύτερη εικόνα) και τη νότα F4 για το φλάουτο (τρίτη και τέταρτη εικόνα).

τον τρόπο που μεταβάλλεται (βλ. επίσης Εν. 3.2.1). Το φαινόμενο αυτό, με συγκεκριμένες διαφοροποιήσεις για τα διαφορετικά όργανα, παρατηρείται ως επί το πλείστον στα ξύλινα και στα χάλκινα πνευστά και αρχίζει να αναπτύσσεται σε συχνότητες περίπου 260 Hz και πάνω.

Στην επόμενη ενότητα παρουσιάζεται συμπληρωματική ανάλυση για την εξάρτηση αυτή του MFD από τη συχνότητα, ενώ διερευνούμε περαιτέρω και άλλα χαρακτηριστικά που έχουν ήδη συζητηθεί, χρησιμοποιώντας συνθετικά σήματα αποτελούμενα από ημίτονα. Ωστόσο, πρέπει να αναφέρουμε πως από την έως τώρα ανάλυση προκύπτουν ενδείξεις ότι τα MFD θα μπορούσαν πιθανώς να είναι χρήσιμα όχι μόνο στη διάκριση των διάφορων κατηγοριών οργάνων, αλλά ενδεχομένως και στην εκτίμηση της ακουστικής συχνότητας των σημάτων.



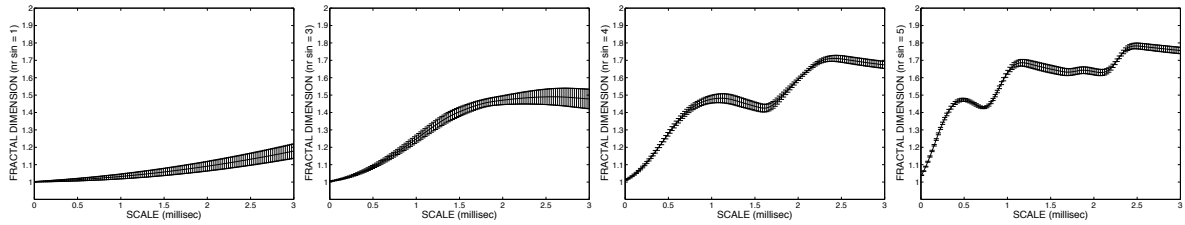

*Σχήμα 3.7:* Μέση τιμή του MFD και τυπική απόκλιση (error bars) απλών ημιτονοειδών σημάτων με συχνότητες 5, 100, 300 και 500 Hz (για παράθυρα ανάλυσης 30 ms και επικάλυψη 15 ms).

## 3.3 Ανάλυση των MFD σε Συνθετικά Σήματα

Εφαρμόζουμε τον MFD αλγόριθμο σε συνθετικά σήματα, δηλαδή απλά αλλά και πιο πολύπλοκα ημίτονα, προκειμένου να αξιολογήσουμε προηγούμενες παρατηρήσεις μας, όπως για παράδειγμα την απόκλιση του MFD στα διαφορετικά πλαίσια ανάλυσης της ίδιας νότας, τη μεταβλητότητα του MFD προφίλ σε διαφορετικές νότες του ίδιου οργάνου κ.ά. Στην πειραματική αυτή ανάλυση απομονώνουμε και μεταβάλλουμε επιμέρους παραμέτρους των ημιτόνων, ενώ κρατάμε όλες τις υπόλοιπες σταθερές, εξαλείφοντας έτσι χαρακτηριστικά των σημάτων που θεωρούμε πως μας εμποδίζουν στο να να βγάλουμε σωστά συμπεράσματα. Οι περιπτώσεις που εξετάζονται είναι οι εξής: (i) Απλά ημιτονοειδή σήματα διαφορετικών συχνοτήτων. (ii) Σύνθετα ημιτονοειδή σήματα καθώς προσθέτουμε ημίτονα μεγαλύτερης συχνότητας. (iii) Προσομοίωση μιας «νότας» συγκεκριμένης συχνότητας ενώ προσθέτουμε ημίτονα συχνοτήτων ίσων με τις αρμονικές της και τέλος, (iv) προσομοίωση της ίδιας «νότας», ενώ μεμονωμένες αρμονικές λείπουν, προκειμένου να μιμηθούμε όργανα όπως το κλαρινέτο, το οποίο αποτελείται μόνο από τις περιττές αρμονικές, π.χ., $f_0$, $3f_0$, $5f_0$ κ.λπ. Η παραμετροποίηση που χρησιμοποιείται για τον παρακάτω πειραματισμό είναι παρόμοια με αυτή της προηγηθείσας ανάλυσης.

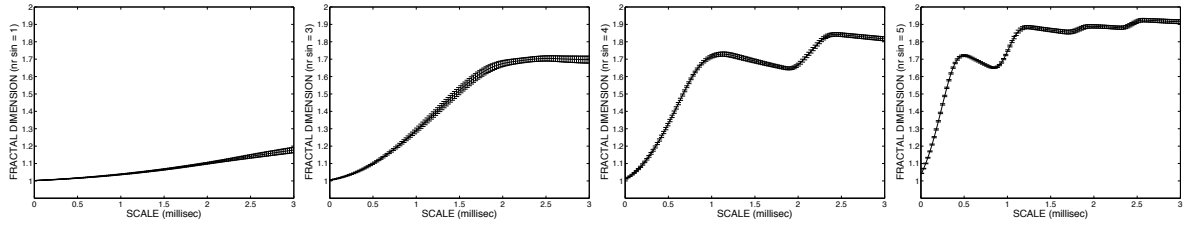
*Απλά Ημίτονα:* Στο Σχήμα 3.7 παρουσιάζεται η μέση τιμή του MFD προφίλ και η τυπική απόκλιση (error bars) για την απλούστερη περίπτωση, δηλαδή απλών ημιτονοειδών σημάτων διαφορετικών συχνοτήτων. Οι συχνότητες που χρησιμοποιούνται είναι: 5, 100, 300 και 500 Hz. Το πλάτος και η φάση διατηρούνται σταθερά και ισούνται με 1 και 3/4 του κύκλου, αντίστοιχα. Παρατηρούμε την εξάρτηση του MFD προφίλ από τη συχνότητα του σήματος. Συγκεκριμένα, η πρώτη κορυφή παρουσιάζεται στη μισή περίοδο (βλ. στο τελευταίο σχήμα, όπου η συχνότητα είναι 500 Hz, πως η πρώτη κορυφή του MFD προφίλ είναι περίπου στο 1 ms).

*Πολύπλοκα Σήματα με Ημίτονα Διπλάσιας Συχνότητας:* Στο Σχήμα 3.8 παρουσιάζεται η μέση τιμή του MFD προφίλ και η τυπική απόκλιση ημιτονοειδών σημάτων στα οποία προσθέτουμε διαδοχικά ημίτονα διπλάσιας συχνότητας σε σχέση με την αρχική, η οποία είναι 50 Hz. Οι συχνότητες των προστιθέμενων ημιτόνων είναι: 100, 200, 400 και 800 Hz. Το πλάτος και η φάση παραμένουν σταθερά. Εδώ παρατηρούμε ότι η δομή του




(α′) (β′) (γ′) (δ′)

*Σχήμα 3.8:* Μέση τιμή του MFD και τυπική απόκλιση συνθετικών ημιτονοειδών σημάτων. (α′) Αρχικό ημίτονο $x_0 = x_{50}$ (50 Hz), (β′) $x = x_{50} + x_{100} + x_{200}$, (γ′) $x = x_{50} + x_{100} + x_{200} + x_{400}$ και (δ′) $x = x_{50} + x_{100} + x_{200} + x_{400} + x_{800}$.



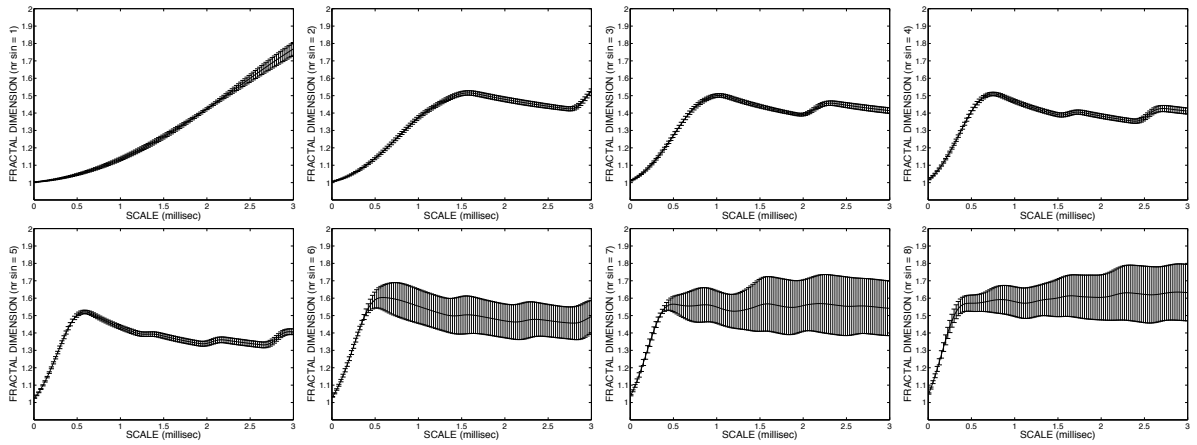

(α′) (β′) (γ′) (δ′)

*Σχήμα 3.9:* Μέση τιμή του MFD και τυπική απόκλιση συνθετικών ημιτονοειδών σημάτων ενώ προστίθενται ημίτονα διπλάσιας συχνότητας με πλάτος το οποίο μειώνεται γεωμετρικά. (α′) Αρχικό ημίτονο $x_0 = x_{50,1}$ (όπου 50 σε Hz και 1 το πλάτος), (β′) $x = x_{50,1} + x_{100,1/2} + x_{200,1/4}$, (γ′) $x = x_{50,1} + x_{100,1/2} + x_{200,1/4} + x_{400,1/8}$, και (δ′) $x = x_{50,1} + x_{100,1/2} + x_{200,1/4} + x_{400,1/8} + x_{800,1/16}$. Η φάση μεταβάλλεται με τυχαίο τρόπο.

MFD προφίλ εμφανίζει περισσότερες μεταβολές, καθώς ο αριθμός των ημίτονων αυξάνεται. Ωστόσο, η μορφή του παραμένει αρκετά απλή.

*Πολύπλοκα Σήματα Αποτελούμενα από Ημίτονα Διαφορετικής Συχνότητας, Διαφορετικού Πλάτος και Τυχαία Επιλεγμένης Φάσης:* Στο αρχικό σήμα $x_0$ συχνότητας 50 Hz και πλάτους 1 προσθέτουμε ημίτονα διαφορετικών συχνοτήτων, διαφορετικού πλάτους και τυχαία επιλεγμένης φάσης [0, $2\pi$], βλ. Σχήμα 3.9. Οι συχνότητες (σε Hz) και τα πλάτη των προστιθέμενων ημιτόνων είναι: $x_1$ = 100, 1/2, $x_2$ = 200, 1/4, $x_3$ = 400, 1/8, και $x_4$ = 800, 1/16. Εδώ παρατηρούμε ότι η μείωση του πλάτους δεν επηρεάζει το προφίλ, ενώ η μεταβολή της φάσης αυξάνει κάπως τις τιμές του MFD για κλίμακες μεγαλύτερες της $D(s = 1)$. Παρ′ όλα αυτά, το σχήμα και η δομή του προφίλ αναπτύσσονται με παρόμοιο τρόπο σε σχέση με την προηγούμενη περίπτωση. Οι παρατηρήσεις μας αυτές είναι σύμφωνες με το γεγονός ότι η φάση δεν συμβάλλει στην αντίληψη του ηχοχρώματος, αλλά παράγει μόνο μικρές αλλαγές στον ήχο που αντιλαμβάνεται ο ακροατής [46, 134].

*«Προσομοίωση της Νότας C3»:* Σε αυτή την περίπτωση προσπαθήσαμε να προσομοιώσουμε τη νότα C3 με θεμελιώδη συχνότητα $f_0$ = 131 Hz, προσθέτοντας

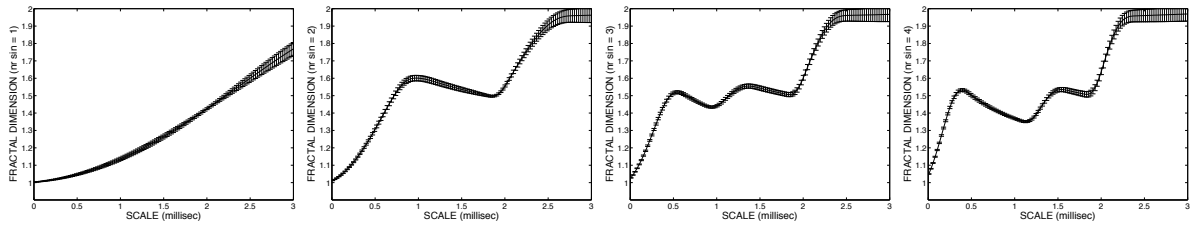



*Σχήμα 3.10:* Μέση τιμή του MFD και τυπική απόκλιση συνθετικών ημιτονοειδών σημάτων καθώς προστίθενται ημίτονα με συχνότητα ίση με τις αρμονικές της νότας C3 ($f_0 = 131$ Hz). Το πλάτος και η φάση παραμένουν σταθερά.

διαδοχικά ημίτονα συχνότητας ίσης με τις αρμονικές της (το πλάτος και η φάση παραμένουν σταθερά). Άρα οι συχνότητες των ημιτόνων που προστίθενται είναι ακέραια πολλαπλάσια της $f_0$, δηλαδή, $f$= 262, 393, 524, 655, 789, 917, και 1046 Hz. Το αποτέλεσμα αυτής της προσομοίωσης φαίνεται στο Σχήμα 3.10 παρουσιάζοντας τη μέση τιμή του MFD και την τυπική απόκλιση, όπου και παρατηρούμε μεγάλη διακύμανση της τιμής $D$ στα διαφορετικά πλαίσια ανάλυσης για σήματα που αποτελούνται από έξι ή περισσότερα ημίτονα.

*«Προσομοίωση της Νότας C3 Προσθέτοντας Ημίτονα με Συχνότητα Ίση με τις Περιττές Αρμονικές της Νότας»:* Το Σχήμα 3.11 δείχνει τη μέση τιμή του MFD και την τυπική απόκλιση αυτής της προσομοίωσης. Σε αυτή την περίπτωση προσπαθούμε να μιμηθούμε όργανα όπως το κλαρινέτο, σε μια απόπειρα να εξακριβώσουμε αν χαρακτηριστικά του συχνοτικού περιεχομένου του οργάνου είναι δυνατόν να φανούν στη μορφή του MFD προφίλ. Οι συχνότητες των ημιτόνων («*αρμονικών*») που προστίθενται είναι ίσες με $f$ = 131, 393, 655 και 917 Hz (το πλάτος και η φάση παραμένουν σταθερά). Διαπιστώνουμε πως τα MFD προφίλ διαφοροποιούνται όταν μεμονωμένες αρμονικές λείπουν, ενώ παρατηρούμε υψηλότερη φράκταλ διάσταση και πιο πολύπλοκες δομές. Τέλος, όταν το πλάτος των αρμονικών μειώθηκε στο 1/2 οι αλλαγές που παρατηρήθηκαν ήταν αμελητέες.

Μετά την ανάλυση του MFD σε συνθετικά σήματα καταλήγουμε πως όντως υπάρχει εξάρτηση από τη συχνότητα του σήματος. Για υψηλότερες συχνότητες διαπιστώνουμε πως η μεγαλύτερη κορυφή παρουσιάζεται σε μικρότερες κλίμακες $s_t$ (περίπου στο μισό της περιόδου για απλά ημίτονα), κάτι που εκδηλώνεται τόσο σε απλά όσο και σε πιο σύνθετα σήματα. Επιπλέον, ο αριθμός των ημιτόνων που προστίθενται στο αρχικό σήμα επηρεάζει την πολυπλοκότητα του MFD προφίλ, ενώ διακρίνουμε επίσης αυξημένη διακύμανση του $D$ στα διαφορετικά πλαίσια ανάλυσης. Τέλος, η βραχέος χρόνου φράκταλ διάσταση στη




*Σχήμα 3.11:* Μέση τιμή του MFD και τυπική απόκλιση συνθετικών ημιτονοειδών σημάτων καθώς προστίθενται ημίτονα με συχνότητα ίση με τις περιττές αρμονικές της νότας C3 ($f_0 = 131$ Hz). Το πλάτος και η φάση παραμένουν σταθερά.

μικρότερη διακριτή κλίμακα $D(s = 1)$ παίρνει υψηλότερες τιμές όταν προσθέτουμε στο αρχικό σήμα ένα τυχαίο σήμα θορύβου.

Εν κατακλείδι, πρέπει να επισημάνουμε πως τέτοιου είδους «συνθετικοί τόνοι», οι οποίοι αποτελούνται μόνο από τη λεγόμενη «σταθερή κατάσταση της νότας», δεν μπορούν στην πραγματικότητα να συγκριθούν με πραγματικούς ήχους μουσικών οργάνων, δεδομένου ότι αυτού του είδους η σύνθεση δεν καταφέρνει να παράγει τις δυναμικές εναλλαγές των μουσικών ήχων. Μολαταύτα, καταλήγουμε στο συμπέρασμα πως τα συγκεκριμένα πειράματα, μας βοήθησαν στην καλύτερη κατανόηση μερικών εκ των χαρακτηριστικών των οργάνων. Για παράδειγμα, το γεγονός ότι το attack ορισμένων τόνων παρουσιάζει υψηλότερη φράκταλ διάσταση στις μικρότερες κλίμακες $D(s = 1)$ θα μπορούσε ενδεχομένως να σημαίνει την ύπαρξη θορύβου η άλλων παραγόντων, προερχόμενων π.χ. από το γλωσσίδι των πνευστών ή τον ήχο του δοξαριού πάνω στις χορδές των έγχορδων. Επιπλέον, η αυξημένη απόκλιση του $D$ στις χαμηλότερες οκτάβες, όπως για παράδειγμα είδαμε στην τούμπα, θα μπορούσε να σημαίνει πιο πλούσιο αρμονικό περιεχόμενο, το οποίο επιβεβαιώνεται σύμφωνα με τον Olson [116]. Το γεγονός ότι τα MFD προφίλ διαφέρουν όταν το συχνοτικό περιεχόμενο του ήχου αλλάζει (π.χ., στις υψηλότερες συχνότητες) θα μπορούσε να μας δώσει μια ένδειξη για τη σχετική θέση μιας νότας στη μουσική κλίμακα και να επιτρέψει μια προσέγγιση της κατανομής του συχνοτικού της περιεχομένου.

## 3.4 Πειράματα Αναγνώρισης Μουσικών Οργάνων

**Βάση Δεδομένων**

Σε αυτό το σημείο ενσωματώνουμε τη φράκταλ διάσταση σε πολλαπλές κλίμακες ως χαρακτηριστικό σε πειράματα αναγνώρισης μουσικών οργάνων, προκειμένου να αξιολογήσουμε τα αποτελέσματα της ανάλυσης που προηγήθηκε. Τα πειράματα αυτά διεξήχθησαν με τη χρήση 1331 μεμονωμένων μουσικών τόνων εφτά οργάνων: του κοντραμπάσου, του φαγκότου, του τσέλου, του B♭ κλαρινέτου, του φλάουτου, του κόρνου

και της τούμπας, από τη βάση δεδομένων του Πανεπιστημίου της IOWA [181]. Τα δεδομένα αποτελούνται από ολόκληρο το εύρος συχνοτήτων κάθε οργάνου και καλύπτουν τις δυναμικές ενδείξεις από piano ως forte. Τα σήματα αναλύθηκαν με τη χρήση παραθύρων 30 ms με επικάλυψη 15 ms.

### 3.4.1 Πειραματική Αξιολόγηση: Σύνολα Χαρακτηριστικών

Για να επιτύχουμε καλά αποτελέσματα αναγνώρισης, είναι απαραίτητη η ενσωμάτωση και η παραγωγή περιγραφικών αλλά σύντομων αναπαραστάσεων του σήματος, με ιδιαίτερη έμφαση στην ευρωστία (*robustness*), την αμεταβλητότητα (*invariance*) και την επιλογή/μείωση της διάστασης των χαρακτηριστικών (*dimensionality reduction*). Για να επιτευχθεί αυτό επιλέξαμε να μειώσουμε τον αριθμό του αρχικού συνόλου χαρακτηριστικών (124 MFD) χρησιμοποιώντας PCA ανάλυση. Με αυτό τον τρόπο τα χαρακτηριστικά είναι ασυσχέτιστα και παρουσιάζουν τη μέγιστη διακύμανση μεταξύ τους. Επιπλέον σύνολα χαρακτηριστικών (με λογαριθμική δειγματοληψία ή βάσει παρατήρησης) προέκυψαν ύστερα από εκτενή πειραματισμό.

Τα τελικά σύνολα αλλά και οι συνδυασμοί τους με καθιερωμένα σύνολα χαρακτηριστικών, π.χ. MFCC, αξιολογήθηκαν με τη χρήση στατικών Γκαουσιανών μοντέλων (GMM), και κρυφών Μαρκοβιανών μοντέλων (HMM) τα οποία δύνανται να μοντελοποιήσουν τη δυναμικότητα στον χρόνο, χρησιμοποιώντας ποικίλους συνδυασμούς καταστάσεων $N$ και/ή μειγμάτων $M$ για λόγους πειραματισμού. Για την εφαρμογή των Μαρκοβιανών μοντέλων χρησιμοποιήθηκε το εργαλείο HTK [193] χρησιμοποιώντας τον αλγόριθμο EM και Viterbi. Η απόδοση των επιλεγμένων χαρακτηριστικών συγκρίθηκε με τα MFCC (13 μαζί με την ενέργεια), τα οποία επιλέχθηκαν τόσο για την καλή τους απόδοση όσο και λόγω της αποδοχής τους σε παρόμοιες μελέτες. Η ανάλυση των MFCC διεξήχθη με χρήση παραθύρων των 30 ms με 15 ms επικάλυψη, και με 24 τριγωνικά ζωνοπερατά φίλτρα. Σε όλες τις περιπτώσεις τα δεδομένα εκπαίδευσης έχουν επιλεγεί τυχαία να είναι το 70% των διαθέσιμων ηχητικών σημάτων, και τα αποτελέσματα που παρουσιάζονται έχουν προκύψει ύστερα από πενταπλή συγκριτική τεκμηρίωση (*cross-validation*).

Έχουν διενεργηθεί δύο διαφορετικά σετ πειραμάτων και τα σύνολα χαρακτηριστικών που αξιολογήθηκαν φαίνονται στους Πίνακες 3.2 και 3.3. Πριν από την τελική επιλογή των συνόλων αυτών υπήρξε εκτενής πειραματισμός σε σχέση με την απόδοσή τους.

Τα MFD χαρακτηριστικά που χρησιμοποιήθηκαν στα πειράματα δημιουργήθηκαν με δειγματοληψία του αρχικού συνόλου με δύο διαφορετικούς τρόπους: α) με λογαριθμική δειγματοληψία (με αποτέλεσμα δεκατρία χαρακτηριστικά - MFDLG–σύνολο), παράδειγμα της οποίας φαίνεται στο Σχήμα 3.12, και β) το σύνολο MFDLGOB που αποτελείται από είκοσι τέσσερα χαρακτηριστικά, δηλαδή τα MFDLG συν έντεκα σημεία επιλεγμένα ύστερα από παρατήρηση του MFD προφιλ για σήματα διαφορετικών

*Πίνακας 3.2:* Λίστα MFD χαρακτηριστικών του πρώτου σετ πειραμάτων για αναγνώριση μουσικών οργάνων.

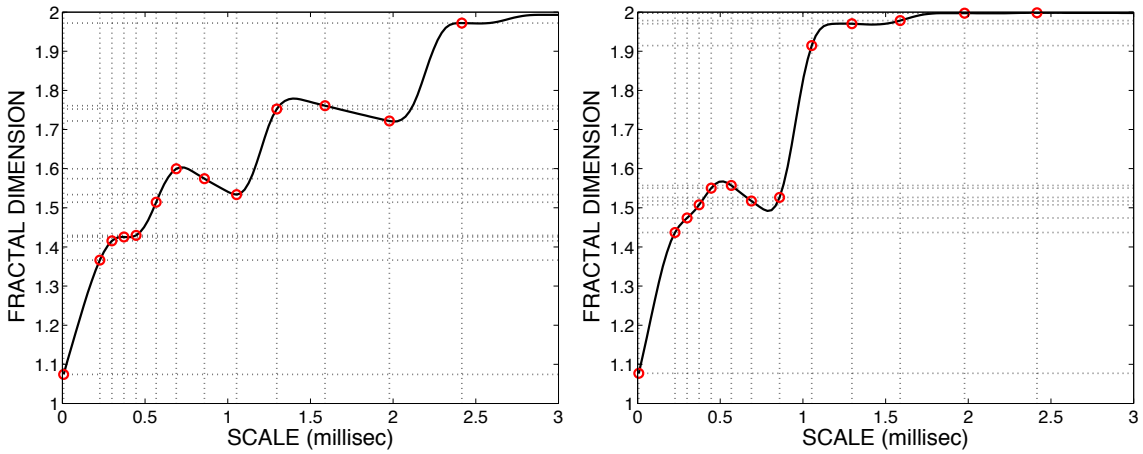
| **Λίστα Χαρακτηριστικών 1ου σετ Πειραμάτων** | |
|---|---|
| 1 | 6 MFD μετά από ανάλυση PCA (MFDPC) |
| 2 | 13 λογαριθμικά επιλεγμένα MFD χαρακτηριστικά (MFDLG) |
| 3 | 13 MFCC |
| **Multi-stream cases** | |
| 4 | 6 MFDPC + 13 MFCC |
| 5 | 13 MFDLG + 13 MFCC |

*Πίνακας 3.3:* Λίστα MFD χαρακτηριστικών του δεύτερου σετ πειραμάτων με την προσθήκη των χρονικών παραγώγων $\Delta$ για αναγνώριση μουσικών οργάνων. (Το σύνολο $\text{MFDPC}_{\Delta\text{i}}$ υποδηλώνει τα MFD χαρακτηριστικά υστέρα από την ανάλυση PCA στα ξεχωριστά σύνολα χαρακτηριστικών, ενώ το σύνολο $\text{MFDPC}_{\Delta\text{f}}$ υποδηλώνει τα MFD χαρακτηριστικά υστέρα από την ανάλυση PCA στο συνολικό ενωμένο σύνολο χαρακτηριστικών.)

| **Λίστα Χαρακτηριστικών 2ου σετ Πειραμάτων** | |
|---|---|
| 1 | 13 MFDPC + 13 $\Delta$ + 13 $\Delta\Delta$ ($\text{MFDPC}_{\Delta}$) |
| 2 | 13 MFDLG + 13 $\Delta$ + 13 $\Delta\Delta$ ($\text{MFDLG}_{\Delta}$) |
| 3 | 13 MFCC + 13 $\Delta$ + 13 $\Delta\Delta$ ($\text{MFCC}_{\Delta}$) |
| **Multi-stream cases** | |
| 4 | 39 $\text{MFDPC}_{\Delta\text{i}}$ + 39 $\text{MFCC}_{\Delta}$ |
| 5 | 39 $\text{MFDPC}_{\Delta\text{f}}$ + 39 $\text{MFCC}_{\Delta}$ |
| 6 | 39 $\text{MFDLG}_{\Delta}$ + 39 $\text{MFCC}_{\Delta}$ |

οργάνων. Το διάνυσμα MFDLG αποτελείται από τη διάσταση $D$ σε κλίμακες $s = 1, 10, 13, 16, 19, 24, 29, 36, 44, 54, 66, 82$ και $100$, ενώ το διάνυσμα MFDLGOB ενισχύθηκε με τη διάσταση $D$ στις κλίμακες $s = 3, 5, 6, 39, 48, 57, 64, 74, 91, 115$ και $122$. Και τα δύο σύνολα περιλαμβάνουν τη φράκταλ διάσταση στη μικρότερη κλίμακα $D(s = 1)$. Πειραματισμός διεξήχθη όσον αφορά και την ανάλυση PCA. Οι δύο περιπτώσεις που εξετάστηκαν είναι: i) PCA στο ενιαίο διάνυσμα των χαρακτηριστικών MFD με τις παραγώγους του και ii) στα τρία επιμέρους σύνολα χαρακτηριστικών: το διάνυσμα MFD, την πρώτη και τη δεύτερη παράγωγό τους ξεχωριστά. Ύστερα από αρκετές αξιολογήσεις των χαρακτηριστικών και δεδομένου ότι το $\text{MFDLGOB}_{\Delta}$ είχε συγκρίσιμα αποτελέσματα με το $\text{MFDLG}_{\Delta}$, αναφέρουμε μόνο τα αποτελέσματα για το $\text{MFDLG}_{\Delta}$. Όσον αφορά την PCA ανάλυση, παρατηρούμε ότι η εφαρμογή της στα επιμέρους διανύσματα, καταλήγοντας σε ένα 13-διάστατο διάνυσμα από κάθε σύνολο (συνολικά 39 χαρακτηριστικά), οδηγεί σε βελτίωση των αποτελεσμάτων αναγνώρισης. Ωστόσο, καλά αποτελέσματα επιτύχαμε και από την εφαρμογή της PCA ανάλυσης στα ενιαία διανύσματα χαρακτηριστικών, με τελικό αριθμό συνιστωσών 30, 32 ή 39 (μερικά από τα οποία αναφέρονται παρακάτω).

Η αξιολόγηση περιλαμβάνει τη μεταβολή του αριθμού των καταστάσεων $N$ [3-9] και των μειγμάτων $M$ [1-5] με τη χρήση GMM έως 5 μείγματα και HMM έως 9 καταστάσεις,




*Σχήμα 3.12:* Παράδειγμα των δεκατριών λογαριθμικά επιλεγμένων σημείων του MFD για τη νότα A3 του B♭ κλαρινέτου και τη νότα A4 για το φαγκότο, τα οποία δημιουργούν το σύνολο χαρακτηριστικών MFDLG.

υιοθετώντας left-right τοπολογία για τη μοντελοποίηση. Επιπρόσθετα, διεξαγάγαμε multi-stream πειράματα για τη μοντελοποίηση των δύο διαφορετικών συνόλων χαρακτηριστικών (δηλαδή, MFD vs. MFCC) χρησιμοποιώντας διαφορετικούς εκθέτες βαρών οι οποίοι υποδηλώνουν την αξιοπιστία κάθε ροής (*stream*). Τα βάρη είναι δυνατόν να καθοριστούν από μας σε κάποιες τιμές που αντανακλούν τη σχετική εμπιστοσύνη στη συγκεκριμένη ροή ή μπορούν να εκτιμηθούν και να βελτιστοποιηθούν, π.χ., [54, 139]. Η βελτιστοποίηση των βαρών πραγματοποιήθηκε σε ένα hold-out σετ, το οποίο επιλέχθηκε από το αρχικό σύνολο εκπαίδευσης (το 70% του αρχικού συνόλου εκπαίδευσης χωρίστηκε σε 60% για την εκπαίδευση ενώ το 10% αποτέλεσε το hold-out σύνολο). Υποθέτουμε ότι οι δύο εκθέτες $w_1, w_2$ ικανοποιούν τον περιορισμό $0 \leq w_1, w_2 \leq 1$ και $w_1 + w_2 = 1$, ενώ τα βάρη που μεγιστοποιούν τα αποτελέσματα αναγνώρισης (*accuracy*) για το hold-out σύνολο επιλέχθηκαν και εφαρμόστηκαν στο τελικό σύνολο δοκιμής. Στο Σχ. 3.13 φαίνονται τα αποτελέσματα αναγνώρισης για το hold-out σύνολο (με 5 cross-validation), ενώ τα συνολικά ποσοστά αναγνώρισης για το σύνολο δοκιμής, τα οποία πρόκειται να συζητηθούν στη συνέχεια, παρουσιάζονται στους Πίνακες 3.6 – 3.7.

### 3.4.2 Πειραματική Αξιολόγηση: Αποτελέσματα

#### Πρώτο σετ πειραμάτων

Τα αποτελέσματα του συνδυασμού των προτεινόμενων χαρακτηριστικών με τα MFCC αποδείχθηκαν ελαφρώς καλύτερα από τα αποτελέσματα των MFCC για τις περισσότερες περιπτώσεις (ακόμα και γι' αυτές που δεν παρουσιάζονται εδώ), αν και τα MFD μόνα τους παρουσιάζουν χαμηλότερο διαχωρισμό. Το μειονέκτημα των MFD είναι ο χαμηλός διαχωρισμός μεταξύ του B♭ κλαρινέτου και του φλάουτου, τα οποία παρουσιάζουν τα

*Πίνακας 3.4:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) για τα MFD με HMM και GMM, όπου $N$ ο αριθμός των καταστάσεων και $M$ ο αριθμός των μειγμάτων. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 3.2 (πρώτο σετ πειραμάτων).

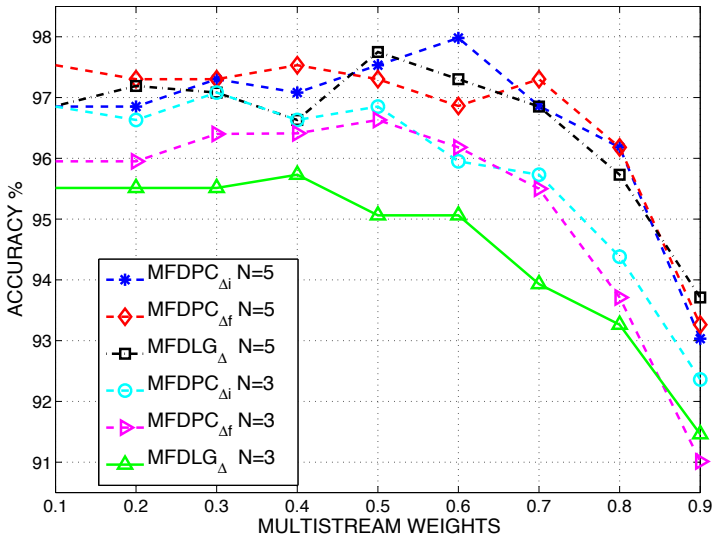
| Ποσοστά επιτυχίας κατηγοριοποίησης | | | | |
|---|---|---|---|---|
| | Βάρη | GMM | HMM | |
| Χαρακτηριστικά | MFD-MFCC | $M=3$ | $N=3$ $M=3$ | $N=5$ $M=3$ |
| MFDPC-MFCC | 0.2 - 0.8 | **86.01** | **93.01** | **94.68** |
| | 0.5 - 1.0 | 85.78 | 92.31 | 94.22 |
| | 0.5 - 0.5 | **86.01** | 92.78 | **94.68** |
| MFDLG-MFCC | 0.2 - 0.8 | 85.25 | 92.93 | **94.98** |
| | 0.5 - 1.0 | **85.78** | **93.16** | 94.91 |
| | 0.5 - 0.5 | 85.40 | 92.47 | 94.45 |
| MFCC | - | 87.07 | 92.93 | 94.40 |
| MFDPC | - | 69.35 | 75.59 | 76.51 |
| MFDLG | - | 64.95 | 71.79 | 72.11 |

*Πίνακας 3.5:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) για τα MFD ανά όργανο για τους δύο καλύτερους συνδυασμούς χαρακτηριστικών σε σύγκριση με τα MFCC (πρώτο σετ πειραμάτων).

| Ποσοστά επιτυχίας κατηγοριοποίησης | | | |
|---|---|---|---|
| Κατηγορία Μουσικού οργάνου | MFDPC + MFCC | MFDLG + MFCC | MFCC |
| Double Bass | **100** | **100** | **100** |
| Bassoon | 93.32 | **95.84** | 88.52 |
| B♭ Clarinet | **78.54** | 77.07 | 72.25 |
| Cello | 93.64 | 93.94 | **96.73** |
| Horn | 97.9 | **100** | 92.08 |
| Tuba | **100** | **100** | **100** |
| Flute | 96.02 | 95.58 | **97.25** |

χαμηλότερα αποτελέσματα αναγνώρισης μεταξύ των οργάνων (περίπου 55%). Η έως τώρα ανάλυση επισήμανε τις ομοιότητές τους για τους ήχους υψηλότερων συχνοτήτων (βλ. Ενότητα 3.2.1), και αυτός είναι πιθανώς και ο λόγος των χαμηλών ποσοστών. Άρα, υπολογίζουμε το median μέσο, αντί για τον κλασικό μέσο (mean), και τα αποτελέσματα για την καλύτερη περίπτωση του συνόλου MFDPC ($N=5$, $M=3$) διαμορφώνονται στο 79,7% και για την καλύτερη περίπτωση του συνόλου MFDLG ($N=5$, $M=3$) στο 75,8%. Παρ' όλα αυτά, οφείλουμε να επισημάνουμε ότι η τούμπα, το φαγκότο και το κοντραμπάσο είναι τα όργανα με την καλύτερη αναγνώριση, κάτι που συμφωνεί με την ανάλυσή μας.

Ο Πίνακας 3.4 δείχνει τα ποσοστά επιτυχίας αναγνώρισης, όπου το μέτρο αξιολόγησης είναι το ποσοστό μέσης ακρίβειας αναγνώρισης (mean accuracy %), για τα διαφορετικά σύνολα χαρακτηριστικών με HMM μοντελοποίηση ($N=5$ και $M=3$). Για τα ποσοστά αναγνώρισης των προτεινόμενων χαρακτηριστικών σε σύγκριση με τα MFCC για τις διαφορετικές κατηγορίες οργάνων, βλ. Πίνακα 3.5. Παρατηρούμε πως η προσθήκη των MFD ενισχύει την αναγνώριση του φαγκότου, του κλαρινέτου και του κόρνου, μειώνει την




*Σχήμα 3.13:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) κατά τη βελτιστοποίηση των βαρών για τις multi-stream δοκιμές των HMM για $N = 3, 5$ και $M = 5$. Ο $x$-άξονας δείχνει το βάρος $w_1$ για τα MFD (όπου $w_1 + w_2 = 1$).

αναγνώριση των MFCC για το τσέλο και το φλάουτο, ενώ το κοντραμπάσο και η τούμπα διατηρούν την ήδη καλή απόδοση των MFCC.

### Δεύτερο σετ πειραμάτων

Το Σχήμα 3.13 δείχνει τα ποσοστά επιτυχίας κατηγοριοποίησης (%) που επιτυγχάνονται στο hold-out σύνολο για τα τρία διαφορετικά MFD σύνολα χαρακτηριστικών ύστερα από τη σύμμειξή τους με τα MFCC. Παρατηρούμε πως βάρος μεταξύ $0.5 - 0.8$ για τα MFCC οδηγεί στις περισσότερες περιπτώσεις σε καλύτερα αποτελέσματα για τρεις ή πέντε καταστάσεις και πέντε μείγματα. Ως εκ τούτου, επιλέγουμε τρεις περιπτώσεις, όπου $w_2 = 0.8, 0.6, 0.5$ για τα MFCC, και στον Πίνακα 3.6 παρουσιάζουμε τα αποτελέσματα της αναγνώρισης στο τελικό σύνολο δοκιμής για διάφορα σύνολα χαρακτηριστικών με την προσθήκη των παραγώγων $\Delta$. Λόγω της απόλυτης αύξησης του ποσοστού αναγνώρισης κατά 10% για τη μοντελοποίηση με GMM όπου $M = 5$, σε σύγκριση με $M = 3$, συνεχίζουμε τη συζήτηση για $M = 5$ μείγματα. Διακρίνουμε μείωση του σφάλματος έως και 26% για τα MFDPC$_{\Delta \mathrm{i}}$ ($N = 3$, $M = 5$) και μέχρι 32% και 10% για τα MFDPC$_{\Delta \mathrm{f}}$ και MFDLG$_{\Delta}$ ($N = 5$, $M = 5$), αντίστοιχα. Σε σύγκριση με τα προηγούμενα πειράματα, χωρίς την προσθήκη παραγώγων (βλ. Πίνακα 3.4), διαπιστώνουμε μείωση του σφάλματος για τα MFDLG μέχρι 50%, ενώ για τα MFDPC έως 35%. Ο Πίνακας 3.7 δείχνει το ποσοστό επιτυχίας κατηγοριοποίησης για τα διαφορετικά όργανα, για τις περιπτώσεις των συνδυασμένων συνόλων χαρακτηριστικών συγκριτικά με τα MFCC$_{\Delta}$ (με HMM). Υπογραμμίζουμε τη βελτίωση της αναγνώρισης

*Πίνακας 3.6:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) για τα MFD με HMM και GMM, όπου $N$ ο αριθμός των καταστάσεων και $M$ ο αριθμός των μειγμάτων. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 3.3 (δεύτερο σετ πειραμάτων).

| Ποσοστά επιτυχίας κατηγοριοποίησης | | | | | |
|---|---|---|---|---|---|
| | Βάρη | GMM | | HMM | |
| Χαρακτηριστικά | MFD-MFCC | $M=3$ | $M=5$ | $N=3$ $M=5$ | $N=5$ $M=5$ |
| $MFDPC_{\Delta i}$ – $MFCC_{\Delta}$ | 0.2 - 0.8 | 84.36 | **94.70** | 97.52 | **97.93** |
| | 0.4 - 0.6 | 84.80 | 94.65 | **98.03** | 97.67 |
| | 0.5 - 0.5 | **85.75** | 94.59 | 97.67 | 97.83 |
| $MFDPC_{\Delta f}$– $MFCC_{\Delta}$ | 0.2 - 0.8 | **86.75** | 94.29 | **97.63** | **98.18** |
| | 0.4 - 0.6 | 85.20 | **94.54** | 97.42 | 97.93 |
| | 0.5 - 0.5 | 84.50 | 93.88 | 97.17 | 97.93 |
| $MFDLG_{\Delta}$– $MFCC_{\Delta}$ | 0.2 - 0.8 | 83.74 | **93.79** | **96.82** | **97.58** |
| | 0.4 - 0.6 | 83.84 | 92.73 | 96.77 | 97.42 |
| | 0.5 - 0.5 | **83.94** | 91.82 | 96.72 | 97.43 |
| $MFCC_{\Delta}$ | - | 83.64 | 93.23 | 96.41 | 97.32 |
| $MFDPC_{\Delta i}$ | - | 70.10 | 77.88 | 85.91 | 88.08 |
| $MFDPC_{\Delta f}$ | - | 67.42 | 75.35 | 85.66 | 88.53 |
| $MFDLG_{\Delta}$ | - | 68.08 | 75.25 | 86.41 | 87.43 |

*Πίνακας 3.7:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) για τα MFD ανά όργανο για τους τρεις καλύτερους συνδυασμούς χαρακτηριστικών, $MFDPC_{\Delta i}$, $MFDPC_{\Delta f}$, $MFDLG_{\Delta}$, σε σύγκριση με τα $MFCC_{\Delta}$ (για $N=5$, $M=5$, εκτός από το $MFDPC_{\Delta i}$ σύνολο το οποίο το δείχνουμε για $N=3$, $M=5$).

| Ποσοστά επιτυχίας κατηγοριοποίησης | | | | |
|---|---|---|---|---|
| Κατηγορία Μουσικού Οργάνου | $MFDPC_{\Delta i}$ $MFCC_{\Delta}$ 0.4-0.6 | $MFDPC_{\Delta f}$ $MFCC_{\Delta}$ 0.2-0.8 | $MFDLG_{\Delta}$ $MFCC_{\Delta}$ 0.2-0.8 | $MFCC_{\Delta}$ |
| Double Bass | 99.76 | 99.52 | 95.52 | **100** |
| Bassoon | 96.66 | **97.20** | 96.08 | 94.98 |
| B♭ Clarinet | **94.64** | 88.78 | 88.80 | 91.68 |
| Cello | 97.88 | 94.24 | 93.10 | **98.26** |
| Horn | **99.28** | 87.84 | 86.42 | 94.30 |
| Tuba | **100** | 99.40 | 98.18 | **100** |
| Flute | **97.34** | 89.70 | 90.26 | 97.06 |

όλων των οργάνων και συγκεκριμένα του διαχωρισμού των B♭ κλαρινέτου και φλάουτου. Σημειώνουμε, δε, ότι το κοντραμπάσο και η τούμπα είναι και πάλι ανάμεσα στα όργανα με το μεγαλύτερο ποσοστό αναγνώρισης, ενώ η μόνη περίπτωση μείωσης των αποτελεσμάτων είναι το τσέλο.

Τέλος, όσον αφορά τα $MFDPC_{\Delta}$ και $MFDLG_{\Delta}$, παρατηρούμε πως τα λογαριθμικά επιλεγμένα χαρακτηριστικά έχουν εξίσου καλή αν όχι καλύτερη αναγνώριση, πράγμα που σημαίνει μείωση της υπολογιστικής πολυπλοκότητας για την εξαγωγή χαρακτηριστικών.

**Συμπεράσματα**

Στο κεφάλαιο αυτό, παρουσιάσαμε τα αποτελέσματα της ανάλυσης και της πειραματικής αξιολόγησης σε ήχους μουσικών οργάνων με τη χρήση της φράκταλ διάστασης σε πολλαπλές κλίμακες. Βάσει της αρχικής μας ανάλυσης παρατηρούμε πως τα MFD μπορούν να διακρίνουν διάφορα χαρακτηριστικά των μουσικών οργάνων ενώ η πειραματική αξιολόγηση των χαρακτηριστικών έδειξε μείωση του σφάλματος αναγνώρισης η οποία φτάνει έως και 32%.

# Κεφάλαιο 4

# Μη-Γραμμικά Μοντέλα Βασισμένα στις Διαμορφώσεις Πλάτους και Συχνότητας (AM-FM)

Στο κεφάλαιο αυτό εξετάζουμε το μη-γραμμικό μοντέλο διαμορφώσεων (*AM-FM modulations*) και εξάγουμε χαρακτηριστικά για την αναγνώριση και τη μοντελοποίηση των μουσικών σημάτων.

## 4.1 Θεωρία Διαμορφώσεων

Οι μικρο-διαμορφώσεις στη συχνότητα εμφανίζονται με φυσικό τρόπο τόσο στην ανθρώπινη φωνή όσο και στα μουσικά όργανα. Σύμφωνα με τον Bregman [18], τέτοιου είδους διαμορφώσεις του σήματος είναι συχνά πολύ μικρές – κυμαίνονται σε ποσοστά μικρότερα του 1% για ήχους που προέρχονται από το κλαρινέτο, έως περίπου 1% για ήχους φωνής με σταθερό τονικό ύψος (*pitch*), ενώ μεγαλύτερες διακυμάνσεις έως 20% παρατηρούνται κατά τη διάρκεια του βιμπράτο. Ο Bregman αναφέρει επίσης ότι ακόμη μικρότερες διαμορφώσεις της συχνότητας έχουν σημαντικές επιπτώσεις στην αντιληπτική ομαδοποίηση των αρμονικών ενός ήχου.

Θεωρούμε, λοιπόν, ότι το μουσικό σήμα μπορεί να αναπαρασταθεί ως ένας συνδυασμός από «συντονισμούς» (δηλ. ομάδες αρμονικών). Οι συντονισμοί αυτοί αντιστοιχούν προσεγγιστικά στα συστήματα ταλάντωσης που προκύπτουν από χαρακτηριστικά του μουσικού οργάνου ή/και τη γενικότερη διαδικασία παραγωγής των ήχων (γεωμετρία του οργάνου, τρόπος εκτέλεσης ενός μουσικού κομματιού). Έτσι, ορισμένες συχνότητες ενισχύονται ενώ άλλες εξασθενούν.

Προτείνουμε τη μοντελοποίηση και την αναπαράσταση κάθε συντονισμού των σημάτων μουσικής ως ένα σήμα διαμόρφωσης πλάτους και συχνότητας AM-FM:

$$S(t) = \sum_{i=1}^{K} \alpha_i(t) \cos\left(\phi_i(t)\right) \tag{4.1}$$

όπου $\alpha_i$ και $\phi_i$ το στιγμιαίο πλάτος και η φάση της συνιστώσας $i$, και την αναπαράσταση του συνολικού μουσικού σήματος ως μια αθροιστική επαλληλία τέτοιων AM-FM σημάτων.

Σε κάθε AM-FM σήμα, η στιγμιαία συχνότητα μοντελοποιεί τη χρονικά-μεταβαλλόμενη συχνότητα του συντονισμού, ενώ το στιγμιαίο πλάτος ακολουθεί τη χρονικά-μεταβαλλόμενη ενέργεια της ακουστικής πηγής. Το AM-FM μοντέλο μπορεί να εκτιμήσει τη μέση τιμή της συχνότητας, τη στιγμιαία ένταση του πλάτους των συντονισμών καθώς και τη στιγμιαία απόκλιση της συχνότητας. Το πλεονέκτημα της ανάλυσης αυτής είναι ότι οι διαμορφώσεις ανιχνεύουν τη λεπτή δομή και τις γρήγορες διακυμάνσεις των μουσικών σημάτων. Το μοντέλο αυτό μπορεί να εφαρμοστεί σε μικρότερα ή μεγαλύτερα παράθυρα ανάλυσης, διερευνώντας τη δυνατότητα μοντελοποίησης μουσικών χαρακτηριστικών μικρο-, μεσο- και μακροδομών.

### 4.1.1 Χαρακτηριστικά Διαμόρφωσης

Τα AM-FM χαρακτηριστικά που εξετάζονται είναι: (α) το **μέσο Στιγμιαίο Πλάτος** (*mean Instantaneous Amplitude, m-IAM*), το οποίο ορίζεται ως το βραχέος χρόνου μέσο στιγμιαίο πλάτος $|\alpha_i(t)|$ για κάθε συντονισμό $i$. Παραμετροποιεί τα πλάτη των συντονισμών, ενώ επιπλέον μπορεί να ανιχνεύσει μέρος της μη-γραμμικής συμπεριφοράς του σήματος. (β) Η **μέση Στιγμιαία Συχνότητα** (*mean Instantaneous Frequency, m-IFM*) είναι η βραχέος χρόνου σταθμισμένη μέση τιμή της στιγμιαίας συχνότητας $f_i(t)$, η οποία δίνει πληροφορίες σχετικά με τη λεπτή δομή του σήματος. (γ) Το **ποσοστό διαμόρφωσης της συχνότητας** (*Frequency Modulation Percentage, FMP*), το οποίο ορίζεται ως ο λόγος της μέσης διακύμανσης (ή bandwidth) $B_i$ του σήματος στιγμιαίας συχνότητας $f_i(t)$ προς τη μέση τιμή του και παραμετροποιεί τη μέγιστη μεταβολή από τη μέση τιμή της διαμόρφωσης της συχνότητας (το μοντέλο AM-FM θεωρεί πως κατά τη διάρκεια μιας περιόδου η συχνότητα των συντονισμών μεταβάλλεται γύρω από κάποια κεντρική συχνότητα).

Ο υπολογισμός των χαρακτηριστικών αυτών στηρίζεται στον αλγόριθμο διαχωρισμού ενέργειας ESA (*Energy Separation Algorithm*) [104], ο οποίος παρουσιάζει μεγάλη χρονική ακρίβεια.

Έστω ότι το αρχικό μη-στατικό σήμα φιλτράρεται μέσω μιας συστοιχίας ζωνοπερατών φίλτρων και η έξοδος μοντελοποιείται ως ένα AM-FM σήμα. Εφαρμόζοντας τον τελεστή ενέργειας Teager (*Teager Energy Operator*) [174] σε αυτό το σήμα προκύπτει η στιγμιαία ενέργεια πηγής καθώς και το στιγμιαίο πλάτος και η στιγμιαία συχνότητα ως:

$$f(t) \approx \frac{1}{2\pi}\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \tag{4.2}$$

$$|\alpha(t)| \approx \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \tag{4.3}$$

όπου $\Psi[x] = \dot{x}^2 - x\ddot{x}$ και $\dot{x} = dx/dt$.

Τα FMP υπολογίζονται από τη σταθμισμένη μέση τιμή και διακύμανση του σήματος της στιγμιαίας συχνότητας [33], οι εξισώσεις των οποίων δίνονται από [137]:

$$F_i = \frac{\int_0^T f_i(t)\alpha_i^2(t)dt}{\int_0^T \alpha_i^2(t)dt} \tag{4.4}$$

$$B_i = \sqrt{\frac{\int_0^T \left[\dot{\alpha}_i^2(t) + (f_i(t) - F_i)^2\alpha_i^2(t)\right] dt}{\int_0^T \alpha_i^2(t)dt}} \tag{4.5}$$

όπου $i = 1, 2, \ldots, N$ ο αριθμός του συντονισμού, και $T$ το μήκος του παραθύρου ανάλυσης. Έχει βρεθεί ότι η στάθμιση των σημάτων $f_i(t)$ με τα σήματα $|\alpha_i(t)|^2$ δίνει πιο εύρωστες και πιο ομαλές αναπαραστάσεις των μεγεθών $B_i$ και $F_i$ σε περιπτώσεις που οι τιμές του στιγμιαίου πλάτους είναι πολύ μικρές κάτι που έχει ως επακόλουθο να παρουσιάζουν μεγάλες μεταβολές και ασυνέχειες στη στιγμιαία συχνότητα [137].

Στη μελέτη αυτή χρησιμοποιούμε τον αλγόριθμο Gabor-ESA [33], ο οποίος είναι συνδυασμός του συνεχούς χρόνου ESA με ζωνοπερατό φιλτράρισμα (*bandpass filtering*) του σήματος με Gabor φίλτρα. Η επιλογή των φίλτρων Gabor βασίζεται στην καλή χρονο-συχνοτική τους ευκρίνεια [104], ενώ ο αλγόριθμος αυτός δίνει ομαλότερες στιγμιαίες εκτιμήσεις. Με τη χρήση μιας ζωνοπερατής mel-spaced Gabor συστοιχίας 12 φίλτρων (με επικάλυψη των διαδοχικών φίλτρων 50%) δημιουργούνται τα ζωνοπερατά μουσικά σήματα (baseline παράμετροι για τον υπολογισμό των μουσικών σημάτων). Στη συνέχεια αποδιαμορφώνονται και υπολογίζονται τα σήματα στιγμιαίου πλάτους και στιγμιαίας συχνότητας, τα οποία παραθυροποιούνται και υπολογίζονται οι μη-γραμμικές μέσες τιμές τους. Το ζωνοπερατό φιλτράρισμα και η χρονική παραγώγηση με τον τελεστή $\Psi$ συνδυάζονται σε μια συνέλιξη με τη χρονική παράγωγο της κρουστικής απόκρισης του φίλτρου Gabor:

$$\begin{aligned}\Psi[x(t) * g(t)] = &\left[x(t) * \frac{dg(t)}{dt}\right]^2 - \\ &(x(t) * g(t))\left[x(t) * \frac{d^2g(t)}{dt^2}\right],\end{aligned} \tag{4.6}$$

όπου $x(t)$ το σήμα εισόδου και $g(t)$ η κρουστική απόκριση του φίλτρου Gabor.

Καταλήγοντας σημειώνουμε πως για να αποφύγουμε τις πολύ μικρές τιμές και κατά συνέπεια τις ασυνέχειες του στιγμιαίου πλάτους και της στιγμιαίας συχνότητας ορίζουμε ένα κατώφλι, έτσι ώστε τιμές της στιγμιαίας ενέργειας μικρότερες του κατωφλίου αυτού μηδενίζονται. Ασυνέχειες τέτοιου είδους παρατηρήσαμε κυρίως κατά τον πειραματισμό μας με συστοιχίες Gabor φίλτρων, οι οποίες αποτελούνται από φίλτρα με πολύ μικρές κεντρικές συχνότητες και bandwidth, όπως και περιγράφονται στην Εν. 4.3.1. Τέλος, αντισταθμίζουμε την ύπαρξη μηδενικών του στιγμιαίου πλάτος, υπολογίζοντας τις μέσες τιμές του, σε κάθε παράθυρο ανάλυσης, μόνο από τις «υγιείς» τιμές του σήματος, και άρα μη λαμβάνοντας υπόψη τις όποιες μηδενικές τιμές.

Στη συνέχεια του κεφαλαίου, παρουσιάζουμε πειραματικά αποτελέσματα τα οποία βασίζονται στον υπολογισμό μετρήσεων των χαρακτηριστικών διαμόρφωσης για την κατηγοριοποίηση των διαφορετικών μουσικών οργάνων αλλά και των διαφορετικών ειδών μουσικής.

## 4.2 Πειράματα Αναγνώρισης Μουσικών Οργάνων

**Βάση Δεδομένων**

Έχουν διεξαχθεί δύο σετ πειραμάτων όπου χρησιμοποιήθηκαν (1) επτά διαφορετικά όργανα (1331 νότες), τα οποία είναι τα εξής: το κοντραμπάσο, το φαγκότο, το τσέλο, το B♭ κλαρινέτο, το φλάουτο, το κόρνο και η τούμπα και (2) πέντε επιπλέον όργανα (738 νότες): το άλτο σαξόφωνο, το μπάσο τρομπόνι, το τενόρο τρομπόνι, η τρομπέτα και το όμποε (σύνολο 12 όργανα και 2369 νότες). Οι νότες που χρησιμοποιήθηκαν αποτελούν το πλήρες εύρος συχνοτήτων των οργάνων και καλύπτουν τη δυναμική ένδειξη από piano έως forte (Fs = 44.1 kHz). Η παραμετροποίηση των δύο σετ πειραμάτων είναι ίδια και περιγράφεται παρακάτω.

### 4.2.1 Πειραματική Αξιολόγηση: Σύνολα Χαρακτηριστικών

Στα πειράματα που ακολουθούν αξιολογούμε την απόδοση των διαφορετικών συνόλων χαρακτηριστικών, τα οποία παρατίθενται στον Πίνακα 4.1. Τα χαρακτηριστικά m-IAM και m-IFM υπολογίζονται σε πλαίσια των 30 ms με επικάλυψη των 15 ms από τα σήματα στιγμιαίου πλάτους και στιγμιαίας συχνότητας (βλ. Εν. 4.1). Για το ζωνοπερατό φιλτράρισμα των μουσικών σημάτων χρησιμοποιήθηκαν δώδεκα φίλτρα Gabor, καθώς ύστερα από εκτενή πειραματισμό καταλήξαμε ότι είναι η καλύτερη επιλογή αριθμού φίλτρων. Εκτός των m-IAM και m-IFM υπολογίζουμε και την πρώτη και δεύτερη χρονική παράγωγο τους με αποτέλεσμα το διάνυσμα $\text{AMFM}_{\Delta}$ 72 χαρακτηριστικών, η διάσταση του οποίου μειώνεται με τη χρήση PCA ανάλυσης. Εξετάστηκαν αρκετοί διαφορετικοί

*Πίνακας 4.1:* Λίστα AM-FM χαρακτηριστικών για αναγνώριση μουσικών οργάνων.

| **Σύνολα χαρακτηριστικών** | |
|---|---|
| 1 | AMFM (12 m-IAM + 12 m-IFM) |
| 2 | $\text{AMFM}_{\Delta}$ (12 m-IAM+12 m-IFM (+ their 12 $\Delta$ + 12 $\Delta\Delta$)) |
| 3 | $\text{AMFM}_{50}$ (50 AMFM χαρακτηριστικά μετά από ανάλυση PCA) |
| 4 | $\text{AMFM}_{39}$ (39 AMFM χαρακτηριστικά μετά από ανάλυση PCA) |
| 5 | $\text{MFCC}_{\Delta}$ (13 MFCC + 13 $\Delta$ + 13 $\Delta\Delta$) |
| Multi-Stream Cases | |
| 1 | $\text{AMFM}_{\Delta}$ + $\text{MFCC}_{\Delta}$ |
| 2 | $\text{AMFM}_{39}$ + $\text{MFCC}_{\Delta}$ |
| 3 | $\text{AMFM}_{50}$ + $\text{MFCC}_{\Delta}$ |

συνδυασμοί του αριθμού των PCA συνιστωσών ώστε να βρεθεί ο καλύτερος. Η μελέτη μας έδειξε ότι τα m-IFM ήταν λιγότερο συσχετισμένα, γι' αυτό και κρατήθηκαν περισσότερα. Τα σύνολα των χαρακτηριστικών τα οποία παρουσιάζονται στη συνέχεια απέδωσαν τη μεγαλύτερη μείωση λάθους σε σύγκριση με τα $\text{MFCC}_{\Delta}$.

Συγκεκριμένα, χρησιμοποιήθηκαν δύο σύνολα χαρακτηριστικών. Το πρώτο αποτελείται συνολικά από 50 PCA συνιστώσες και συγκεκριμένα από 18 m-IAM (6 m-IAM, 6 m-$\text{IAM}_{\Delta}$, 6 m-$\text{IAM}_{\Delta\Delta}$) και 32 m-IFM (12 m-IFM, 10 m-$\text{IFM}_{\Delta}$, 10 m-$\text{IFM}_{\Delta\Delta}$). Το δεύτερο αποτελείται από 39 PCA συνιστώσες, μιας και σκοπός μας είναι ο αριθμός χαρακτηριστικών που τελικά χρησιμοποιείται να είναι συγκρίσιμος σε μέγεθος με τα 39 $\text{MFCC}_{\Delta}$. Το δεύτερο αυτό σύνολο αποτελείται από 12-m-IAM (4 m-IAM, 4 m-$\text{IAM}_{\Delta}$, 4 m-$\text{IAM}_{\Delta\Delta}$) και 27 m-IFM (12 m-IFM, 8 m-$\text{IFM}_{\Delta}$, 7 m-$\text{IFM}_{\Delta\Delta}$).

Τα διαφορετικά σύνολα χαρακτηριστικών αξιολογήθηκαν με τη χρήση στατικών (GMM) και δυναμικών (HMM) ταξινομητών, με ποικίλους συνδυασμούς καταστάσεων $N$ [3-9] και μειγμάτων $M$ [1-3]. Σε όλες τις περιπτώσεις τα δεδομένα εκπαίδευσης έχουν επιλεγεί τυχαία και αποτελούν το 70% των διαθέσιμων ήχων, και τα αποτελέσματα που παρουσιάζονται έχουν προκύψει ύστερα από πέντε cross-validation. Η απόδοση των χαρακτηριστικών που τελικά επιλέχθηκαν συγκρίθηκε με τα MFCC (13 μαζί με την ενέργεια), σε ένα πρώτο σετ πειραμάτων μόνα τους, ενώ στη συνέχεια προσθέσαμε την πρώτη και τη δεύτερη χρονική τους παράγωγο. Η ανάλυση των MFCC διεξήχθη με τη χρήση παραθύρων των 30 ms με 15 ms επικάλυψη, και με 24 τριγωνικά ζωνοπερατά φίλτρα. Επιπλέον, διεξαγάγαμε multi-stream πειράματα για τη μοντελοποίηση και σύμμειξη των δύο διαφορετικών συνόλων χαρακτηριστικών μεταβάλλοντας το βάρος κάθε ροής.

### 4.2.2 Πειραματική Αξιολόγηση: Αποτελέσματα

Βάσει των αποτελεσμάτων αναγνώρισης των διαφορετικών χαρακτηριστικών, αποδείχθηκε πως τα AMFM μπορούν να αποδώσουν καλύτερα από ό,τι τα $\text{MFCC}_{\Delta}$

*Πίνακας 4.2:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) για 7 και 12 μουσικά όργανα, όπου $N$ ο αριθμός των καταστάσεων και $M$ ο αριθμός των μειγμάτων. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 4.1.

| **Ποσοστά επιτυχίας κατηγοριοποίησης για 7 όργανα** | | | | |
|---|---|---|---|---|
| | Βάρη | GMM | HMM | |
| Χαρακτηριστικά | MFCC-AMFM | $M=3$ | $N=3$ $M=3$ | $N=5$ $M=3$ |
| AMFM | - | 88.74 | 94.90 | 95.00 |
| $\text{AMFM}_{\Delta}$ | - | 89.14 | 94.60 | 96.72 |
| $\text{AMFM}_{50}$ | - | **95.30** | 96.31 | 96.06 |
| $\text{AMFM}_{39}$ | - | 94.29 | **96.41** | **96.77** |
| $\text{MFCC}_{\Delta}$ | - | 86.06 | 94.65 | 96.16 |
| Multi-Stream Cases | | | | |
| $\text{MFCC}_{\Delta}$ - $\text{AMFM}_{\Delta}$ | 1.0 - 0.5 | **91.57** | **96.67** | 97.57 |
| | 0.5 - 1.0 | 90.50 | 95.66 | 97.17 |
| | 0.5 - 0.5 | 90.91 | 96.61 | **97.93** |
| | 1.0 - 0.1 | 90.10 | 96.52 | 96.97 |
| $\text{MFCC}_{\Delta}$ - $\text{AMFM}_{50}$ | 1.0 - 0.5 | 95.81 | **98.33** | **98.64** |
| | 0.5 - 1.0 | **96.26** | 97.78 | 97.67 |
| | 0.5 - 0.5 | **96.26** | 97.88 | 97.83 |
| | 1.0 - 0.1 | 89.60 | 96.46 | 97.73 |
| $\text{MFCC}_{\Delta}$ - $\text{AMFM}_{39}$ | 1.0 - 0.5 | 95.46 | **98.48** | **98.68** |
| | 0.5 - 1.0 | **96.31** | 97.82 | 98.13 |
| | 0.5 - 0.5 | 96.16 | 98.18 | 98.18 |
| | 1.0 - 0.1 | 90.71 | 96.61 | 97.37 |
| **Ποσοστά επιτυχίας κατηγοριοποίησης για 12 όργανα** | | | | |
| $\text{AMFM}_{50}$ | - | **85.46** | 91.74 | **93.72** |
| $\text{AMFM}_{39}$ | - | 82.38 | **92.68** | **93.72** |
| $\text{MFCC}_{\Delta}$ | - | 79.09 | 88.23 | 90.60 |
| Multi-Stream Cases | | | | |
| $\text{MFCC}_{\Delta}$ - $\text{AMFM}_{50}$ | 1.0 - 0.5 | 88.55 | 94.37 | 95.32 |
| | 0.5 - 1.0 | 87.19 | **94.73** | 94.96 |
| | 0.5 - 0.5 | **88.03** | 94.50 | **95.55** |
| | 1.0 - 0.1 | 86.18 | 92.33 | 94.02 |
| $\text{MFCC}_{\Delta}$ - $\text{AMFM}_{39}$ | 1.0 - 0.5 | **87.64** | **95.19** | **95.89** |
| | 0.5 - 1.0 | 85.33 | 94.67 | 95.45 |
| | 0.5 - 0.5 | 86.70 | 94.93 | 95.67 |
| | 1.0 - 0.1 | 85.73 | 92.55 | 93.33 |

στις περισσότερες περιπτώσεις (ακόμα και σε αυτές που δεν παρουσιάζονται εδώ). Τα πιο αντιπροσωπευτικά αποτελέσματα (ποσοστά μέσης ακρίβειας αναγνώρισης %, mean accuracy) για τα δύο σετ πειραμάτων παρουσιάζονται στον Πίνακα 4.2. Παρατηρούμε ότι τα $\text{AMFM}_{\Delta}$ έδειξαν μεγαλύτερη ικανότητα αναγνώρισης απ' ό,τι τα $\text{MFCC}_{\Delta}$ με μείωση σφάλματος 15% για $N=5$, $M=3$. Η καλύτερη περίπτωση των $\text{AMFM}_{39}$ απέδωσε μείωση σφάλματος έως 60% (15%), 33% (38%) και 16% (33%) για τα GMM και τα HMM για $N=3,5$ και $M=3$, για 7 και 12 όργανα (η μείωση σφάλματος για 12 όργανα φαίνεται στις παρενθέσεις). Δεδομένων των αποτελεσμάτων αυτών, θεωρούμε πως

τα AMFM χαρακτηριστικά ευνοούν την αναγνώριση μεταξύ των εξεταζόμενων οργάνων.

Ο συνδυασμός των προτεινόμενων χαρακτηριστικών $\text{AMFM}_{39}$ με τα $\text{MFCC}_{\Delta}$ επιφέρει ακόμη μεγαλύτερη μείωση του σφάλματος, η οποία φτάνει περίπου το 60% και το 56% σε σύγκριση με τα $\text{MFCC}_{\Delta}$ για 7 και 12 μουσικά όργανα, αντίστοιχα. Τα αποτελέσματα σε σχέση με τα βάρη που χρησιμοποιήθηκαν για το πειραματισμό κατά τη σύμμειξη των χαρακτηριστικών είναι συγκρίσιμα, με ελαφρώς καλύτερη την περίπτωση όπου τα βάρη ισούνται με $s_1 = 1.0$ για τα $\text{MFCC}_{\Delta}$ και $s_2 = 0.5$ για τα διάφορα σύνολα AMFM. Ωστόσο, για την περίπτωση όπου τα $\text{MFCC}_{\Delta}$ έχουν βάρος ίσο με $s_1 = 1.0$ και τα AMFM $s_2 = 0.1$ πρέπει να αναφέρουμε πως η ακρίβεια αναγνώρισης είναι πολύ χαμηλότερη, κάτι που ενισχύει το γεγονός ότι τα AMFM χαρακτηριστικά συμβάλλουν σημαντικά στην αναγνώριση των οργάνων. Επιπλέον, διαπιστώνουμε ότι η μοντελοποίηση με HMM επιτυγχάνει καλύτερα αποτελέσματα, αν και η μείωση του σφάλματος για τα προτεινόμενα χαρακτηριστικά σε σύγκριση με τα MFCC είναι υψηλότερη στις περιπτώσεις ταξινόμησης με τη χρήση των GMM.

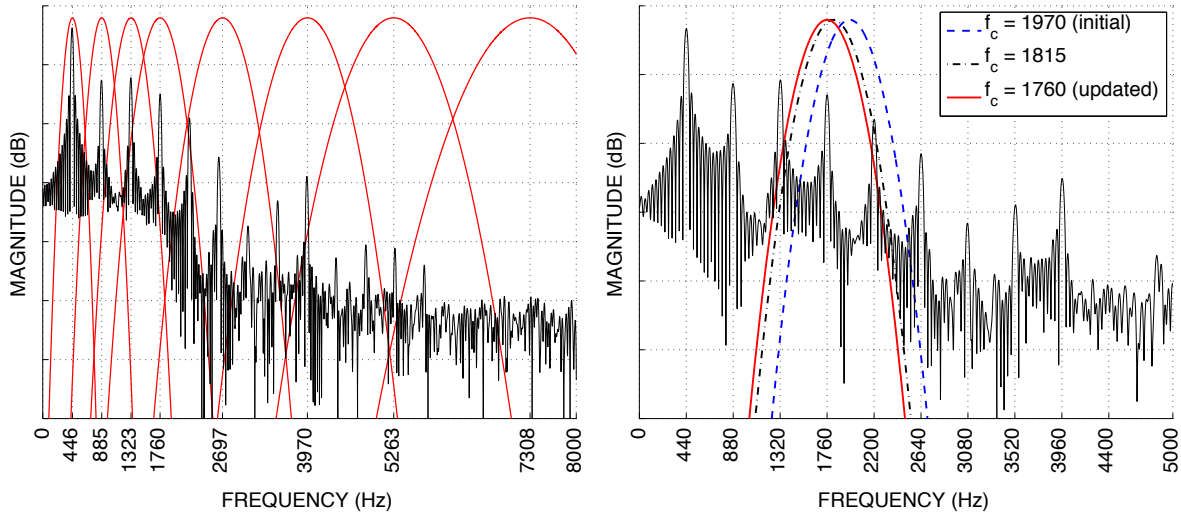
Αξίζει επίσης να σχολιάσουμε πως τα προτεινόμενα χαρακτηριστικά στο δεύτερο σετ πειραμάτων είχαν αξιοσημείωτη αναγνώριση, παρά την ύπαρξη των οργάνων μπάσο τρομπόνι και τενόρο τρομπόνι, που πρακτικά θα μπορούσαν να θεωρηθούν το ίδιο όργανο.

### 4.2.3 Επαναληπτικός ESA (*Iterative-ESA*)

Στο σημείο αυτό εφαρμόζουμε τον Επαναληπτικό-ESA (*Iterative-ESA*) για την εκτίμηση των κεντρικών συχνοτήτων $f_c$ των φίλτρων Gabor [59]. Η μέθοδος αυτή συνεπάγεται την επαναληπτική εφαρμογή του ESA στο φιλτραρισμένο σήμα και την προσαρμογή της κεντρικής συχνότητας κάθε φίλτρου υστέρα από κάθε επανάληψη. Η μέθοδος είναι σημαντική γιατί μειώνει τη σπουδαιότητα καλών αρχικών εκτιμήσεων των κεντρικών συχνοτήτων των φίλτρων.

Για την ανάλυση με τον επαναληπτικό-ESA υπολογίζουμε τη βραχέος χρόνου στιγμιαία συχνότητα των μουσικών ήχων χρησιμοποιώντας πλαίσια ανάλυσης των 30 ms. Χρησιμοποιούμε τις νότες A3 και A4, με θεμελιώδη συχνότητα ίση με 220 Hz και 440 Hz αντίστοιχα, από τα όργανα B♭ κλαρινέτο, σαξόφωνο, βιολί και φλάουτο. Ξεκινάμε τον αλγόριθμο χρησιμοποιώντας κεντρικές συχνότητες σύμφωνες με την κλίμακα mel, ενημερώνοντας (*updating*) την καθεμία ύστερα από κάθε επανάληψη του ESA, και διατηρώντας το εύρος ζώνης σταθερό. Θεωρούμε πως ο αλγόριθμος έχει συγκλίνει όταν η κεντρική συχνότητα κάθε φίλτρου δεν μεταβάλλεται περισσότερο από 1% ή όταν φτάσει έναν ορισμένο αριθμό επαναλήψεων. Σύγκλιση επιτυγχάνεται κατά μέσο όρο ύστερα από τέσσερις επαναλήψεις για τα χαμηλότερης συχνότητας φίλτρα, ενώ περισσότερες επαναλήψεις απαιτούνται για φίλτρα υψηλής συχνότητας.

Παρατηρούμε ότι κατά τη διάρκεια αυτής της επαναληπτικής διαδικασίας, οι κεντρικές




*Σχήμα 4.1:* (α′) Συστοιχία φίλτρων Gabor (φίλτρα 2–9) με τις εκτιμώμενες κεντρικές συχνότητες $f_c$ μετά την εφαρμογή του επαναληπτικού–ESA μαζί με το φάσμα της νότας A4 του B♭ κλαρινέτου για ένα πλαίσιο ανάλυσης 30 ms ($Fs$ = 44.1 Hz). (β′) Φάσμα της νότας A4 του B♭ κλαρινέτου για ένα πλαίσιο ανάλυσης των 30 ms μαζί με το πέμπτο φίλτρο Gabor. Ο αλγόριθμος Iterative-ESA εφαρμόστηκε για το πέμπτο φίλτρο με αρχική κεντρική συχνότητα $f_c = 1970$ Hz και υστέρα από δύο επαναλήψεις συνέκλινε στη συχνότητα $f_c = 1760$ Hz η οποία είναι $4f_0$ για τη νότα A4 (με διαφορά 210 Hz).

συχνότητες τείνουν να συγκλίνουν σε συχνότητες που βρίσκονται κοντά σε ακέραια πολλαπλάσια της θεμελιώδους συχνότητας των τόνων ανάλυσης, δηλ. στις αρμονικές. Το Σχήμα 4.1 (α) δείχνει τη συστοιχία φίλτρων Gabor για τα φίλτρα 2–9, με τις ανανεωμένες κεντρικές συχνότητες $f_c$ μαζί με το φάσμα της νότας A4 του B♭ κλαρινέτου σε ένα πλαίσιο ανάλυσης των 30 ms ($f_0$: 440 Hz). Διαπιστώνουμε πως οι συχνότητες αυτές είναι όντως πολύ κοντινά ακέραια πολλαπλάσια της $f_0$ και συγκεκριμένα συμφωνούν με τις εξής αρμονικές του σήματος: $f_0$, $2f_0$, $3f_0$, $4f_0$, $6f_0$, $9f_0$, $12f_0$ και $17f_0$. Στο Σχήμα 4.1 (β) βλέπουμε τη διαδικασία της σύγκλισης για το πέμπτο φίλτρο Gabor μαζί με το φάσμα της νότας A4 του B♭ κλαρινέτου σε ένα πλαίσιο ανάλυσης των 30 ms. Η αρχική κεντρική συχνότητα ισούται με 1970 Hz, και ύστερα από δύο επαναλήψεις συγκλίνει στη συχνότητα $f_{c5}$ = 1760 Hz που είναι η τέταρτη αρμονική ($4f_0$) της νότας A4. Παρόμοια αποτελέσματα λάβαμε από την ανάλυση και των υπόλοιπων οργάνων.

Δεύτερη σημαντική παρατήρηση της ανάλυσής μας είναι ότι μερικά από τα φίλτρα Gabor συγκλίνουν στην ίδια κεντρική συχνότητα. Θωρούμε πως αυτό οφείλεται είτε στις αρχικά επιλεγμένες κεντρικές συχνότητες είτε σχετίζεται με τις ιδιότητες του σήματος στις συγκεκριμένες συχνότητες, όπου παρατηρήσαμε αρμονικές μικρότερου πλάτους. Ωστόσο, θεωρούμε ότι τα συμπεράσματά μας είναι σημαντικά για δύο λόγους. Η μέθοδος αυτή θα μπορούσε (α) να έχει ως αποτέλεσμα καλύτερες εκτιμήσεις των $|\alpha(t)|$ και $f(t)$ και (β) να

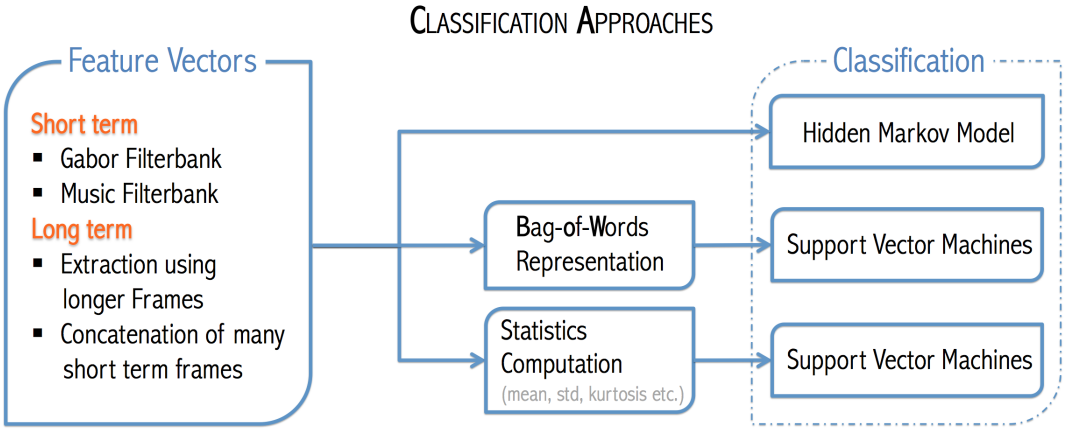
εκτιμήσει πιθανώς το αρμονικό περιεχόμενο της νότας δεδομένης της σύγκλιση των $f_c$ σε συχνότητες πολύ κοντινές με τις αρμονικές του μουσικού σήματος, μολονότι δεν υπάρχει πρότερη γνώση του σήματος που αναλύεται.

**Συμπεράσματα**

Παρουσιάσαμε και προτείναμε τη χρήση ενός μη-γραμμικού μοντέλου διαμορφώσεων AM-FM των μουσικών σημάτων, υποκινούμενοι από παρόμοιες ιδέες οι οποίες με επιτυχία χρησιμοποιήθηκαν και εφαρμόστηκαν στην αναγνώριση φωνής καθώς και στην κατηγοριοποίηση φωνής και μουσικής. Με βάση τα αποτελέσματα αναγνώρισης και τη δεδομένη μείωση του σφάλματος έως και 56% για 12 όργανα για τον συνδυασμό των AMFM χαρακτηριστικών με τα MFCC, θεωρούμε ότι η προτεινόμενη μέθοδος είναι σε θέση να προσδιορίσει τη δομή και τα χαρακτηριστικά των διαφορετικών μουσικών οργάνων. Επιπλέον, η ανάλυση με τον επαναληπτικό-ESA είναι πολλά υποσχόμενη, και με περαιτέρω διερεύνηση θα μπορέσουμε να δούμε εάν εκτός της κατηγοριοποίησης των μουσικών οργάνων μπορεί επίσης να εκτιμήσει και το αρμονικό περιεχόμενο των ήχων αυτών.

## 4.3 Πειράματα Αναγνώρισης Ειδών Μουσικής

Το πρόβλημα της αναγνώρισης και κατηγοριοποίησης των διαφορετικών ειδών μουσικής εμπεριέχει μεγάλη πολυπλοκότητα, όπως αναδείξαμε σε προηγούμενο κεφάλαιο. Παρ' όλα αυτά οι πιο πρόσφατες ερευνητικές εργασίες επιτυγχάνουν ιδιαίτερα καλά ποσοστά αναγνώρισης. Οι δυσκολίες που παρουσιάζονται στην αυτόματη κατηγοριοποίηση των ειδών οφείλονται τόσο στην ύπαρξη επικαλύψεων των διαφορετικών ειδών και των υποκατηγοριών τους, όσο και στον τρόπο δημιουργίας των σχετικών βάσεων. Οι περισσότερες ευρέως διαδεδομένες βάσεις δεδομένων βασίζονται κυρίως στις μουσικές γνώσεις του κάθε ερευνητή και όχι σε κάποια διεξοδική και εμπεριστατωμένη αξιολόγηση. Σε αυτό το κεφάλαιο παρουσιάζουμε τις διαφορετικές μεθοδολογίες που εξετάστηκαν και αξιολογήθηκαν για την εφαρμογή της αναγνώρισης των ειδών της μουσικής. Στο Σχήμα 4.2 παρουσιάζονται οι διαφορετικές προσεγγίσεις που εξετάστηκαν τόσο όσον αφορά τις εναλλακτικές αναπαραστάσεις των μουσικών σημάτων όσο και τις τεχνικές αναγνώρισης.

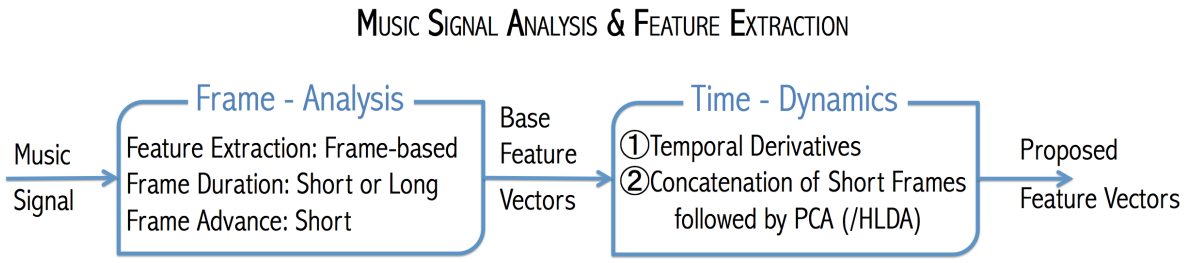



*Σχήμα 4.2:* Διάγραμμα με τις εναλλακτικές προσεγγίσεις όσον αφορά τις διαφορετικές αναπαραστάσεις των μουσικών σημάτων διαφορετικών ειδών μουσικής και τις τεχνικές αναγνώρισης.

### Βάση Δεδομένων

Στα πειράματα που έχουν διεξαχθεί πειραματιζόμαστε με τη βάση δεδομένων GTZAN [179], για λεπτομέρειες βλ. Ενότητα 2.5. Συνοπτικά αναφέρουμε πως αποτελείται από 10 διαφορετικά είδη μουσικής τα οποία περιγράφονται ως: blues, country, classical, disco, hiphop, jazz, metal, pop, reggae, και rock.

Η βάση GTZAN αποτελείται από αποσπάσματα μουσικών κομματιών διάρκειας 30 sec το καθένα. Για να μειώσουμε την πολυπλοκότητα των συστημάτων εκπαίδευσης αλλά και την πολυπλοκότητα κατά την εξαγωγή των χαρακτηριστικών (όπως για παράδειγμα κατά




*Σχήμα 4.3:* Διαφορετικές εναλλακτικές μεθοδολογίες για την ανάλυση των μουσικών σημάτων και την εξαγωγή των προτεινόμενων χαρακτηριστικών.

την επεξεργασία με τα μοντέλα διαμορφώσεων, όπου το στιγμιαίο πλάτος και η στιγμιαία συχνότητα υπολογίζονται σε ολόκληρο το αρχείο ήχου) επεξεργαζόμαστε μικρότερα αποσπάσματα των κομματιών, χωρίζοντάς τα σε τρία τμήματα των 10 δευτερολέπτων το καθένα. Κατ' αυτόν τον τρόπο επιτυγχάνουμε την επέκταση της βάσης (1000×3), θεωρώντας τις τριπλέτες που έχουν δημιουργηθεί ως «διαφορετική εκδοχή» του ίδιου μουσικού κομματιού.

### 4.3.1 Ανάλυση των AM-FM Χαρακτηριστικών σε Πολυφωνικά Σήματα Διαφορετικών Ειδών Μουσικής

Για την πειραματική αξιολόγηση των χαρακτηριστικών έχουν χρησιμοποιηθεί διαφορετικές προσεγγίσεις όσον αφορά την ανάλυση των μουσικών σημάτων και κατ' επέκταση την εξαγωγή αναπαραστάσεων και συνόλων χαρακτηριστικών. Το βασικό σύνολο χαρακτηριστικών που χρησιμοποιήθηκε σε όλες τις περιπτώσεις αξιολόγησης αποτελείται από το λογαριθμικό μέσο στιγμιαίο πλάτος (log(m-IAM)), τη μέση στιγμιαία συχνότητα (m-IFM) και το ποσοστό διαμόρφωσης της συχνότητας (FMP).

Συγκεκριμένα, στόχος της επεξεργασίας που ακολουθούμε είναι η ανάλυση τόσο των μικροδομών όσο και των μακροδομών των μουσικών σημάτων. Αυτό πρακτικά μας ωθεί στην εύρεση διαφορετικών μεθοδολογιών για την εξαγωγή χαρακτηριστικών, βλ. Σχήμα 4.3. Η κλασική προσέγγιση για τη δημιουργία αναπαραστάσεων του σήματος είναι οι διάφορες μετρήσεις να λαμβάνονται με βραχέος χρόνου παράθυρα ανάλυσης τα οποία, στην επεξεργασία μουσικών σημάτων, συνήθως κυμαίνονται μεταξύ 25-50 ms (*short term analysis*). Ένας τρόπος επομένως για να εξετάσουμε τις μακροδομές των σημάτων είναι η λήψη μετρήσεων και επεξεργασία των μουσικών σημάτων με παράθυρα ανάλυσης μεγαλύτερης διάρκειας. Στην προκειμένη περίπτωση, τα προτεινόμενα χαρακτηριστικά τα οποία και παρουσιάζουμε στη συνέχεια έχουν εξαχθεί είτε με τη χρήση παραθύρων των 30 ms με 50% επικάλυψη, είτε με τη χρήση παραθύρων των 125 ή 200 ms με 80% επικάλυψη. Και στις δύο περιπτώσεις τα διανύσματα χαρακτηριστικών έχουν επαυξηθεί

με τη χρονική πληροφορία μέσω των πρώτων και δεύτερων χρονικών παραγώγων τους. Για τη δημιουργία εύρωστων αναπαραστάσεων, μειώνουμε τη διάσταση των χαρακτηριστικών χρησιμοποιώντας PCA ανάλυση. Σημειώνουμε πως σε κάποιες περιπτώσεις διενεργούμε την ανάλυση στα ξεχωριστά μεμονωμένα σύνολα χαρακτηριστικών (δηλ. μόνο στα m-IAM ή στα m-IFM και ξεχωριστά στα σύνολα της κάθε παραγώγου), ενώ σε άλλες περιπτώσεις στο συνολικό συνενωμένο σύνολο χαρακτηριστικών. Με τη χρήση της PCA ανάλυσης δημιουργούνται σετ χαρακτηριστικών τα οποία είναι ασυσχέτιστα ενώ ταυτοχρόνως παρουσιάζουν την μέγιστη διακύμανση μεταξύ τους.

Η χρονική πληροφορία των σημάτων (καθώς και οι μακροδομές τους) είναι εφικτό να εντοπιστεί και με έναν δεύτερο τρόπο. Σε αυτή την περίπτωση πραγματοποιούμε τη συνένωση πολλών διαδοχικών παραθύρων της βραχέος χρόνου ανάλυσης και τα αντιμετωπίζουμε ως ένα. Εν συνεχεία εφαρμόζουμε κάποια τεχνική μείωσης της διάστασης του χώρου χαρακτηριστικών (για παράδειγμα PCA ή HLDA (*Heteroscedastic Linear Discriminant Analysis*). Ο συγκεκριμένος τρόπος προσέγγισης δημιουργίας αναπαραστάσεων, όπου η μείωση της διάστασης γίνεται με HLDA, έχει χρησιμοποιηθεί σε εφαρμογές αναγνώρισης φωνής με επιτυχία [67, 160]. Στη διδακτορική αυτή έρευνα πειραματιστήκαμε συνενώνοντας διαδοχικά παράθυρα με συνολική διάρκεια 1/8, 1/4, 1/2 και 1 sec τα οποία και αναλογούν σε 8, 15, 33 και 65 πλαίσια ανάλυσης αντίστοιχα. Για τη μείωση της διάστασης των χαρακτηριστικών εφαρμόσαμε PCA ανάλυση και πειραματιστήκαμε με τον τελικό αριθμό κύριων συνιστωσών. Σημειώνουμε πως για τη δημιουργία αυτών των συνόλων χαρακτηριστικών δεν χρησιμοποιούμε τις χρονικές παραγώγους των σημάτων.

Παράλληλα, μία άλλη παράμετρος την οποία επιλέξαμε να αξιολογήσουμε, κατά την επεξεργασία των σημάτων μουσικής, είναι το πλήθος και το εύρος των Gabor φίλτρων που χρησιμοποιήθηκαν για τη δημιουργία των ζωνοπερατών μουσικών σημάτων. Στην πρώτη περίπτωση χρησιμοποιήθηκε η baseline παραμετροποίηση με τη χρήση μιας ζωνοπερατής mel-spaced Gabor συστοιχίας 12 φίλτρων (με επικάλυψη των διαδοχικών φίλτρων 50%) ενώ στη δεύτερη περίπτωση εξετάσαμε την κατασκευή μιας «*μουσικής*» συστοιχίας φίλτρων Gabor, η κεντρική συχνότητα των οποίων ορίζεται από τη συχνότητα των μουσικών τόνων. Δημιουργήθηκαν δύο διαφορετικές μουσικές συστοιχίες φίλτρων, η πρώτη εκ των οποίων ξεκινάει από τη δεύτερη οκτάβα και αποτελείται από 89 φίλτρα, περιλαμβάνοντας τις συχνότητες από 65.41Hz (C2) - 10548Hz, ενώ η δεύτερη ξεκινάει από τη πρώτη οκτάβα και αποτελείται από 111 φίλτρα, περιλαμβάνοντας τις συχνότητες από 32.70Hz (C1) - 10548Hz. Επιπλέον πειραματισμός διενεργήθηκε σε σχέση με το εύρος των φίλτρων το οποίο αρχικά επιλέγεται να εκτείνεται από την κεντρική συχνότητα της προηγούμενης νότας ως την κεντρική συχνότητα της επόμενης. Δηλαδή για ένα φίλτρο με συχνότητα $f_i$, το εύρος ορίζεται ως $b_{1i} = [f_{i-1}, f_{i+1}]$. Για λόγους εύρεσης του κατάλληλου μεγέθους

*Πίνακας 4.3:* Θεμελιώδεις συχνότητες των μουσικών τόνων και για τις 8 οκτάβες.

| | **Νότα** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **C** | **C#** | **D** | **D#** | **E** | **F** | **F#** | **G** | **G#** | **A** | **A#** | **B** |
| **0** | 16.35 | 17.32 | 18.35 | 19.45 | 20.60 | 21.83 | 23.12 | 24.50 | 25.96 | 27.50 | 29.14 | 30.87 |
| **1** | **32.70** | 34.65 | 36.71 | 38.89 | 41.20 | 43.65 | 46.25 | 49.00 | 51.91 | 55.00 | 58.27 | 61.74 |
| **2** | **65.41** | 69.30 | 73.42 | 77.78 | 82.41 | 87.31 | 92.50 | 98.00 | 103.8 | 110.0 | 116.5 | 123.5 |
| **3** | 130.8 | 138.6 | 146.8 | 155.6 | 164.8 | 174.6 | 185.0 | 196.0 | 207.7 | 220.0 | 233.1 | 246.9 |
| **4** | 261.6 | 277.2 | 293.7 | 311.1 | 329.6 | 349.2 | 370.0 | 392.0 | 415.3 | 440.0 | 466.2 | 493.9 |
| **5** | 523.3 | 554.4 | 587.3 | 622.3 | 659.3 | 698.5 | 740.0 | 784.0 | 830.6 | 880.0 | 932.3 | 987.8 |
| **6** | 1046 | 1108 | 1174 | 1244 | 1318 | 1396 | 1479 | 1567 | 1661 | 1760 | 1864 | 1975 |
| **7** | 2093 | 2217 | 2349 | 2489 | 2637 | 2793 | 2959 | 3135 | 3322 | 3520 | 3729 | 3951 |
| **8** | 4186 | 4434 | 4698 | 4978 | 5274 | 5587 | 5919 | 6271 | 6644 | 7040 | 7458 | 7902 |

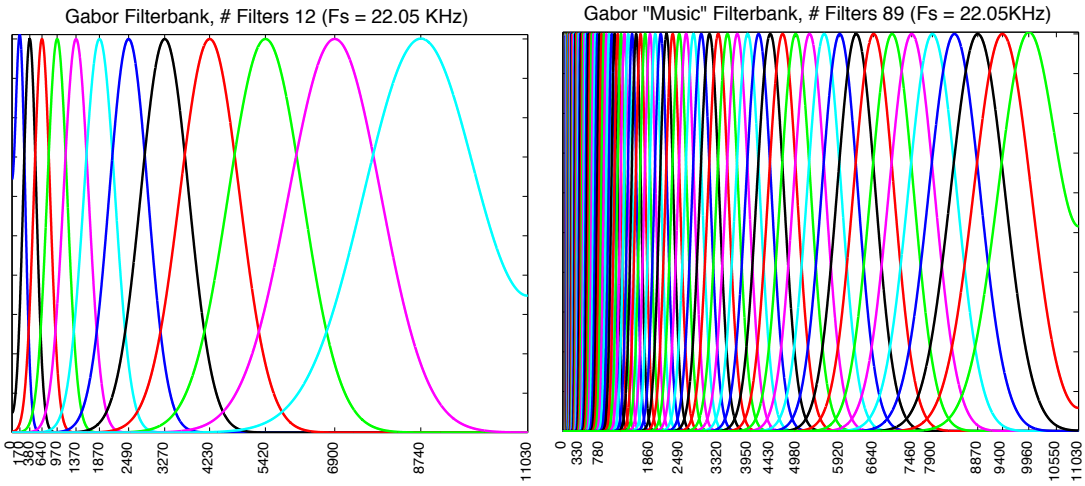
εύρους πειραματιστήκαμε περαιτέρω μεταβάλλοντάς το και συγκεκριμένα διπλασιάζοντας, τετραπλασιάζοντας ή εξαπλασιάζοντάς το. Τα χαρακτηριστικά που εξάγονται είναι το μέσο στιγμιαίο πλάτος ($\log$(m-IAM)) και η μέση στιγμιαία συχνότητα (m-IFM), υπολογισμένα σε βραχέος χρόνου παράθυρα ανάλυσης των 30 ms με επικάλυψη 15 ms από τα σήματα στιγμιαίου πλάτους και στιγμιαίας συχνότητας. Εκτός των m-IAM και m-IFM υπολογίζεται η πρώτη και η δεύτερη χρονική παράγωγος των χαρακτηριστικών. Λόγω του πολυπληθή χώρου χαρακτηριστικών, 534 (89×6) και 606 (101×6) χαρακτηριστικά αντίστοιχα για τις δύο διαφορετικές συστοιχίες φίλτρων, η διάστασή τους μειώνεται με τη χρήση της PCA ανάλυσης και διαφορετικοί αριθμοί PCA συνιστωσών εξετάζονται για την εύρεση του καλύτερου συνόλου χαρακτηριστικών.

Στον Πίνακα 4.3 παρουσιάζονται οι θεμελιώδεις συχνότητες των μουσικών τόνων που χρησιμοποιήθηκαν για τη δημιουργία της μουσικής συστοιχίας φίλτρων και για τις 8 οκτάβες. Παρατηρώντας τις συχνότητες αυτές βλέπουμε πως αυξάνονται λογαριθμικά. Ως γνωστόν στη σύγχρονη δυτική μουσική το φάσμα των συχνοτήτων χωρίζεται σε οκτάβες. Οι νότες που απέχουν μία οκτάβα έχουν το ίδιο όνομα και την ίδια ακουστική αίσθηση, ενώ η μία είναι διπλάσια της άλλης. Το μουσικό σύστημα το οποίο χρησιμοποιείται είναι το ισοσυγκερασμένο σύστημα των 12 τόνων, που διαχωρίζει την κάθε οκτάβα σε 12 λογαριθμικά ίσα μέρη, τα ημιτόνια. Ο λόγος μεταξύ δύο διαδοχικών ημιτονίων είναι $2^{(1/12)}$ και κατά συνέπεια οι 12 νότες οι οποίες περιλαμβάνονται σε μία οκτάβα είναι λογαριθμικά κατανεμημένες στον άξονα των συχνοτήτων.

Ο υπολογισμός της θεμελιώδους συχνότητας $f$ μιας νότας $n$ δίνεται από την εξίσωση:

$$f(n) = (\sqrt[12]{2})^{n-49} * 440\text{Hz} = 2^{\frac{n-49}{12}} * 440\text{Hz} \tag{4.7}$$

όπου σαν αναφορά για τον υπολογισμό χρησιμοποιούμε την νότα A4 (49η νότα στο πιάνο)




*Σχήμα 4.4:* Δύο διαφορετικές συστοιχίες φίλτρων Gabor, η πρώτη με 12 mel-spaced φίλτρα και επικάλυψη 50% (αριστερά) και η δεύτερη με 89 Gabor φίλτρα κεντραρισμένα στις θεμελιώδεις συχνότητες των μουσικών τόνων ξεκινώντας από την δεύτερη οκτάβα με το εύρος του κάθε φίλτρου να εκτείνεται από την κεντρική συχνότητα του προηγούμενου φίλτρου ως την κεντρική συχνότητα του επόμενου (δεξιά).

με συχνότητα 440Hz[1]. Εναλλακτικά, για την εύρεση του αριθμού της νότας η εξίσωση γράφεται ως εξής:

$$n = 12 \log_2 \left( \frac{f}{440\text{Hz}} \right) + 49. \tag{4.8}$$

Στο Σχήμα 4.4 παρουσιάζονται οι δύο διαφορετικές συστοιχίες φίλτρων Gabor, η πρώτη με 12 mel-spaced φίλτρα και επικάλυψη 50% και η δεύτερη με 89 Gabor φίλτρα κεντραρισμένα στις θεμελιώδεις συχνότητες των μουσικών τόνων, ξεκινώντας από τη δεύτερη οκτάβα με το εύρος του κάθε φίλτρου να εκτείνεται από την κεντρική συχνότητα του προηγούμενου φίλτρου ως την κεντρική συχνότητα του επόμενου.

Τα τελικά σύνολα χαρακτηριστικών αλλά και οι συνδυασμοί τους αξιολογήθηκαν με τη χρήση των κρυφών Μαρκοβιανών μοντέλων (HMMs), όπου πειραματιστήκαμε με διαφορετικό αριθμό μειγμάτων $M$ και ενίοτε καταστάσεων $N$. Η απόδοση των διαφόρων χαρακτηριστικών συγκρίθηκε με τα MFCC. Σε όλα τα πειράματα που ακολουθούν τα δεδομένα εκπαίδευσης έχουν επιλεγεί τυχαία να είναι το 90% των διαθέσιμων ηχητικών σημάτων, και τα αποτελέσματα που παρουσιάζονται έχουν προκύψει με 5 cross validation. Έχουμε ήδη αναφέρει πως τα ηχητικά αρχεία της βάσης (διάρκειας 30 δευτερολέπτων) χωρίζονται σε τρία τμήματα των 10 δευτερολέπτων. Σημειώνουμε πως

[1]Αντί του 49 χρησιμοποιούμε το 69 για τον υπολογισμό MIDI συχνοτήτων (το MIDI (Musical Instrument Digital Interface) [188] είναι ένα πρωτόκολλο το οποίο επιτρέπει την επικοινωνία, τον έλεγχο και τη σύνδεση μεταξύ διαφόρων ηλεκτρονικών συσκευών όπως υπολογιστές, ηλεκτρονικά μουσικά όργανα κ.ά.), το 57 αν θέλουμε οι συχνότητες να ξεκινάνε από τη συχνότητα 16.35Hz και 49 για να ξεκινάνε από τη συχνότητα 27.5Hz.

στην τυχαία επιλογή των δεδομένων εκπαίδευσης και δοκιμής φροντίζουμε οι τριπλέτες των αρχείων να συμπεριλαμβάνονται είτε στα δεδομένα εκπαίδευσης είτε στα δεδομένα δοκιμής, μιας και η εντελώς τυχαία επιλογή και μίξη τους θα είχε ως παρεπόμενο την πολύ καλύτερη αναγνώριση. Επιπλέον διεξαγάγαμε multi-stream μοντελοποίηση όπου τα διαφορετικά σύνολα χαρακτηριστικών συνδυάζονται, μεταβάλλοντας το βάρος της κάθε ηχητικής ροής.

Στην προσπάθειά μας να εξακριβώσουμε περαιτέρω την ευρωστία των εξαχθέντων χαρακτηριστικών, αναπτύσσουμε και ένα δεύτερο πλαίσιο για την αξιολόγησή τους, το οποίο βασίζεται στη μέθοδο Bag-of-Words, όπου τα χαρακτηριστικά αξιολογούνται με τη χρήση των SVMs. Περισσότερες λεπτομέρειες για τη μεθοδολογία που ακολουθήθηκε και τα πειράματα αξιολόγησης μπορούν να βρεθούν στην Ενότητα 4.4.

Στην συνέχεια του κεφαλαίου αυτού παρουσιάζουμε τα χαρακτηριστικά τα οποία εξάγαμε βάσει της κάθε διαφορετικής παραμετροποίησης καθώς και την απόδοσή τους στα πειράματα αναγνώρισης.

### 4.3.2 Πειραματική Αξιολόγηση: Σύνολα Χαρακτηριστικών

Στα πειράματα που ακολουθούν αξιολογούμε την απόδοση των διαφορετικών συνόλων χαρακτηριστικών. Το βασικό σύνολο αποτελείται από χαρακτηριστικά τα οποία έχουν εξαχθεί βάσει της θεωρίας διαμορφώσεων αλλά σε πολλές περιπτώσεις έχουν συνενωθεί με τα MFD ή τη φράκταλ διάσταση MFD$[s = 1]$, ή/και με τα MFCC τα οποία κυρίως χρησιμοποιούνται για λόγους σύγκρισης. Τα κύρια χαρακτηριστικά αυτής της ανάλυσης και οι μεθοδολογίες με τις οποίες προσεγγίστηκαν παρουσιάζονται στη συνέχεια:

1) Αναγνώριση των μικροδομών των μουσικών σημάτων χρησιμοποιώντας βραχέος χρόνου ανάλυση με τη baseline ζωνοπερατή mel-spaced Gabor συστοιχία 12 φίλτρων με επικάλυψη των διαδοχικών φίλτρων 50%, βλέπε Πίνακα 4.4.

2a) Αναγνώριση των μικροδομών των μουσικών σημάτων χρησιμοποιώντας βραχέος χρόνου ανάλυση με τη ζωνοπερατή «μουσική» συστοιχία 89 Gabor φίλτρων με την εφαρμογή ανάλυσης PCA για τη μείωση της διάστασης των συνόλων χαρακτηριστικών, βλ. Πίνακα 4.5.

2b) Αναγνώριση των μικροδομών των μουσικών σημάτων χρησιμοποιώντας βραχέος χρόνου ανάλυση με τη ζωνοπερατή «μουσική» συστοιχία 101 Gabor φίλτρων με την εφαρμογή ανάλυσης PCA για τη μείωση της διάστασης των συνόλων χαρακτηριστικών, βλ. Πίνακα 4.5.

3) Αναγνώριση των μακροδομών των μουσικών σημάτων χρησιμοποιώντας μεγαλύτερης διάρκειας παράθυρα ανάλυσης για τον υπολογισμό των μέσων τιμών των

*Πίνακας 4.4:* Λίστα χαρακτηριστικών χρησιμοποιώντας βραχέος χρόνου ανάλυση με την baseline ζωνοπερατή mel-spaced Gabor συστοιχία 12 φίλτρων με επικάλυψη των διαδοχικών φίλτρων 50% για την αναγνώριση ειδών μουσικής.

| **Σύνολα χαρακτηριστικών** | | **Περιγραφή** |
|---|---|---|
| 1 | $\text{LMF}_{\Delta}$ | 72 AMFM χαρακτηριστικά, δηλ. 12 $\log$(m-IAM) + 12 m-IFM (+ 12 $\Delta$ + 12 $\Delta\Delta$). |
| 2 | $\text{LMFPf}_{39}$ | 39 AMFM+FMP χαρακτηριστικά μετά από PCA ανάλυση στο συνολικό συνενωμένο διάνυσμα 108 χαρακτηριστικών. * |
| 3 | $\text{LMFi}_{39}$ | 39 AMFM χαρακτηριστικά μετά από PCA ανάλυση. |
| 4 | $\text{LMFiPf}_{52}$ | 39 AMFM + 13 FMP χαρακτηριστικά μετά από PCA ανάλυση. |
| 5 | $\text{LMFiD}_{40}$ | 39 AMFM χαρακτηριστικά μετά από PCA ανάλυση + MFD$[s = 1]$. |
| 6 | $\text{MFC}_{\Delta}$ | 13 MFCC + 13 $\Delta$ + 13 $\Delta\Delta$ |

* Όπου i στην ονοματολογία των χαρακτηριστικών υποδηλώνει πως η PCA ανάλυση έγινε στα ξεχωριστά διανύσματα χαρακτηριστικών και όπου f στο συνολικό συνενωμένο σύνολο χαρακτηριστικών.

**Multi-Stream Cases**

| | | |
|---|---|---|
| 1 | $\text{LMFi}_{39}$ + $\text{MFC}_{\Delta}$ | Αξιολόγηση των διαφόρων συνόλων χαρακτηριστικών με multi-stream μοντελοποίηση μαζί με τα $\text{MFC}_{\Delta}$. |
| 2 | $\text{LMFiP}_{52}$ + $\text{MFC}_{\Delta}$ | |
| 3 | $\text{LMFMiD}_{40}$ + $\text{MFC}_{\Delta}$ | |

χαρακτηριστικών και την baseline ζωνοπερατή mel-spaced Gabor συστοιχία 12 φίλτρων με επικάλυψη των διαδοχικών φίλτρων 50%, βλ. Πίνακα 4.6.

4) Αναγνώριση των δομών των μουσικών σημάτων με τη συνένωση πολλών διαδοχικών παραθύρων βραχέος χρόνου, χωρίς τις παραγώγους τους, και τη μείωση της διάστασης τους με την εφαρμογή ανάλυσης PCA, βλ. Πίνακα 4.7.

### 4.3.3 Πειραματική Αξιολόγηση: Αποτελέσματα

Βάσει των αποτελεσμάτων αναγνώρισης των διαφορετικών χαρακτηριστικών, αποδεικνύουμε πως τα AM-FM χαρακτηριστικά παρουσιάζουν σχεδόν την ίδια απόδοση με τα $\text{MFC}_{\Delta}$, ενώ ο συνδυασμός τους επιτυγχάνει αξιόλογη αναγνώριση των 10 διαφορετικών

*Πίνακας 4.5:* Λίστα AM-FM χαρακτηριστικών χρησιμοποιώντας βραχέος χρόνου ανάλυση με τη ζωνοπερατή «μουσική» συστοιχία 89 ή 101 φίλτρων Gabor, για την αναγνώριση ειδών μουσικής.

| **Σύνολα χαρακτηριστικών** | | **Περιγραφή** |
|---|---|---|
| 1 | LMF89$b1_{198}$ | Τα σύνολα 1-5 έχουν προκύψει χρησιμοποιώντας 89 φίλτρα και τα 6-9 χρησιμοποιώντας 101 φίλτρα. Το $b$ υποδηλώνει το εύρος του κάθε φίλτρου, όπου $b1$ = $[f_{n-1}, f_{n+1}]$ για κεντρική συχνότητα $f_n$, ενώ το $b2$ = $b1$x2 και αντίστοιχα για τα υπόλοιπα. Ο τελικός αριθμός χαρακτηριστικών έχει προκύψει μετά από PCA ανάλυση των 89 ή 101 log(m-IAM) + 89 ή 101 m-IFM και των παραγώγων τους (89 ή 101Δ + 89 ή 101ΔΔ). |
| 2 | LMF89$b1_{240}$ | |
| 3 | LMF89$b2_{122}$ | |
| 4 | LMF89$b4_{52}$ | |
| 5 | LMF89$b6_{198}$ | |
| 6 | LMF101$b1_{105}$ | |
| 7 | LMF101$b1_{265}$ | |
| 8 | LMF101$b2_{133}$ | |
| 9 | LMF101$b4_{58}$ | |
| **Multi-Stream Cases** | | |
| 1 | LMF89$b1_{198}$MFC$_{39}$ | Για την multi-stream μοντελοποίηση των χαρακτηριστικών χρησιμοποιούμε 13 συντελεστές MFCC με τις παραγώγους τους. |
| 2 | LMF89$b1_{240}$MFC$_{39}$ | |
| 3 | LMF101$b1_{105}$MFC$_{39}$ | |
| 4 | LMF101$b1_{265}$MFC$_{39}$ | |

*Πίνακας 4.6:* Λίστα AM-FM χαρακτηριστικών τα οποία έχουν προκύψει χρησιμοποιώντας παράθυρα ανάλυσης μεγαλύτερης διάρκειας, για την αναγνώριση ειδών μουσικής.

| **Σύνολα χαρακτηριστικών** | | **Περιγραφή** |
|---|---|---|
| 1 | LMF$_{\Delta(125)}$ | Τα σύνολα 1-2 αποτελούνται από 12 log(m-IAM) + 12 m-IFM + τις παραγώγους τους (72 χαρακτηριστικά σύνολο). Το νούμερο στην παρένθεση υποδηλώνει το μέγεθος του παραθύρου ανάλυσης το οποίο είναι 125 ή 250 ms με επικάλυψη 80%, δηλαδή 100 και 200 ms αντίστοιχα. |
| 2 | LMF$_{\Delta(250)}$ | |
| 3 | LMFP$_{108(125)}$ | Τα σύνολα 3-4 αποτελούνται από 12 log(m-IAM) + 12 m-IFM + 12 FMP + τις παραγώγους τους (108 χαρακτηριστικά σύνολο). Το νούμερο στην παρένθεση υποδηλώνει το μέγεθος του παραθύρου ανάλυσης το οποίο είναι 125 ή 250 ms. |
| 4 | LMFP$_{108(250)}$ | |

*Πίνακας 4.7:* Λίστα χαρακτηριστικών τα οποία έχουν προκύψει μετά από τη συνένωση διαδοχικών πλαισίων ανάλυσης χωρίς παραγώγους, για την αναγνώριση ειδών μουσικής.

| **Σύνολα χαρακτηριστικών** | | **Περιγραφή** |
|---|---|---|
| 1 | $\text{LMFPC}_{392(8F)}$ | Τα σύνολα προκύπτουν από συνένωση των 12 log(m-IAM) + 12 m-IFM + 12 FMP + 13 MFCC, τα οποία και υποδηλώνονται με τα αρχικά LM,F,P,C αντίστοιχα. |
| 2 | $\text{LMFPC}_{368(15F)}$ | |
| 3 | $\text{LMFPC}_{735(15F)}$ | |
| 4 | $\text{LMFPCD}_{172(8F)}$ | Στα σύνολα 4-11 εκτός των χαρακτηριστικών που αναφέρονται για τα 1-3 προσθέτουμε και τα MFD χαρακτηριστικά (υποδηλώνονται ως D). Στην παρένθεση βλέπουμε των αριθμό των διαδοχικών πλαισίων που έχουν συνενωθεί για τη δημιουργία των συνόλων, π.χ το $8F$ υποδηλώνει τη συνένωση 8 διαδοχικών πλαισίων ενώ το $8FH$ τη συνένωση 8 πλαισίων, χρησιμοποιώντας επικάλυψη 50%. Ο τελικός αριθμός χαρακτηριστικών έχει προκύψει μετά από PCA ανάλυση, και σε πολλές περιπτώσεις έχει επιλεχθεί για λόγους σύγκρισης ενώ σε άλλες αποτελεί ένα συγκεκριμένο ποσοστό (%) του αρχικού αριθμού χαρακτηριστικών. Παράδειγμα για το σύνολο LMFPCD το οποίο αρχικά αποτελείται από 107 χαρακτηριστικά και σύνολο 856 (107×8 πλαίσια), παρατηρούμε πως τα 172 χαρακτηριστικά είναι το 20%, τα 214 το 25% και τα 428 το 50% του συνολικού αριθμού χαρακτηριστικών. Κατ' αντιστοιχία στο σύνολο LMFPC τα 368 χαρακτηριστικά αποτελούν το 50% και τα 735 το 100% επί του συνόλου των χαρακτηριστικών. |
| 5 | $\text{LMFPCD}_{172(8FH)}$ | |
| 6 | $\text{LMFPCD}_{214(8F)}$ | |
| 7 | $\text{LMFPCD}_{214(8FH)}$ | |
| 8 | $\text{LMFPCD}_{428(8F)}$ | |
| 9 | $\text{LMFPCD}_{428(8FH)}$ | |
| 10 | $\text{LMFPCD}_{735(8F)}$ | |
| 11 | $\text{LMFPCD}_{735(8FH)}$ | |

ειδών της βάσης που χρησιμοποιήθηκε.

Σε κάθε πίνακα αποτελεσμάτων που ακολουθεί το μέτρο αξιολόγησης που παρουσιάζουμε είναι το ποσοστό μέσης ακρίβειας αναγνώρισης (accuracy) (%) των μουσικών ειδών. Για τους λόγους που αναφέρθηκαν στην Ενότητα 2.5 (αναφορικά με την εγκυρότητα της GTZAN βάσης) στη συζήτηση που ακολουθεί σε πολλές περιπτώσεις αναφέρουμε το μέγιστο ποσοστό αναγνώρισης κάποιου fold για τα χαρακτηριστικά που απέδωσαν καλύτερα. Τα διαφορετικά σύνολα χαρακτηριστικών αξιολογήθηκαν με τη

*Πίνακας 4.8:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) για 10 μουσικά είδη με HMM και χαρακτηριστικά βραχέος χρόνου, όπου $N$ ο αριθμός των καταστάσεων. Στην παρένθεση φαίνεται ο αριθμός των μειγμάτων $M$ για τον οποίο επιτεύχθηκε το καλύτερο αποτέλεσμα αναγνώρισης. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 4.4.

| **Ποσοστά επιτυχίας κατηγοριοποίησης** | | | | |
|---|---|---|---|---|
| Χαρακτηριστικά | | HMM | | |
| | # Καταστάσεων $N$: | $N=5$ | $N=7$ | $N=9$ |
| $\text{LMF}_{\Delta}$ | - | 76.46 (16) | 76.46 (16) | 78.07 (16) |
| $\text{LMFPf}_{39}$ | - | 76.99 (16) | - | - |
| $\text{LMFi}_{39}$ | - | 77.12 (16) | - | - |
| $\text{LMFiPf}_{52}$ | - | 77.86 (16) | **79.87 (15)** | - |
| $\text{LMFiD}_{40}$ | - | 76.12 (15) | 76.86 (15) | - |
| $\text{MFC}_{\Delta}$ | - | **78.06** (14) | 78.35 (14) | 77.59 (15) |
| **Multi-Stream Cases** | | | | |
| Χαρακτηριστικά | Βάρη | HMM | | |
| $\text{LMFi}_{39}\text{MFC}_{\Delta}$ | 0.3 - 0.7 | 81.40 (16) | 81.47 (12) | - |
| | 0.5 - 0.5 | **82.81** (16) | 82.27 (14) | - |
| $\text{LMFiPf}_{52}\text{MFC}_{\Delta}$ | 0.3 - 0.7 | 81.54 (16) | 81.87 (12) | - |
| | 0.5 - 0.5 | 82.08 (12) | **82.61** (14) | - |
| $\text{LMFiD}_{40}\text{MFC}_{\Delta}$ | 0.3 - 0.7 | 80.80 (16) | 80.67 (13) | 81.60 (16) |
| | 0.5 - 0.5 | 82.34 (16) | 82.21 (16) | 81.34 (10) |
| | 0.7 - 0.3 | 80.81 (16) | 80.80 (12) | - |

χρήση HMM ταξινομητών, με $N = 5$ και/ή $7, 9$ καταστάσεις, μεταβάλλοντας τον αριθμό των Γκαουσιανών μειγμάτων $M = [1 - 16]$. Επιπλέον, κατά την παρουσίαση των αποτελεσμάτων στους πίνακες που ακολουθούν σε παρένθεση φαίνεται ο αριθμός των μειγμάτων $M$ για τον οποίο επιτεύχθηκε το καλύτερο αποτέλεσμα αναγνώρισης.

**Πειραματικά αποτελέσματα για την πρώτη κατηγορία χαρακτηριστικών.**

Συγκεκριμένα, στα πειραματικά αποτελέσματα για την πρώτη κατηγορία χαρακτηριστικών (βλ. Πίνακα 4.8 για ποσοστά επιτυχίας κατηγοριοποίησης), δηλ. αυτά τα οποία έχουν εξαχθεί με ανάλυση βραχέος χρόνου και την baseline παραμετροποίηση (για χαρακτηριστικά βλ. Πίνακα 4.4), ο συνδυασμός των προτεινόμενων χαρακτηριστικών με τα MFCC αποδείχθηκε καλύτερος από ό,τι τα MFCC σε όλες τις περιπτώσεις. Παρ' όλα αυτά τα AM-FM μόνα τους παρουσιάζουν χαμηλότερο διαχωρισμό σε ένα ποσοστό περίπου 0.5-2% για $N = 5$ καταστάσεις, κάτι που αλλάζει όταν οι καταστάσεις του HMM μοντέλου αυξάνονται για $N = 7$ ή $N = 9$. Για παράδειγμα, παρατηρούμε πως το σύνολο $\text{LMFiPf}_{52}$ παρουσιάζει μείωση σφάλματος 7% για $N = 7$. Η καλύτερη περίπτωση συνδυασμού των $\text{LMFi}_{39}\text{MFC}_{\Delta}$ στα multi-stream πειράματα πέτυχε μείωση σφάλματος περίπου 20%, για HMMs με $N = 5$ καταστάσεις και ίσα βάρη στις δύο ροές $s_1 = s_2 = 0.5$. Επίσης διακρίνουμε πως και οι υπόλοιποι συνδυασμοί των χαρακτηριστικών απέδωσαν καλά με μείωση σφάλματος έως περίπου 18%.

*Πίνακας 4.9:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) για 10 μουσικά είδη με HMM και χαρακτηριστικά βραχέος χρόνου με τη μουσική συστοιχία φίλτρων Gabor. Όπου $N$ ο αριθμός των καταστάσεων. Στη παρένθεση φαίνεται ο αριθμό των μειγμάτων $M$ για τον οποίο επιτεύχθηκε το καλύτερο αποτέλεσμα αναγνώρισης. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 4.5.

| **Ποσοστά επιτυχίας κατηγοριοποίησης** | | |
|---|---|---|
| Χαρακτηριστικά | # Καταστάσεων: | $N = 5$ |
| LMF89$b1_{198}$ | - | 81.68 (14) |
| LMF89$b1_{240}$ | - | 81.81 (14) |
| LMF89$b2_{122}$ | - | 81.14 (16) |
| LMF89$b4_{52}$ | - | 74.78 (16) |
| LMF89$b6_{198}$ | - | 74.58 ( 7) |
| LMF101$b1_{105}$ | - | 80.87 (16) |
| LMF101$b1_{265}$ | - | **83.22** (14) |
| LMF101$b2_{133}$ | - | 81.47 (16) |
| LMF101$b4_{58}$ | - | 76.72 (15) |
| **Multi-Stream Cases** | | |
| Χαρακτηριστικά | Βάρη | HMM |
| LMF89b$1_{198}$MFC$_\Delta$ (237) | 0.3 - 0.7 | **84.22** (16) |
| | 0.5 - 0.5 | 83.50 (12) |
| | 0.7 - 0.3 | 83.13 (16) |
| LMF89b$1_{240}$MFC$_\Delta$ (279) | 0.3 - 0.7 | 83.48 (15) |
| | 0.5 - 0.5 | **84.15** (11) |
| | 0.7 - 0.3 | 82.74 (15) |
| LMF101b$1_{105}$MFC$_\Delta$ (144) | 0.3 - 0.7 | 82.34 (15) |
| | 0.5 - 0.5 | 81.21 (16) |
| | 0.7 - 0.3 | 82.81 (14) |
| LMF101b$1_{265}$MFC$_\Delta$ (304) | 0.3 - 0.7 | **84.41** (14) |
| | 0.5 - 0.5 | 83.68 (14) |
| | 0.7 - 0.3 | 83.28 (14) |

Αναφορικά με τα AM-FM χαρακτηριστικά διαπιστώνουμε πως η προσθήκη των FMP μετά από την ανάλυση σε κύριες συνιστώσες, όπως στο σύνολο LMFiPf$_{52}$, μειώνει το σφάλμα σε σχέση με τα LMF$_\Delta$ περίπου 5% για $N = 5$ και 14% για $N = 7$ καταστάσεις. Από την άλλη, η προσθήκη της φράκταλ διάστασης των χαρακτηριστικών δεν επιτυγχάνει την καλυτέρευση των αποτελεσμάτων.

**Πειραματικά αποτελέσματα για τη δεύτερη κατηγορία χαρακτηριστικών.**

Στον Πίνακα 4.9 παρουσιάζονται τα ποσοστά επιτυχίας κατηγοριοποίησης για τη δεύτερη κατηγορία χαρακτηριστικών, και συγκεκριμένα τα χαρακτηριστικά βραχέος χρόνου, για την εξαγωγή των οποίων χρησιμοποιήσαμε τη μουσική συστοιχία φίλτρων. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 4.5. Τα ποσοστά επιτυχίας για αυτή την κατηγορία αναπαραστάσεων των μουσικών σημάτων αυξάνονται ενώ η μείωση σφάλματος ανέρχεται περίπου στο 24% για το σύνολο LMF101$b1_{265}$ σε σύγκριση με το

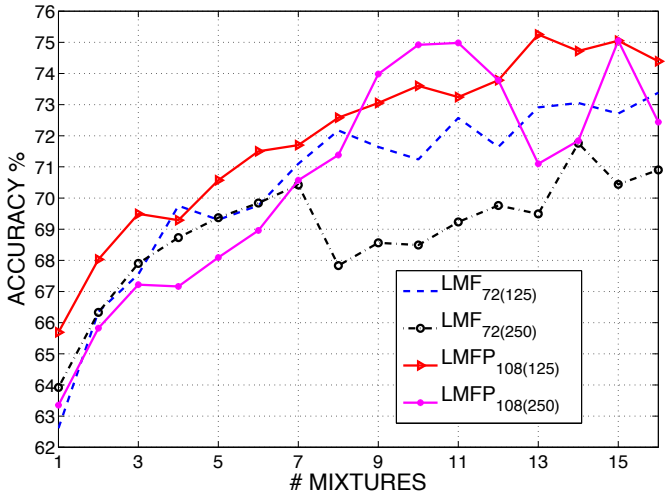
*Πίνακας 4.10:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) για 10 μουσικά είδη με HMM και χαρακτηριστικά που έχουν προκύψει με τη συνένωση των βραχέος χρόνου χαρακτηριστικών (χωρίς παραγώγους), για $N = 5, 7,$ και $9$ καταστάσεις. Στην παρένθεση βλέπουμε τον αριθμό των μειγμάτων για τον οποίο επιτεύχθηκε το καλύτερο ποσοστό αναγνώρισης. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 4.7.

| **Ποσοστά επιτυχίας κατηγοριοποίησης** | | | |
|---|---|---|---|
| | HMM | | |
| Χαρακτηριστικά | $N = 5$ | $N = 7$ | $N = 9$ |
| $\text{LMFPC}_{392(8F)}$ | 82.61 (15) | 81.20 (8) | 81.20 (12) |
| $\text{LMFPC}_{368(15F)}$ | 78.92 (5) | 78.52 (10) | 79.21 (5) |
| $\text{LMFPC}_{735(15F)}$ | 80.40 (13) | 78.46 (9) | 79.33 (4) |
| $\text{LMFPCD}_{172(8F)}$ | 80.60 (14) | 81.00 (16) | 80.07 (10) |
| $\text{LMFPCD}_{172(8FH)}$ | 79.67 (7) | 80.66 (11) | 80.20 (16) |
| $\text{LMFPCD}_{214(8F)}$ | 82.28 (14) | 80.88 (15) | 81.94 (11) |
| $\text{LMFPCD}_{214(8FH)}$ | **82.88** (10) | 82.34 (15) | 82.01 (11) |
| $\text{LMFPCD}_{428(8F)}$ | 82.48 (16) | 81.20 (15) | 80.74 (10) |
| $\text{LMFPCD}_{428(8FH)}$ | 81.68 (16) | **82.35** (14) | **82.61** (13) |
| $\text{LMFPCD}_{735(8F)}$ | 79.87 (5) | 78.59 (7) | 78.39 (8) |
| $\text{LMFPCD}_{735(8FH)}$ | 79.07 (14) | 80.20 (10) | 79.20 (8) |

σύνολο $\text{LMFiPf}_{52}$ και το $\text{MFC}_{\Delta}$, ενώ ο συνδυασμός των χαρακτηριστικών αυτών με τα $\text{MFC}_{\Delta}$ επιφέρει μείωση σφάλματος 28% σε σύγκριση με τα $\text{MFC}_{\Delta}$ και περίπου 8% σε σύγκριση με τον καλύτερο συνδυασμό των χαρακτηριστικών της πρώτης κατηγορίας, δηλ. το σύνολο $\text{LMFi}_{39}\text{MFC}_{\Delta}$. Αξίζει να αναφέρουμε πως το καλύτερο μεμονωμένο αποτέλεσμα κάποιου fold σε αυτή την κατηγορία φτάνει το 87.21% για το σύνολο $\text{LMF101}b1_{265}$ είτε αυτό αξιολογείται μόνο του, είτε σε συνδυασμό με τα $\text{MFC}_{\Delta}$. Παρατηρούμε επίσης, πως στις περισσότερες περιπτώσεις τα καλύτερα αποτελέσματα επιτεύχθηκαν όταν το βάρος των AM-FM χαρακτηριστικών είναι μεγαλύτερο ή ίσο με αυτό των MFCC, κάτι που ενισχύει το γεγονός πως οι διαμορφώσεις συμβάλλουν σημαντικά στην αναγνώριση των ειδών. Τέλος, υπογραμμίζουμε πως μεγαλύτερο ποσοστό αναγνώρισης επιτεύχθηκε όταν το εύρος της συστοιχίας φίλτρων ισούται με ένα.

**Πειραματικά αποτελέσματα για την τέταρτη κατηγορία χαρακτηριστικών.**

Στον Πίνακα 4.10 φαίνονται τα ποσοστά επιτυχίας για την τέταρτη κατηγορία χαρακτηριστικών και συγκεκριμένα για τη συνένωση των βραχέος χρόνου χαρακτηριστικών, για HMM και για $N = 5, 7,$ και $9$ καταστάσεις όπου ο αριθμός των μειγμάτων μεταβάλλεται από $M = [1 - 16]$. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 4.7. Το σύνολο χαρακτηριστικών $\text{LMFPCD}_{214(8FH)}$ έδειξε τη μεγαλύτερη ικανότητα αναγνώρισης 82.88% για $N = 5$ καταστάσεις και $M = 10$ μείγματα. Παρ' όλα αυτά, τα περισσότερα σύνολα χαρακτηριστικών τα




*Σχήμα 4.5:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) 10 μουσικών ειδών, με AM-FM χαρακτηριστικά τα οποία έχουν εξαχθεί χρησιμοποιώντας παράθυρα ανάλυσης διάρκειας 125 ή 250 ms αντίστοιχα. Τα διαφορετικά σύνολα χαρακτηριστικών αξιολογήθηκαν με την χρήση HMM ταξινομητών, με $N = 5$ καταστάσεις και μεταβάλλοντας των αριθμό των Γκαουσιανών μειγμάτων $M = [1 - 16]$. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 4.6.

οποία δημιουργήθηκαν με τη συγκεκριμένη μεθοδολογία παρουσιάζονται ιδιαίτερα ανταγωνιστικά όχι μόνο για τα αξιόλογα ποσοστά αναγνώρισης αλλά για τον απλό λόγο ότι παρουσιάζουν τους μικρότερους χρόνους εκμάθησης του συστήματος. Επιπλέον, επιτυγχάνουν ένα αξιόλογο ποσοστό επιτυχίας με τη χρήση λίγων Γκαουσιανών μειγμάτων, παρόλο που στα αποτελέσματα που δείχνουμε δεν γίνεται φανερό, για λόγους παρουσίασης του απόλυτου μέγιστου του κάθε σετ. Σημειώνουμε πως παρ' όλο που ο αριθμός των χαρακτηριστικών που χρειάζεται να κρατήσουμε μετά από την εφαρμογή της ανάλυσης PCA είναι αρκετά μεγάλος, λόγω της μείωσης της διάστασης του χρόνου (δηλ. του αριθμού των τελικών παραθύρων) οι αναπαραστάσεις που τελικά δημιουργούνται είναι αρκετά συμπαγείς και όπως βλέπουμε αρκετά αξιόπιστες.

Επιπρόσθετα, σε αντίθεση με τις παρατηρήσεις μας σχετικά με τα baseline χαρακτηριστικά και την χρησιμότητα των MFD και της φράκταλ διάστασης, παρατηρούμε πως τα MFD χαρακτηριστικά συντελούν στην καλύτερη απόδοση του συστήματος αναγνώρισης. Αξίζει να σχολιάσουμε πως το καλύτερο μέγιστο αποτέλεσμα κάποιου fold σε αυτή την κατηγορία χαρακτηριστικών είναι 85.95% για το σύνολο $\text{LMFPCD}_{214(8FH)}$. Αν και στην παρουσίαση των χαρακτηριστικών αναφέραμε τη δημιουργία συνόλων συνενώνοντας διαδοχικά παράθυρα με συνολική διάρκεια 1/8, 1/4, 1/2 και 1 sec, τα αποτελέσματα της αξιολόγησης παρουσίασαν μέγιστη αναγνώριση στις περιπτώσεις συνένωσης 8 ή 15 παραθύρων ανάλυσης.

**Πειραματικά αποτελέσματα για την τρίτη κατηγορία χαρακτηριστικών.**

Στο Σχήμα 4.5 παρουσιάζουμε τα ποσοστά επιτυχία κατηγοριοποίησης για την

*Πίνακας 4.11:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) ανά μουσικό είδος (10 μουσικών ειδών) για τους τέσσερις καλύτερους συνδυασμούς χαρακτηριστικών σε σύγκριση με τα MFCC. Όλα τα αποτελέσματα των multi-stream πειραμάτων αφορούν $N = 5$ καταστάσεις εκτός των $\text{MFC}_\Delta$ για $N = 7$ και 5-fold cross-validation.

| **Ποσοστά επιτυχίας κατηγοριοποίησης** | | | | | |
|---|---|---|---|---|---|
| Είδος Μουσικής | $\text{LMF101b1}_{265}$ $M = 14$ | $\text{LMF101b1}_{265}$ $\text{MFC}_\Delta$ $M = 14$ $s_1 = 0.3, s_2 = 0.7$ | LMFPCD $214(8FH)$ $M = 16$ $s_{1,2} = 0.5$ | $\text{LMFi}_{39}$ $\text{MFC}_\Delta$ $M = 10$ | MFCC $M = 14$ |
| Blues | 89.3 | 90.0 | 87.3 | **92.0** | 88.7 |
| Classical | 95.3 | **96.0** | 93.4 | 95.3 | 93.3 |
| Country | **85.8** | 84.5 | 85.1 | 81.8 | 83.8 |
| Disco | 75.8 | 74.5 | **78.0** | 71.8 | 69.7 |
| Hiphop | 74.0 | **83.3** | 78.7 | 80.0 | 77.3 |
| Jazz | **94.0** | 90.7 | 90.7 | 89.3 | 76.7 |
| Metal | 91.3 | **92.7** | 88.7 | 89.2 | 89.3 |
| Pop | 85.3 | **94.7** | 90.0 | 92.7 | 84.0 |
| Reggae | 73.3 | 74.7 | **75.3** | 72.7 | 69.3 |
| Rock | **67.8** | 63.1 | 61.8 | 61.1 | 51.0 |

τρίτη κατηγορία χαρακτηριστικών, τα οποία έχουν εξαχθεί χρησιμοποιώντας παράθυρα ανάλυσης διάρκειας 125 και 250 ms αντίστοιχα. Τα διαφορετικά σύνολα χαρακτηριστικών αξιολογήθηκαν με τη χρήση HMM ταξινομητών, με $N = 5$ καταστάσεις, μεταβάλλοντας των αριθμό των Γκαουσιανών μειγμάτων $M = [1 - 16]$. Για πληροφορίες σχετικές με τα χαρακτηριστικά βλ. Πίνακα 4.6. Γενικά παρατηρούμε πως η ανάλυση αυτή δεν επιτυγχάνει εξίσου καλά αποτελέσματα. Το καλύτερο ποσοστό αναγνώρισης 75.25% επιτεύχθηκε από το σύνολο $\text{LMFP}_{108(125)}$ με μέγεθος παραθύρου ανάλυσης 125 ms για $M = 13$ μείγματα. Τα χειρότερα αποτελέσματα παρουσιάζονται για τα σύνολα χαρακτηριστικών στα οποία χρησιμοποιήθηκε παράθυρο ανάλυσης των 250 ms με εξαίρεση το σύνολο $\text{LMFP}_{108(250)}$ με αριθμό μειγμάτων $M = [9 - 11, 15]$ όπου και επιτυγχάνει μέγιστη αναγνώριση 75.05%. Παρ' όλα αυτά επισημαίνουμε πως και εδώ τα FMP χαρακτηριστικά λειτουργούν βοηθητικά στην κατηγοριοποίηση ενώ αποδίδουν μείωση σφάλματος ως 7% και 12% για τα σύνολα $\text{LMFP}_{108(125)}$ και $\text{LMFP}_{108(250)}$ αντίστοιχα.

Καταλήγοντας, στον Πίνακα 4.11 παρουσιάζουμε τα ποσοστά επιτυχίας κατηγοριοποίησης ανά μουσικό είδος, για τους τέσσερις καλύτερους συνδυασμούς χαρακτηριστικών σε σύγκριση με τα MFCC. Όλα τα αποτελέσματα των multi-stream πειραμάτων αφορούν $N = 5$ καταστάσεις εκτός των $\text{MFC}_\Delta$ τα οποία και παρουσιάζονται για $N = 7$ (και 5 cross-validation). Παρατηρούμε πως η κλασική μουσική έχει τη μεγαλύτερη επιτυχία αναγνώρισης, ενώ ακολουθούν η pop, η metal και η country. Τα χειρότερα αποτελέσματα αναγνώρισης παρουσιάζονται για τη rock σε όλα τα σύνολα

χαρακτηριστικών που αξιολογήθηκαν. Υπογραμμίζουμε πάντως πως ο συνδυασμός των AM-FM χαρακτηριστικών αποδίδει καλύτερα για όλες τις διαφορετικές κατηγορίες μουσικών ειδών σε σύγκριση με τα MFCC.

**Συμπεράσματα**

Παρουσιάσαμε και προτείναμε τέσσερις διαφορετικές μεθοδολογίες για τη δημιουργία αναπαραστάσεων των μουσικών σημάτων με βάση το μοντέλο των διαμορφώσεων. Η αξιολόγηση των διαφορετικών συνόλων χαρακτηριστικών καταδεικνύει τη δυναμική και τη σημασία των προτεινόμενων μεθόδων στο θέμα της κατηγοριοποίησης διαφορετικών ειδών μουσικής. Τα αποτελέσματα σχεδόν όλων των προτεινόμενων συνόλων παρουσιάζονται ιδιαίτερα ενθαρρυντικά και διαπιστώνουμε πως οι διαμορφώσεις δύνανται να περιγράψουν σημαντικά φαινόμενα των μουσικών σημάτων όπως τις μικρο-μεταβολές που συμβαίνουν λόγω της μελωδίας, του ρυθμού κ.ά.

Σημαντικό συμπέρασμα της διερεύνησης αυτής αποτέλεσε η εισαγωγή της «μουσικής» συστοιχίας φίλτρων, μέσω της οποίας αναπτύχθηκαν σύνολα χαρακτηριστικών, η αξιολόγηση των οποίων επέδειξε ιδιαίτερη διακριτική ικανότητα στην κατηγοριοποίηση των διαφορετικών ειδών μουσικής. Συγκεκριμένα, τα χαρακτηριστικά αυτά απέδωσαν μείωση σφάλματος ως και 28% (84.15% ποσοστό επιτυχίας κατηγοριοποίησης) όταν συνδυάζονται με τα MFCC.

Επιπρόσθετα, οι αναπαραστάσεις του σήματος που βασίζονται στις μακροδομές της μουσικής (μέσω της συνένωσης διαδοχικών πλαισίων ανάλυσης), αν και δεν πέτυχαν το καλύτερο δυνατό αποτέλεσμα (82.88%), θεωρούμε πως είναι πολλά υποσχόμενες. Αυτό γιατί επιφέρουν μείωση της πολυπλοκότητας του συστήματος κατηγοριοποίησης με μικρούς χρόνους εκμάθησης καθώς επίσης επιτυγχάνουν ικανοποιητικά αποτελέσματα με απλούστερα στατιστικά μοντέλα τύπου GMM.

Τέλος αξίζει να αναφέρουμε πως το ποσοστό διαμόρφωσης της συχνότητας (FMP) για τις περισσότερες περιπτώσεις συμβάλλει στην επιτυχή κατηγοριοποίηση μειώνοντας το σφάλμα αναγνώρισης έως και 5% σε σχέση με τα MFCC. Από την άλλη, τα MFD, και συγκεκριμένα η φράκταλ διάσταση MFD$[s = 1]$, παρ' όλο που θεωρούμε πως παρουσιάζουν κάποια διακριτική ικανότητα, στις πειραματικές αξιολογήσεις που πραγματοποιήθηκαν κάποιες φορές λειτουργούν βοηθητικά ενώ άλλες όχι.

**Κατανομές Στιγμιαίων Σημάτων Διαμόρφωσης και Χαρακτηριστικών**

Στο σημείο αυτό παρουσιάζουμε και εξετάζουμε τις κατανομές (ιστογράμματα) των σημάτων του στιγμιαίου πλάτους καθώς και των μη-γραμμικών χαρακτηριστικών, δηλ. του μέσου στιγμιαίου πλάτους και της μέσης στιγμιαίας συχνότητας. Κατά την αξιολόγηση

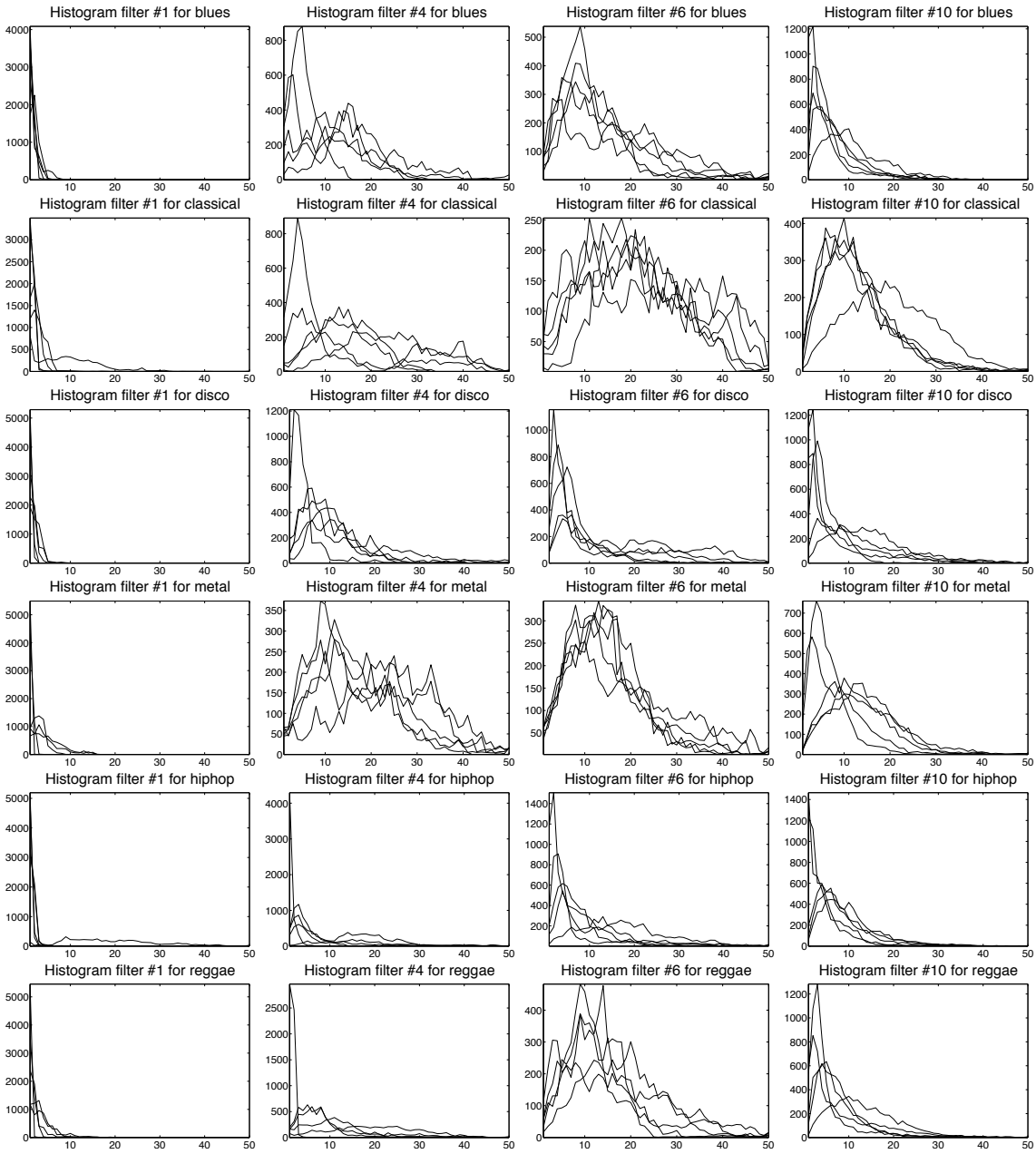
που προηγήθηκε παρατηρήσαμε ότι τα μουσικά είδη με τη χειρότερη κατηγοριοποίηση ήταν τα εξής: rock, reggae, hiphop και disco. Συγκεκριμένα, μουσικά κομμάτια της reggae μουσικής αναγνωρίζονταν για παράδειγμα ως hiphop ή disco, αντίστοιχα η disco μουσική ως hiphop, ενώ η rock ως pop ή country. Από την άλλη, το είδος με την καλύτερη αναγνώριση ήταν σχεδόν πάντα μουσικά κομμάτια της κατηγορίας classical και metal. Με την ανάλυση των κατανομών των αναπαραστάσεων των σημάτων μπορούμε να διαπιστώσουμε κατά πόσο η μειωμένη αναγνώριση των συγκεκριμένων ειδών οφείλεται στις αναπαραστάσεις των χαρακτηριστικών που εξάγουμε.

Στο Σχήμα 4.6 παρουσιάζονται τα ιστογράμματα των κατανομών των σημάτων του στιγμιαίου πλάτους για 5 μουσικά σήματα και για τα φίλτρα 1, 4, 6 και 10, για τα μουσικά είδη blues, classical, disco, metal, hiphop και reggae. Παρατηρούμε πως οι κατανομές για την κλασική μουσική (2η σειρά) και τη metal (4η σειρά) παρουσιάζουν αρκετές διαφοροποιήσεις σε σχέση με τις υπόλοιπες, ενώ ομοιότητες εντοπίζονται στις κατανομές για το πρώτο, δεύτερο και τέταρτο φίλτρο των hiphop και reggae (σειρά 6 και 7), για το πρώτο φίλτρο όλων των ειδών, αλλά και μεταξύ του δεύτερου φίλτρου για τα είδη classical και blues (1η και 2η σειρά). Όντως η hiphop μουσική είναι ένα είδος με αρκετές επιρροές τόσο από τη reggae αλλά και άλλα είδη όπως η jazz, funk κ.ά, σε αντίθεση για παράδειγμα με την κλασική η οποία δομικά και ακουστικά (όσον αφορά τη χροιά) διαφοροποιείται σε σχέση με τα υπόλοιπα μουσικά είδη της αξιολόγησης.

Στο Σχήμα 4.7 παρουσιάζονται οι κατανομές των προτεινόμενων μη-γραμμικών χαρακτηριστικών και συγκεκριμένα του μέσου στιγμιαίου πλάτους για 250 μουσικά σήματα των 10 διαφορετικών ειδών και για τα φίλτρα 1-5. Για καλύτερη παρουσίαση των κατανομών τα bins του ιστογράμματος είναι τοποθετημένα σε λογαριθμική κλίμακα. Και πάλι εντοπίζουμε αρκετές ομοιότητες μεταξύ των διαφορετικών ειδών και συγκεκριμένα για το πρώτο φίλτρο.

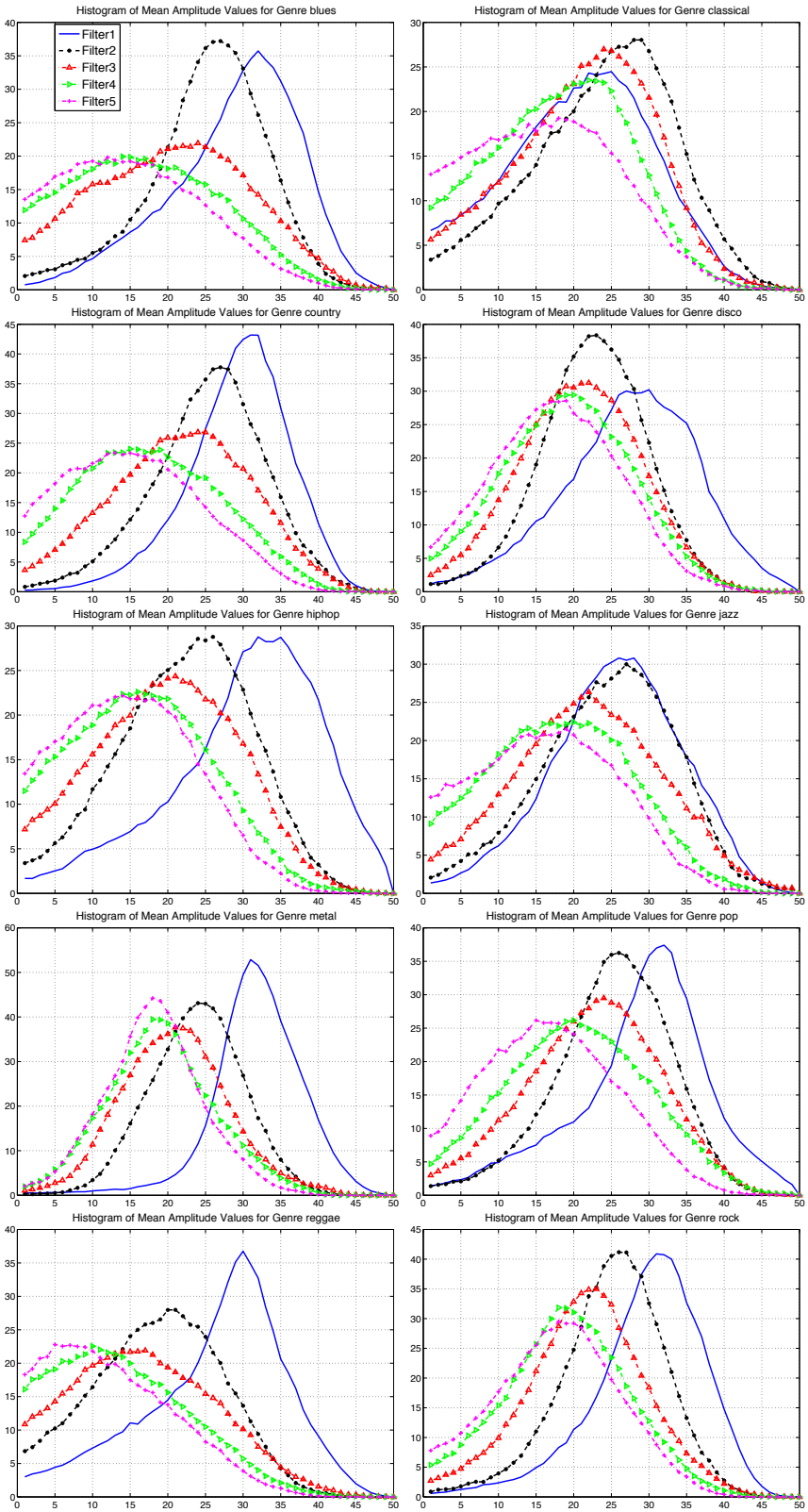
Τέλος, στο Σχήμα 4.8 φαίνονται τα ιστογράμματα των κατανομών της μέσης στιγμιαίας συχνότητας για 250 μουσικά σήματα των 10 διαφορετικών ειδών και για τα φίλτρα 1-5. Η πρώτη μας παρατήρηση αφορά το γεγονός πως οι μέσες τιμές της στιγμιαίας συχνότητας κατανέμονται γύρω από τις κεντρικές συχνότητες των φίλτρων. Αυτό παρατηρήθηκε τόσο για μουσικά σήματα διαφορετικών μουσικών οργάνων όσο και σε προηγούμενες μελέτες που αφορούν τη φωνή και τις κατανομές της συγκεκριμένης μέτρησης σε διαφορετικά φωνήματα [32]. Διαφοροποιήσεις στις κατανομές μεταξύ των ειδών αφορούν κυρίως τη διασπορά των κατανομών και κατά συνέπεια την επικάλυψη μεταξύ των φίλτρων (για παράδειγμα βλ. τις κατανομές της metal και της pop μουσικής, 4η σειρά) αλλά και τη διαφορά στη μέση τιμή της κατανομής για το ίδιο φίλτρο.

Μετά από την ανάλυση των κατανομών διαπιστώνουμε πως η λάθος κατηγοριοποίηση των μουσικών σημάτων οφείλεται πιθανώς στις ομοιότητες των μη-γραμμικών

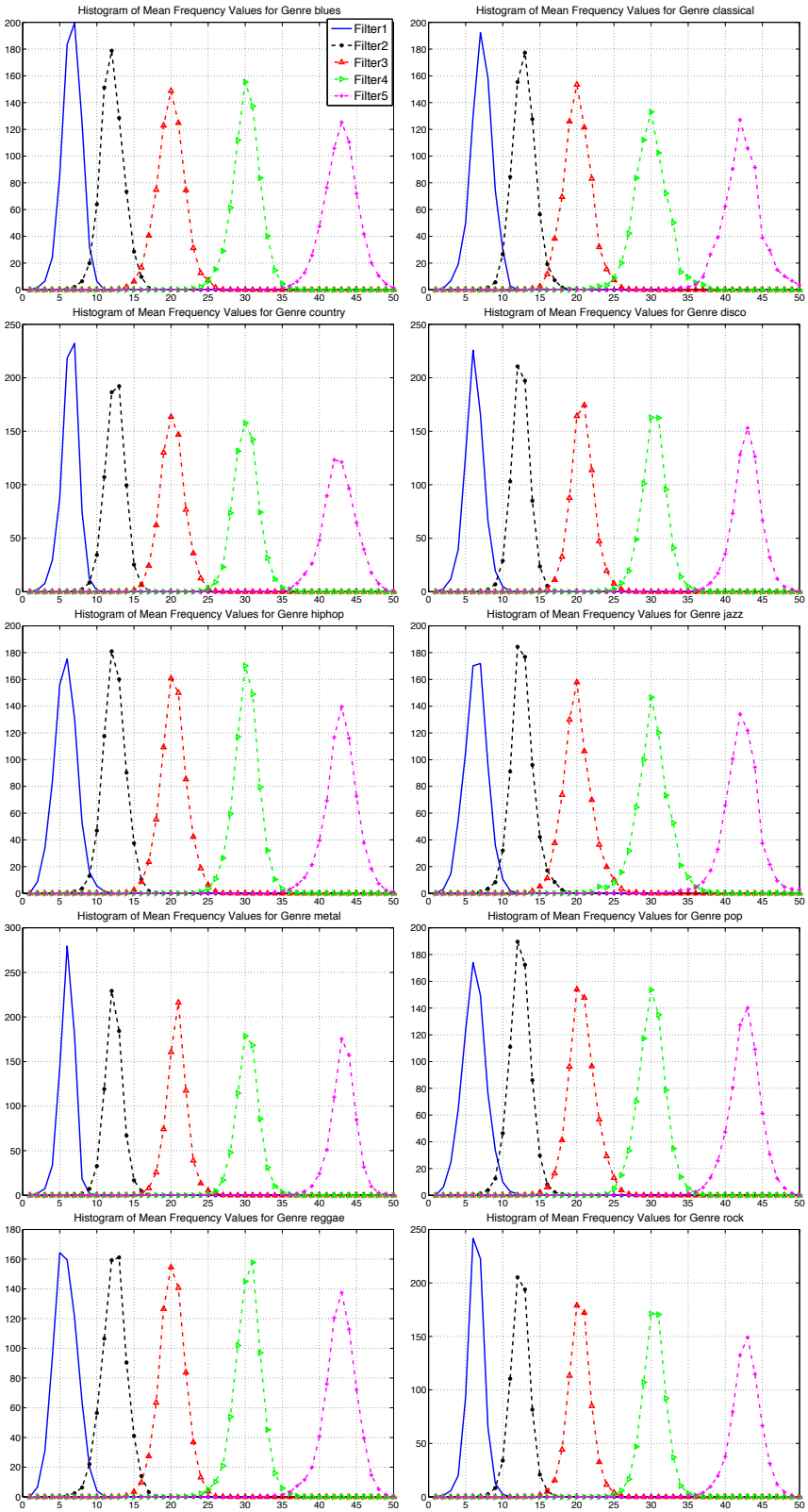



*Σχήμα 4.6:* Ιστογράμματα κατανομών 5 σημάτων στιγμιαίου πλάτους για τα φίλτρα 1, 4, 6 και 10, για τα μουσικά είδη blues, classical, disco, metal, hiphop και reggae (από πάνω προς τα κάτω).

χαρακτηριστικών. Παρ' όλα αυτά πρέπει να τονίσουμε πως οι ομοιότητες αυτές αφορούν μουσικά είδη με παρεμφερή γνωρίσματα.




*Σχήμα 4.7:* Ιστογράμματα κατανομών, 250 μουσικών σημάτων 10 διαφορετικών ειδών, του μέσου στιγμιαίου πλάτους m-IAM για τα 5 πρώτα φίλτρα.




*Σχήμα 4.8:* Ιστογράμματα κατανομών, 250 μουσικών σημάτων 10 διαφορετικών ειδών, της μέσης στιγμιαίας συχνότητας m-IFM για τα 5 πρώτα φίλτρα.

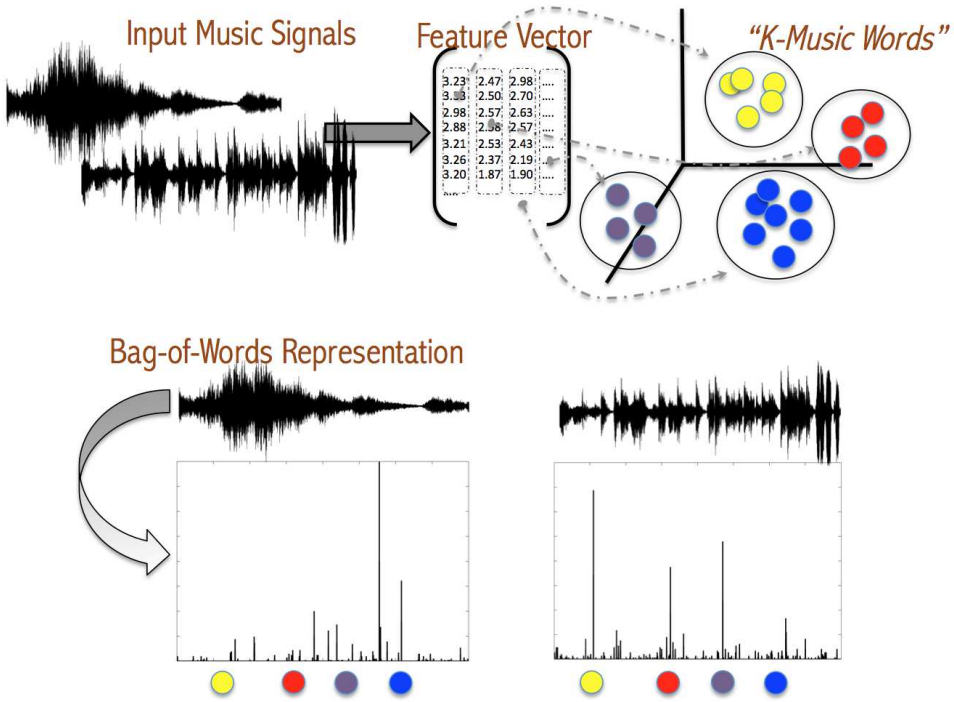
## 4.4 Bag-of-Words (BoW) Μοντέλα για την Ανάλυση Διαφορετικών Ειδών Μουσικής

Στο δεύτερο σετ πειραμάτων αναγνώρισης μουσικών σημάτων διαφορετικών ειδών μουσικής εξετάζουμε μία διαφορετική μοντελοποίηση των χαρακτηριστικών, ιδιαίτερα δημοφιλή σε εφαρμογές της Όρασης Υπολογιστών, γνωστή ως Bag-of-Words (BoW). Η αξιολόγηση πραγματοποιείται με τη χρήση των Support Vector Machines (SVMs) που τα τελευταία χρόνια αποτελεί την κύρια μέθοδο κατηγοριοποίησης σε πολλές διαφορετικές εφαρμογές με υψηλά ποσοστά επιτυχίας.

Τα μοντέλα BoW προτάθηκαν αρχικά για την ανάλυση κειμένων και την κατηγοριοποίησή τους βάσει θέματος [161, 195]. Στη συνέχεια χρησιμοποιήθηκαν σε εφαρμογές της όρασης υπολογιστών, όπως η κατηγοριοποίηση αντικειμένων, εικόνων, σκηνών βίντεο και δράσεων, όπου τα χαρακτηριστικά (όπως για παράδειγμα το χρώμα, η υφή κ.ά.) μοντελοποιούνται ως οπτικές «λέξεις» [69, 83, 192, 196] (για επισκόπηση βλ. [177]).

Η ιδέα για τη δημιουργία των μοντέλων BoW (στην όραση υπολογιστών) βασίζεται στην απεικόνιση κάθε εικόνας ως ένα σύνολο σημείων κλειδιών (*interest points*) και χαρακτηριστικών. Για τη δημιουργία συμπαγών αναπαραστάσεων, τα χαρακτηριστικά αυτά ομαδοποιούνται, συνήθως σε έναν μεγάλο αριθμό κέντρων, μέσω κάποιας μεθόδου ομαδοποίησης (*clustering*). Με αυτό τον τρόπο, κάθε κέντρο περιλαμβάνει χαρακτηριστικά με παρόμοιες ιδιότητες, δημιουργώντας μία «οπτική λέξη» που αναπαριστά ένα συγκεκριμένο «πρότυπο». Αποτέλεσμα της διαδικασίας αυτής είναι η δημιουργία ενός «λεξικού οπτικών λέξεων». Μέσω της χαρτογράφησης των χαρακτηριστικών στις λέξεις του λεξικού η κάθε εικόνα περιγράφεται βάσει της συχνότητας των λέξεων τις οποίες περιλαμβάνει. Τα ιστογράμματα αυτά αποτελούν τις αναπαραστάσεις «*bag-of-words*» διαμορφώνοντας έτσι τα νέα διανύσματα χαρακτηριστικών που αποτελούν είσοδο σε κάποιο συστήματα αναγνώρισης προτύπων (όπως για παράδειγμα τα SVMs). Η μέθοδος αυτή, αν και ιδιαίτερα απλή, έχει φανεί εξαιρετικά αποδοτική σε εφαρμογές κατηγοριοποίησης εικόνων και σκηνών σε βίντεο, παρ' όλο που δεν περιέχει καμία πληροφορία για τη γεωμετρία των εικόνων [69]. Εμπνεόμενοι από την επιτυχία και αποδοτικότητα των BoW τόσο στην ανάλυση κειμένων όσο και στην κατηγοριοποίηση αντικειμένων τα εφαρμόζουμε σε σήματα μουσικής και συγκεκριμένα για την αξιολόγηση της απόδοσής τους στην αναγνώριση των διαφορετικών ειδών μουσικής.

Ήδη στην εισαγωγή αναδείξαμε τις ομοιότητες της δομής της μουσικής τόσο με τη φωνή όσο και με το γραπτό κείμενο. Θεωρίες υποστηρίζουν την παράλληλη εξέλιξή τους, ενώ παρουσιάζουν τα δομικά αυτά στοιχεία τα οποία συνθέτουν τις τελικές οργανωμένες δομές για το σχηματισμό λέξεων/προτάσεων ή μελωδιών. Το κείμενο και η μουσική όπως




*Σχήμα 4.9:* Βήματα για τη δημιουργία Bag-of-Words από μουσικά σήματα.

αναφέρθηκε «είναι σαν δυο αδέλφια, τα οποία προέρχονται από την ομιλία, που κατά την πάροδο των χρόνων εξελίχθηκαν διαφορετικά», το ένα με έμφαση στη διάσταση του ήχου και το άλλο στη διάσταση της νόησης. Για τη δημιουργία Bag-of-Words αναπαραστάσεων από μουσικούς ήχους, θεωρούμε το μουσικό κομμάτι παραπλήσιο του κειμένου, το οποίο, κατ' αναλογία με τη διαφορετική θεματολογία του κειμένου, αποτελείται από μουσικές φράσεις, μοτίβα καθώς και άλλα δομικά στοιχεία (π.χ. ρυθμός, ταχύτητα, αρμονία, μελωδία) τα οποία συνθέτουν το τελικό αποτέλεσμα. Τα δομικά στοιχεία για τη δημιουργία δομημένων ακολουθιών και εν συνεχεία ολοκληρωμένων μουσικών συνθέσεων ακολουθούν συνήθως κάποιους βασικούς κανόνες. Η διαφορετικότητα της μουσικής (και άρα τα διαφορετικά μουσικά είδη) οφείλεται στην παραλλαγή των δομικών στοιχείων που χρησιμοποιούνται. Θεωρούμε άρα, πως τα δομικά αυτά στοιχεία συνθέτουν το τελικό μουσικό κομμάτι, όπως για παράδειγμα οι λέξεις σε ένα κείμενο ορίζουν το τελικό νόημα μιας πρότασης. Ακολουθώντας τη συγκεκριμένη μεθοδολογία αναπαριστάμε το κάθε μουσικό κομμάτι ως μία μίξη διαφορετικών δομικών στοιχείων, θεμάτων, φράσεων ή μοτίβων, η κατανομή των οποίων μοντελοποιείται ως μουσικές λέξεις και χρησιμοποιείται για να το χαρακτηρίσει ως ένα συγκεκριμένο μουσικό είδος.

### Bag-of-Words Πλαίσιο για τη Δημιουργία Μουσικών Λέξεων

Η διαδικασία που απαιτείται για τη δημιουργία των BoW αναπαραστάσεων συνοψίζεται στα ακόλουθα βήματα (βλ. Σχήμα 4.9): (α) εξαγωγή χαρακτηριστικών/εύρωστων

περιγραφέων από το μουσικό σήμα, (β) δημιουργία λέξεων μέσω των αναπαραστάσεων αυτών για το σχηματισμό του μουσικού λεξικού, και (γ) υπολογισμός της συχνότητας των λέξεων (ιστογράμματα) σε κάθε μουσικό κομμάτι για τη δημιουργία των BoW αναπαραστάσεων.

***Βήμα 1: Εξαγωγή Χαρακτηριστικών***. Συγκεκριμένα, τα χαρακτηριστικά τα οποία χρησιμοποιούμε βασίζονται στη θεωρία των διαμορφώσεων, στη φράκταλ διάσταση και στα MFCC. Παρ' όλα αυτά οποιαδήποτε αναπαράσταση του μουσικού σήματος θα μπορούσε να μοντελοποιηθεί ως BoW.

Για τη δημιουργία εύρωστων και περιγραφικών αναπαραστάσεων εφαρμόζουμε αλγόριθμο επιλογής χαρακτηριστικών (*feature selection algorithm*), ο οποίος επιλέγει τα χαρακτηριστικά τα οποία βοηθούν περισσότερο στο συγκεκριμένο πρόβλημα. Στην εργασία αυτή, έχουμε χρησιμοποιήσει τον αλγόριθμο SFFS για την εύρεση του κατάλληλου υποσυνόλου χαρακτηριστικών χρησιμοποιώντας ως κριτήριο αξιολόγησης το άθροισμα της εκτιμώμενης απόστασης Mahalanobis. Η τελική επιλογή των χαρακτηριστικών αποφασίζεται μετά από πειραματική αξιολόγηση.

Επιπλέον, στις εφαρμογές αναγνώρισης και ειδικά στα SVMs είναι πολύ σημαντική η κανονικοποίηση των χαρακτηριστικών ως βήμα προ-επεξεργασίας. Η κανονικοποίηση μεταφέρει όλα τα χαρακτηριστικά στην ίδια κλίμακα και κατά αυτόν τον τρόπο εμποδίζει την ύπαρξη μεγάλων πιθανώς τιμών που συνήθως επηρεάζουν αρνητικά το αποτέλεσμα κατηγοριοποίησης. Για το λόγο αυτό, χρησιμοποιούμε z-score standardization όπου το κάθε χαρακτηριστικό $A_j^n$ μετατρέπεται έτσι ώστε να έχει μηδενική μέση τιμή και μοναδιαία διακύμανση $x_{ji}^* = (x_{ij} - \mu_j)/\sigma_j$, όπου $\mu_j$ και $\sigma_j$ η μέση τιμή και η διακύμανση του χαρακτηριστικού $j$ του $A_j^n$ αντίστοιχα.

***Βήμα 2: Δημιουργία Μουσικών Λέξεων και Μουσικού Λεξικού***. Επόμενο βήμα είναι η επεξεργασία των διανυσμάτων χαρακτηριστικών για τη δημιουργία της αναπαράστασης του μουσικού κομματιού ως ένα ιστόγραμμα «μουσικών λέξεων». Συγκεκριμένα, αυτό επιτυγχάνεται ομαδοποιώντας τα χαρακτηριστικά των μουσικών σημάτων, όλων των δεδομένων εκπαίδευσης σε έναν μεγάλο αριθμό $K$-κέντρων με τον αλγόριθμο K-means. Μέσω της διαδικασίας αυτής το κάθε διάνυσμα χαρακτηριστικών ομαδοποιείται σε κάποιο από τα κέντρα που έχουν υπολογιστεί και αποτελεί μια διαφορετική μουσική λέξη. Το σύνολο των λέξεων απαρτίζει το «μουσικό λεξικό» που περιγράφει το σύνολο των μουσικών δεδομένων και συγκεκριμένα των διαφορετικών ειδών.

Θεωρώντας άρα ένα σύνολο δεδομένων εκπαίδευσης $D = d_1, d_2, ..., d_n$, όπου $d$ τα σύνολα χαρακτηριστικών $n$ μουσικών σημάτων, τότε με την εφαρμογή του K-means πραγματοποιείται η ομαδοποίηση των χαρακτηριστικών $D$ με βάση ένα σταθερό αριθμό κέντρων $K$, δημιουργώντας το μουσικό λεξικό $W = w_1, w_2, ..., w_k$ το οποίο αντιπροσωπεύεται από $K$ μουσικές λέξεις. Εν συνεχεία, το κάθε μουσικό κομμάτι

αναπαριστάται ως ένα ιστόγραμμα, $K \times N$, όπου $N_{ij} = n(w_i, d_j)$ και $n(w_i, d_j)$ η συχνότητα της λέξης $w_i$ σε ένα μουσικό κομμάτι $d_j$ [16].

Το αποτέλεσμα αυτής της αναπαράστασης είναι συνήθως ένα αρκετά αραιό διάνυσμα χαρακτηριστικών, του οποίου η διάσταση ορίζεται από τον αριθμό των κέντρων. Αυτό έχει ως άμεση συνέπεια τη μείωση της υπολογιστικής πολυπλοκότητας καθώς πλέον το πρόβλημα της αναγνώρισης ανάγεται στην εύρεση ομοιοτήτων μεταξύ των μουσικών κομματιών μέσω της αναπαράστασης bag-of-words.

Σημαντική παράμετρος για τη δημιουργία του μουσικού λεξικού αποτελεί η επιλογή του αριθμού των κέντρων [192]. Χρησιμοποιώντας μικρό αριθμό κέντρων υπάρχει περίπτωση ανόμοια χαρακτηριστικά να ομαδοποιηθούν στα ίδια κέντρα και άρα οι «λέξεις» που θα δημιουργηθούν να υστερούν σε διακριτική ικανότητα, ενώ δημιουργώντας ένα πιο λεπτομερές λεξικό με πολλά κέντρα θα έχει ως συνέπεια προβλήματα γενίκευσης, αυξημένη πολυπλοκότητα στην επεξεργασία καθώς και δημιουργία θορύβου αφού παρόμοια διανύσματα χαρακτηριστικών πιθανότατα θα ομαδοποιηθούν σε διαφορετικά κέντρα. Για το λόγο αυτό, ο αριθμός των κέντρων στο state-of-the-art για εφαρμογές της όρασης υπολογιστών ποικίλει από 200 έως και 10000.

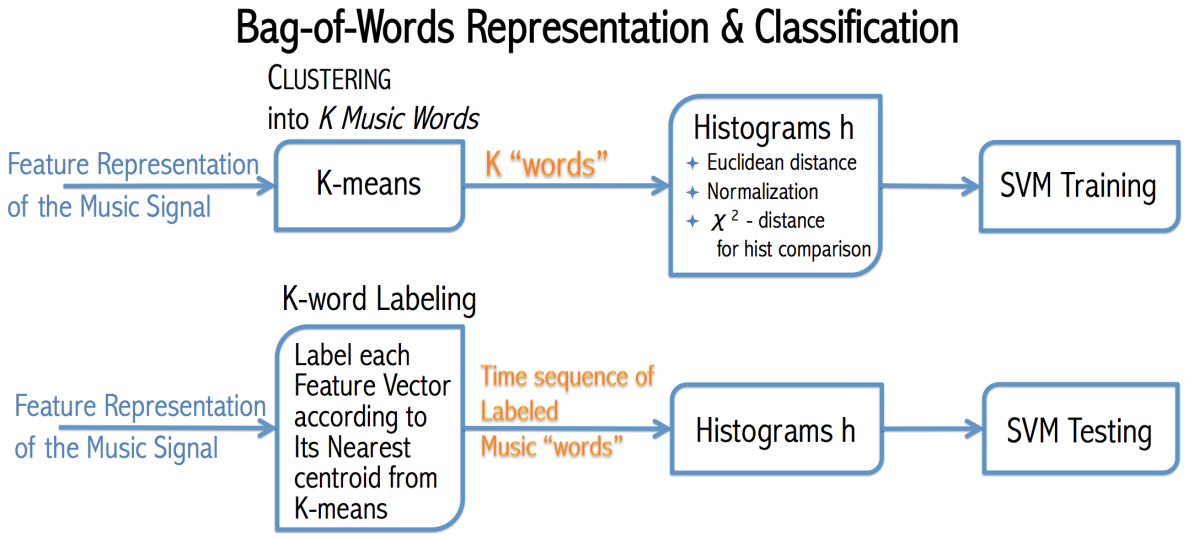
Μετά από αξιολόγηση διαφορετικού αριθμού κέντρων για τη δημιουργία του μουσικού λεξικού βρέθηκε πως τα 4000 κέντρα επιτυγχάνουν τα καλύτερα ποσοστά αναγνώρισης σε σχέση με τα 2000 ή τα 8000 κέντρα. Στα πειράματα που ακολουθούν χρησιμοποιούμε τον αλγόριθμο K-means[2] με 4000 κέντρα.

***Βήμα 3: Απόσταση $\chi^2$ (chi-square) για τη σύγκριση των BoW***. Για την τελική αναγνώριση και κατηγοριοποίηση των διαφορετικών ειδών μουσικής χρησιμοποιούμε μη-γραμμικά SVMs. Τα SVMs αποτελούν έναν από τους πιο δημοφιλείς ταξινομητές για τη μοντελοποίηση αυτή. Σημαντική παράμετρος σε αυτή την περίπτωση είναι η επιλογή του πυρήνα. Αν και υπάρχει πληθώρα διαφορετικών πυρήνων στη βιβλιογραφία, πρόσφατες ερευνητικές εργασίες [69, 195] έχουν δείξει την αποτελεσματικότητα του $\chi^2$ (chi-squared) RBF σε σχέση με τους παραδοσιακούς πυρήνες όπως ο γραμμικός ή ο Gaussian RBF.

Επιπρόσθετα, η απόσταση $\chi^2$ χρησιμοποιείται συχνά σε εφαρμογές της όρασης υπολογιστών για τον υπολογισμό των αποστάσεων μεταξύ των BoW αναπαραστάσεων των εικόνων. Το όνομα της απόστασης προέρχεται από το στατιστικό μετρικό του Pearson (*Pearson's chi squared test statistic*) και χρησιμοποιείται για τη σύγκριση διακριτών κατανομών πιθανοτήτων (δηλ. ιστογραμμάτων). Ωστόσο, σε αντίθεση με το μετρικό του Pearson, η απόσταση $\chi^2(H_i, H_j)$ είναι συμμετρική όσον αφορά τα $H_i, H_j$, κάτι που την καθιστά χρήσιμη για τη δημιουργία πυρήνων, βάσει των αποστάσεων των ιστογραμμάτων, σε εφαρμογές κατηγοριοποίησης με μη γραμμικά SVMs.

Για τους λόγους που προαναφέρθηκαν, στα πειράματα που ακολουθούν, λαμβάνοντας

[2]Για τον αλγόριθμο K-means χρησιμοποιήθηκε η έτοιμη βιβλιοθήκη Yael του πανεπιστημίου Inria [191].




*Σχήμα 4.10:* Συνολική διαδικασία για τη δημιουργία «μουσικών λέξεων» και Bag-of-Words αναπαραστάσεων από την εξαγωγή χαρακτηριστικών ως και την τελική κατηγοριοποίηση.

υπόψη πως τα BoW είναι ιστογράμματα που μετράνε τη συχνότητα των μουσικών λέξεων, χρησιμοποιούμε ένα γενικευμένο Γκαουσιανό $\chi^2$ πυρήνα, όπου $\chi^2(H_i, H_j)$ η $\chi^2$ απόσταση ανάμεσα σε δύο ιστογράμματα $H_i = [h_{i1}, .., h_{ik}]$ και $H_j = [h_{j1}, ..., h_{jk}]$ με $K$ κέντρα [25,70, 83, 151, 195]. Η απόσταση $\chi^2$ για τη σύγκριση δύο (κανονικοποιημένων) ιστογραμμάτων $H_i$ και $H_j$ ορίζεται ως:

$$\chi^2(H_i, H_j) = \frac{1}{2} \sum^{K} \frac{[h_{ik} - h_{jk}]^2}{h_{ik} + h_{jk}} \tag{4.9}$$

όπου $K$ το μέγεθος του μουσικού λεξικού, ενώ ο $\chi^2$ πυρήνας υπολογίζεται ως:

$$\mathcal{K}(H_i, H_j) = \exp\left(-\frac{1}{A}\chi^2(H_i, H_j)\right), \tag{4.10}$$

όπου $H_i = h_{ik}$ και $H_j = h_{jk}$ τα ιστογράμματα και $A$ η μέση τιμή της απόστασης μεταξύ όλων των παραδειγμάτων εκπαίδευσης.

Όσον αφορά τα SVMs, μιας και το πρόβλημα το οποίο εξετάζουμε περιλαμβάνει περισσότερες από δύο κατηγορίες, χρησιμοποιούμε τη μέθοδο one-against-all, δηλ. εκπαιδεύουμε ένα μοντέλο για κάθε κατηγορία σε σχέση με τις υπόλοιπες. Τέλος, διεξάγουμε βελτιστοποίηση και εύρεση της καλύτερης τιμής της παραμέτρου του κόστους. Στο Σχήμα 4.10 παρουσιάζεται η συνολική διαδικασία για τη δημιουργία μουσικών λέξεων και BoW αναπαραστάσεων για μουσικά σήματα, από την εξαγωγή χαρακτηριστικών ως την τελική κατηγοριοποίηση.

*Πίνακας 4.12:* Λίστα χαρακτηριστικών τα οποία χρησιμοποιήθηκαν για τη μοντελοποίηση με Bag-of-Words και αξιολογήθηκαν με SVMs.

| **Χαρακτηριστικά** | | **Περιγραφή** |
|---|---|---|
| 1 | $\text{LMF}_{58}$ | 58 AMFM χαρακτηριστικά μετά από επιλογή στο αρχικό διάνυσμα $\text{LMF}_{72}$ 72 χαρακτηριστικών αποτελούμενο από 12 $\log$(m-IAM) + 12 m-IFM και τις παραγώγους τους. |
| 2 | $\text{LMFP}_{50}$ | 50 AMFM+FMP χαρακτηριστικά μετά από επιλογή στο συνολικό διάνυσμα $\text{LMFP}_{108}$ 108 χαρακτηριστικών αποτελούμενο από 12 $\log$(m-IAM) + 12 m-IFM + 12 FMP και τις παραγώγους τους. |
| 3 | $\text{MFC}_{21}$ | 21 MFCC χαρακτηριστικά μετά από επιλογή στο αρχικό διάνυσμα $\text{MFC}_{\Delta}$ 39 χαρακτηριστικών. |
| 4 | LMF-$\text{MFC}_{74}$ | 74 χαρακτηριστικά μετά από επιλογή στο συνολικό διάνυσμα $\text{LMF}_{72}$ και $\text{MFC}_{\Delta}$. |
| 5 | $\text{LMF}_{58}$-$\text{MFC}_{21}$ | 79 χαρακτηριστικά μετά από επιλογή στα ξεχωριστά διανύσματα $\text{LMF}_{72}$ και $\text{MFC}_{\Delta}$. |
| 6 | LMF-MFC-$\text{D}_{60}$ | 60 χαρακτηριστικά μετά από επιλογή στο συνολικό διάνυσμα $\text{LMF}_{72}$, $\text{MFC}_{\Delta}$ και $\text{MFD}_{58}$. |
| 7 | LMFiD-$\text{MFC}_{66}$ | 66 χαρακτηριστικά μετά από επιλογή στο συνολικό διάνυσμα $\text{LMFiD}_{40}\text{MF}_{72}$ το οποίο έχει προέλθει μετά από PCA ανάλυση των $\text{LFM}_{72}$ + $\text{MFD}[s=1]$ + $\text{MFC}_{\Delta}$ (βλ. Πίνακα 4.4). |

### 4.4.1 Πειραματική Αξιολόγηση: Σύνολα Χαρακτηριστικών

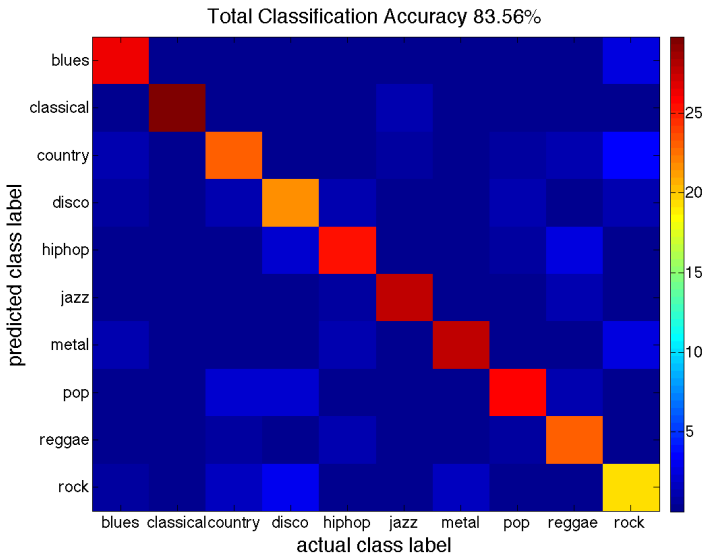
Στον Πίνακα 4.12 φαίνονται κάποια από τα σύνολα χαρακτηριστικών τα οποία χρησιμοποιήθηκαν για τη μοντελοποίηση με BoW μοντέλα και αξιολογήθηκαν με SVMs. Σημειώνουμε πως τα σύνολα αυτά έχουν προέλθει μετά από μείωση της αρχικής τους διάστασης, μέσω αλγορίθμου επιλογής χαρακτηριστικών και πειραματισμό για την εύρεση του βέλτιστου τελικού συνόλου. Στην πειραματική αξιολόγηση που ακολουθεί παρουσιάζουμε μόνο τα σύνολα αυτά τα οποία παρουσίασαν τα μεγαλύτερα ποσοστά επιτυχίας.

### 4.4.2 Πειραματική Αξιολόγηση: Αποτελέσματα

Στον Πίνακα 4.13 φαίνονται τα ποσοστά επιτυχίας κατηγοριοποίησης για τα 10 μουσικά είδη της βάσης GTZAN μετά από 5 cross-validation. Παρατηρούμε πως οι BoW αναπαραστάσεις, με τη χρήση των μη-γραμμικών AM-FM χαρακτηριστικών επιτυγχάνουν καλύτερα αποτελέσματα από τα MFCC και επιφέρουν μείωση του σφάλματος περίπου

*Πίνακας 4.13:* Ποσοστά επιτυχίας κατηγοριοποίησης (%) για 10 μουσικά είδη με Support Vector Machines και χαρακτηριστικά βραχέος χρόνου βασιζόμενοι στα Bag-of-Words μοντέλα.

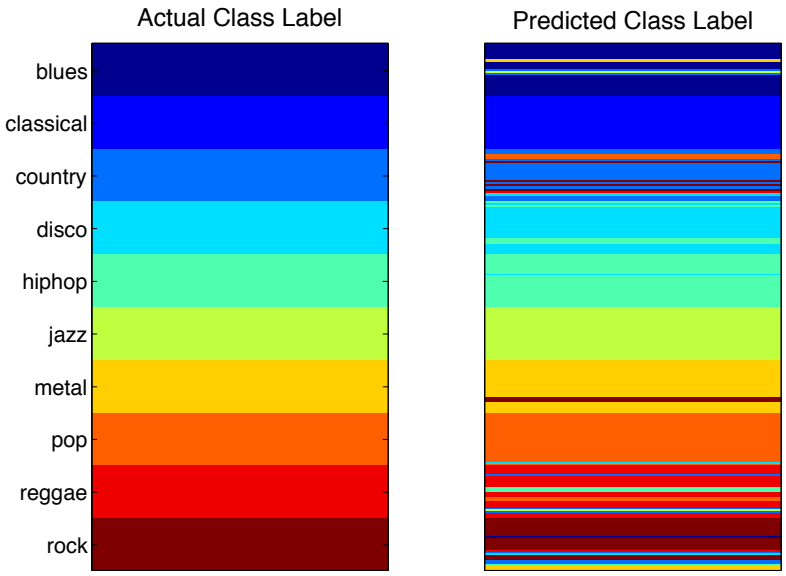
| **Σύνολα χαρακτηριστικών** | **Mean Accuracy** | **Max** |
|---|---|---|
| $LMF_{58}$ | 82.62 | 85.00 |
| $LMFP_{50}$ | 82.42 | 83.67 |
| $MFC_{21}$ | 80.54 | 83.61 |
| LMF-$MFC_{74}$ | 82.88 | 84.33 |
| $LMF_{58}MFC_{21}$ | 82.56 | 83.28 |
| LMF-MFC-$D_{60}$ | 81.48 | 83.33 |
| LMFiD-$MFC_{66}$ | **83.56** | 85.33 |




*Σχήμα 4.11:* Confusion Matrix για τα 10 είδη μουσικής και το σύνολο χαρακτηριστικών LMFiD-$MFC_{66}$ με συνολικό ποσοστό επιτυχίας 83.56%.

11% για το σύνολο $LMF_{58}$ σε σύγκριση με τα MFCC και 16% σε συνδυασμό με τα MFCC (για το σύνολο LMFiD-$MFC_{66}$).

Γενικά παρατηρήσαμε πως τα ιστογράμματα που δημιουργήθηκαν για τα διάφορα μουσικά είδη ήταν σχετικά πυκνά, κάτι που θεωρούμε ως βασικό λόγο για τις περιπτώσεις μη εύστοχης διάκρισης των ειδών. Στο Σχήμα 4.11 παρουσιάζουμε το confusion matrix για 10 είδη μουσικής και το σύνολο χαρακτηριστικών LMFiD-$MFC_{66}$ με συνολικό ποσοστό επιτυχίας κατηγοριοποίησης 83.56%. Καλύτερη απόδοση και σε αυτήν την περίπτωση έχει η κλασική μουσική ενώ χειρότερη η rock. Τέλος, το Σχήμα 4.12 παρουσιάζει την πραγματική έναντι της προβλεπόμενης κατηγορίας για το ίδιο σύνολο χαρακτηριστικών (LMFiD-$MFC_{66}$) και το fold με το μέγιστο ποσοστό επιτυχίας 85.33%.

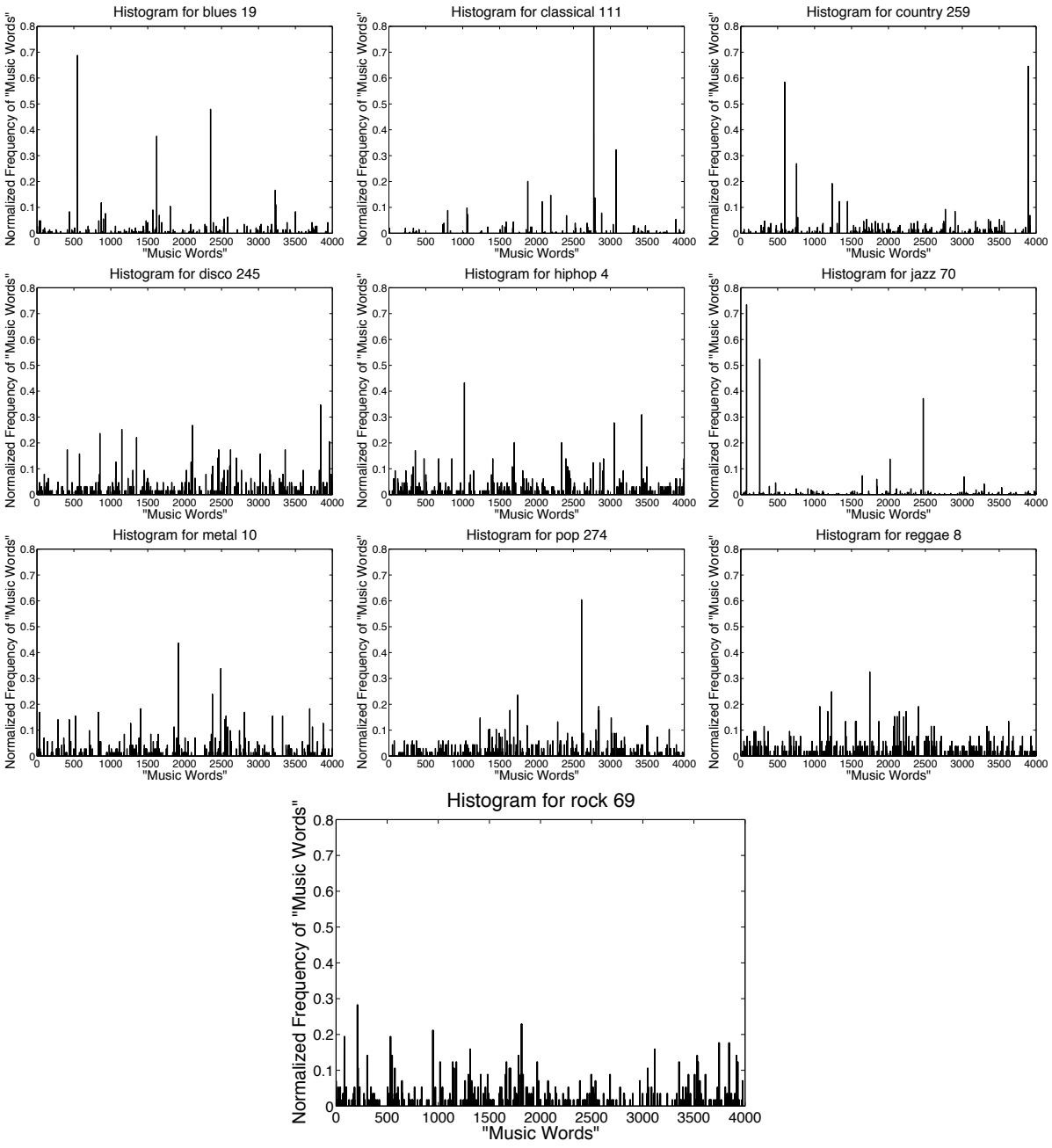



*Σχήμα 4.12:* Πραγματική vs. προβλεπόμενη κατηγορία για δέκα είδη μουσικής και το σύνολο χαρακτηριστικών LMFiD-MFC$_{66}$ και το fold με το μέγιστο ποσοστό επιτυχίας 85.33%.

Τέλος, στο Σχήμα 4.13 φαίνονται παραδείγματα BoW αναπαραστάσεων των 10 διαφορετικών ειδών μουσικής για το σύνολο χαρακτηριστικών LMFiD-MFC$_{66}$. Τα συγκεκριμένα παραδείγματα ιστογραμμάτων είχαν σωστή αναγνώριση για το συγκεκριμένο fold με μέγιστο ποσοστό επιτυχίας.Όντως παρατηρούμε πως τα ιστογράμματα δεν είναι ιδιαίτερα αραιά και περιλαμβάνουν πολλές μουσικές λέξεις έστω και σε μικρή συχνότητα. Παρ' όλα αυτά παρουσιάζουν συγκεκριμένες κορυφές (δηλ. μεγάλο αριθμό συγκεκριμένων μουσικών λέξεων) στις οποίες πιθανώς οφείλεται η επιτυχής αναγνώριση.

**Πειραματικά αποτελέσματα για την βάση ARTISTS.** Στην εφαρμογή της αναγνώρισης των διαφορετικών ειδών μουσικής με μοντέλα BoW πειραματιστήκαμε και με τη βάση ARTISTS, βλ. Ενότητα 2.5, χρησιμοποιώντας έξι διαφορετικά είδη: Classical, Comedy & SpokenWord, Electronic & Dance, HipHop, Latin και Rock & Pop. Παρατηρούμε πως στη συγκεκριμένη βάση υπάρχουν κατηγορίες οι οποίες αποτελούνται από διπλή ονομασία, π.χ. Rock & Pop, Electronic & Dance. Αυτό πιστεύουμε ότι πολύ πιθανώς δυσκολεύει τη διαδικασία αναγνώρισης γι' αυτό και χρησιμοποιούμε έξι μόνο διαφορετικά είδη.

Η μοντελοποίηση των σημάτων πραγματοποιείται όπως περιγράφηκε παραπάνω ενώ τα πειράματα διενεργούνται για 5 cross-validation. Κάποια από τα τα ποσοστά επιτυχίας τα οποία επιτεύχθηκαν για τα σύνολα χαρακτηριστικών είναι τα ακόλουθα: (α) το σύνολο μη-γραμμικών διαμορφώσεων LMF$_{72}$ παρουσίασε ποσοστό επιτυχίας 78.34%, (β) το σύνολο LMFP$_{108}$ 79.31% ενώ (γ) τα MFC$_{\Delta}$ 73.07%. Τα ποσοστά αυτά διαμορφώνουν τη μείωση σφάλματος σε περίπου 23% και 20% με τη χρήση των LMFP$_{108}$ και LMF$_{72}$ αντίστοιχα σε




*Σχήμα 4.13:* Παραδείγματα Bag-of-Words αναπαραστάσεων των 10 ειδών μουσικής με σωστή αναγνώριση για το σύνολο χαρακτηριστικών LMFiD-MFC$_{66}$ και το fold με το μέγιστο ποσοστό επιτυχίας 85.33%.

σχέση με MFC$_{\Delta}$.

Διαπιστώνουμε πως τα μη-γραμμικά AM-FM χαρακτηριστικά ακόμα και χωρίς καμία παραμετροποίηση (π.χ. feature selection, PCA) επιτυγχάνουν καλύτερο ποσοστό κατηγοριοποίησης σε σχέση με τα MFCC, μάλιστα σε ένα πρόβλημα που ασχέτως με τον αριθμό των τάξεων θεωρούμε πως είναι δύσκολο λόγω της μίξης διαφορετικών ειδών σε μία κατηγορία.

**Συμπεράσματα**

Παρουσιάσαμε και προτείναμε τη χρήση των Bag-of-Words μοντέλων, υποκινούμενοι από παρόμοιες ιδέες οι οποίες με επιτυχία χρησιμοποιούνται και εφαρμόζονται στην αναγνώριση κειμένων, και αντικειμένων σε εικόνες και βίντεο. Με βάση τα αποτελέσματα της αξιολόγησης και της δεδομένης μείωσης του σφάλματος ως και 16% για δέκα διαφορετικά μουσικά είδη και τον συνδυασμό των προτεινόμενων χαρακτηριστικών με τα MFCC, και 23% για έξι είδη σε σύγκριση με τα MFCC, θεωρούμε πως η προτεινόμενη μέθοδος είναι ικανή να προσδιορίσει τα γνωρίσματα αυτά της μουσικής που διαφοροποιούν τις διαφορετικές κατηγορίες. Ακολουθώντας την προσέγγιση αυτή επιτυγχάνεται η δημιουργία ενός μουσικού λεξικού, το οποίο χρησιμοποιείται για την περιγραφή των διαφορετικών ειδών μουσικής, διατυπώνοντας έτσι μία εναλλακτική διαδικασία εξαγωγής αναπαραστάσεων των μουσικών σημάτων. Καταλήγοντας, με τη μέθοδο αυτή αντιμετωπίζουμε διάφορα προβλήματα πολυπλοκότητας κατά την κατηγοριοποίηση, λόγω των νέων συμπαγών αναπαραστάσεων.

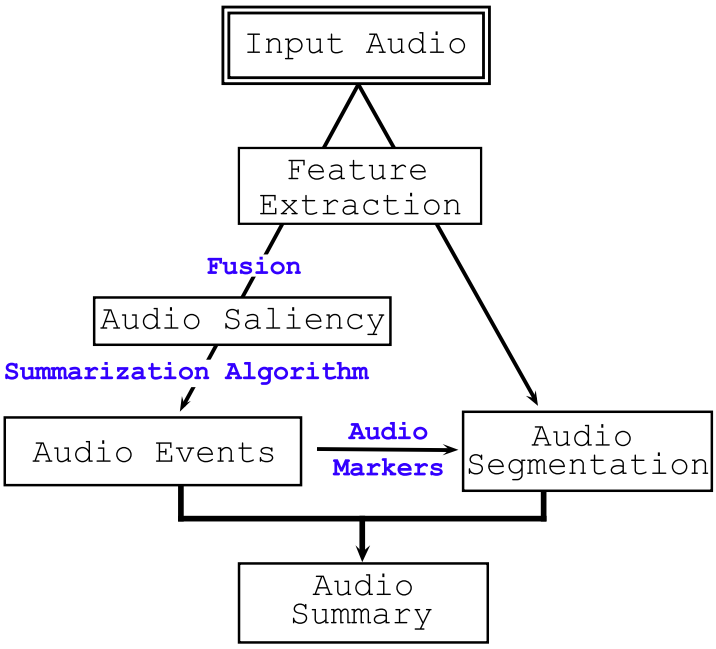
# Κεφάλαιο 5

# Ανίχνευση Σημαντικών Γεγονότων σε Πολυμεσικά Βίντεο με Έμφαση στην Ακουστική Πληροφορία

Στο κεφάλαιο αυτό μελετάμε την καταλληλότητα μη-γραμμικών μοντέλων σε θέματα ανίχνευσης σημαντικών μουσικών και γενικά ακουστικών γεγονότων, δηλαδή χρονικά οριοθετημένων τμημάτων τα οποία εμφανίζουν σημαντική δραστηριότητα ως προς την ανθρώπινη αντίληψη. Βασιζόμαστε στη χαμηλού επιπέδου πληροφορία (*bottom-up processing*) του σήματος, για να προσεγγίζουμε το πρόβλημα της ανάκτησης της ακουστικής πληροφορίας, χρησιμοποιώντας αναπαραστάσεις της ηχητικής κυματομορφής. Συγκεκριμένα, εξετάζουμε κατά πόσο το AM-FM μοντέλο [40, 104] μπορεί να αναπαραστήσει σημαντικά ακουστικά γεγονότα. Επιπλέον, αξιολογούμε διάφορα υπολογιστικά μοντέλα σύμμειξης, τα οποία παρέχουν ένα γενικό πλαίσιο για την ενσωμάτωση της πληροφορίας. Οι μέθοδοι σύμμειξης που μελετώνται αποσκοπούν στον ενδοτροπικό συνδυασμό των χαρακτηριστικών και στη δημιουργία μιας περιγραφικής καμπύλης σημαντικότητας (*saliency curve*) η οποία θα αποτελέσει το κριτήριο δημιουργίας της ηχητικής περίληψης.

## 5.1 Ανάλυση και Μοντελοποίηση Ηχητικών Σημάτων

Προσεγγίζουμε το θέμα του υπολογισμού της σημαντικότητας του ακουστικού σήματος ως πρόβλημα βασισμένο στην προσοχή. Η σημασία των αλλαγών του πλάτους και της συχνότητας έχει αποτελέσει κίνητρο πολλών ερευνητικών μελετών, όπου ζητούμενο είναι η μέτρηση της διαμόρφωσης της συχνότητας και της έντασης [50, 76, 96]. Η ένταση και οι διακυμάνσεις της συχνότητας έχουν επιπλέον αποδειχθεί σημαντικές για την ομαδοποίηση της ακουστικής πληροφορίας [18, 24] αλλά και για την αναγνώριση και την




*Σχήμα 5.1:* Σύνοψη του συστήματος ηχητικών περιλήψεων.

κατηγοριοποίηση των διαφορετικών πηγών ήχου και γεγονότων. Το AM-FM μοντέλο που χρησιμοποιούμε ποσοτικοποιεί τη σημαντικότητα μέσω ενός συνδυασμού των παραμέτρων διαμόρφωσης των μη στάσιμων συνιστωσών του σήματος [104]. Αυτό μας οδηγεί σε μια συμπαγή αναπαράσταση της ηχητικής ροής, παρακολουθώντας τις συνιστώσες με τη μέγιστη ενέργεια [40]. Το Σχήμα 5.1 δείχνει το συνολικό σύστημα από την εξαγωγή χαρακτηριστικών έως τη δημιουργία της περίληψης.

### 5.1.1 Εξαγωγή Χαρακτηριστικών Διαμόρφωσης

Η εξαγωγή ηχητικών χαρακτηριστικών αποτελεί το πρώτο βήμα (μέρος του front-end) σε εφαρμογές ανίχνευσης, κατάτμησης, αναγνώρισης αλλά και ταυτοποίησης του ήχου. Στόχος είναι η παραγωγή σύντομων περιγραφικών αναπαραστάσεων σε επίπεδο σήματος, οι οποίες χρησιμοποιούνται στα επόμενα βήματα της επεξεργασίας.

Στην περίπτωση αυτή η ανάλυση και η μοντελοποίηση της σημαντικότητας του ηχητικού σήματος βασίζεται στο μοντέλο AM-FM [104]:

$$[n] = \sum_{k=1}^{K} A_k[n] \cos\left(\int_0^n \Omega_k[n] dn\right). \tag{5.1}$$

Το στιγμιαίο πλάτος $A_k[n]$ και η συχνότητα $\Omega_k[n]$ εξάγονται με πολυκαναλική ανάλυση (*multiband*) $s[n]$ μέσω συστοιχίας $K$ φίλτρων Gabor $h_k$, την εφαρμογή του ενεργειακού τελεστή Teager $\Psi$ και του αλγορίθμου αποδιαμόρφωσης σε κάθε φίλτρο

εξόδου. Στη συνέχεια, υπολογίζεται το μέσο στιγμιαίο πλάτος $\mathrm{MIA}[m] = (\overline{|A_j[n]|})$ (***Mean multiband Instantaneous Amplitude, MIA***) και η μέση στιγμιαία συχνότητα $\mathrm{MIF}[m] = (\overline{\Omega_j[n]})$ (***Mean multiband Instantaneous Frequency, MIF***) από την κυρίαρχη ενέργεια διαμόρφωσης (*energy-dominant modulation component*) κατά μήκος των διαφορετικών ζωνών συχνοτήτων [40]. Με άλλα λόγια, από τη συνιστώσα $j = j[m]$ η οποία μεγιστοποιεί τη μέση τιμή της Teager ενέργειας $\mathrm{MTE}[m] = \arg\max_k(\overline{\Psi(s * h_k)})$ (***Mean multiband Teager Energy, MTE***), όπου $m$ ο δείκτης του πλαισίου και $(\overline{\cdots})$ η μέση τιμή. Λεπτομέρειες σχετικά με την εξαγωγή των χαρακτηριστικών καθώς και εφαρμογές μπορούν να βρεθούν στα [40–42].

Το ηχητικό σήμα περιγράφεται από ένα τρισδιάστατο διάνυσμα χαρακτηριστικών

$$\vec{F}_{\mathrm{a}}[m] = [\mathrm{MTE}, \mathrm{MIA}, \mathrm{MIF}]\,[m] \tag{5.2}$$

που εκφράζει πληροφορίες ως προς το επίπεδο διέγερσης, το συχνοτικό περιεχόμενο και την ενέργεια, οι οποίες συνδέονται με την παρουσία αλλά και την εξέλιξη των ηχητικών γεγονότων. Στη συνέχεια παρουσιάζουμε διάφορα υπολογιστικά μοντέλα σύμμειξης για τον συνδυασμό των MTE, MIA και MIF, τα οποία έχουν ως αποτέλεσμα τη δημιουργία μιας ενιαίας καμπύλης σημαντικότητας (*saliency curve*).

### 5.1.2 Υπολογιστικά Μοντέλα Ενδοτροπικής Σύμμειξης

Αξιολογήσαμε πειραματικά διάφορα μοντέλα σύμμειξης (*fusion*) των χαρακτηριστικών για τη δημιουργία της μονοδιάστατης καμπύλης σημαντικότητας, η οποία αποτελεί κριτήριο για την επιλογή των αντιληπτικά σημαντικών ηχητικών γεγονότων για τη δημιουργία ουσιαστικών περιλήψεων. Το πρόβλημα που εξετάζουμε είναι η χαμηλού επιπέδου σύμμειξη (*intramodal fusion*), όπου τα χαρακτηριστικά, αφού κανονικοποιηθούν και συνδυαστούν, δημιουργούν μια μονοτροπική (*monomodal*) καμπύλη σημαντικότητας. Κάθε τιμή της αποτελεί μέτρο της αντιληπτικής σημαντικότητας κάθε ξεχωριστού χαρακτηριστικού. Τα χαρακτηριστικά κανονικοποιούνται στο εύρος $[0, 1]$ για να αντισταθμιστούν διαφορές που παρατηρούνται στο δυναμικό τους εύρος, με τη χρήση της μεθόδου των ελαχίστων τετραγώνων. Η καμπύλη αυτή έχει πολλά προτερήματα: (α) είναι μια συνεχής συνάρτηση του χρόνου, κατάλληλα σχεδιασμένη στο εύρος τιμών $[0, 1]$, (β) βασίζεται σε μη επιβλεπόμενα μοντέλα προσοχής (*unsupervised, bottom-up*) και (γ) προσεγγίζει την προκαλούμενη από το ηχητικό σήμα προσοχή στον ακροατή.

Στη συνέχεια τα κανονικοποιημένα χαρακτηριστικά συνδυάζονται μεταξύ τους:

$$S_{\mathrm{A}} \;=\; \mathrm{fusion}(S_1, S_2, S_3). \tag{5.3}$$

Για τη σύμμειξη χαμηλού επιπέδου η διαδικασία συνδυασμού των χαρακτηριστικών μπορεί να είναι: (α) Σταθμισμένοι γραμμικοί συνδυασμοί με ίσα ή άνισα σταθερά βάρη. (β) Προσαρμοζόμενοι γραμμικοί συνδυασμοί, όπου τα βάρη είναι αντιστρόφως ανάλογα της αβεβαιότητας κάθε χαρακτηριστικού (*variance-based weights*). (γ) Μη-γραμμική συνδυασμοί (μέγιστο, ελάχιστο, σταθμισμένο ελάχιστο). (δ) Τέλος, είναι δυνατό τα βάρη να μεταβάλλονται δυναμικά ως προς τον χρόνο (*time-adaptive, dynamic weights*), με τη χρήση της σημασιολογικής δομής του βίντεο (π.χ., πλάνα και σκηνές). Τα μοντέλα που εξετάσαμε προκειμένου να βρεθεί το καλύτερο σχήμα σύμμειξης που θα χρησιμοποιηθεί για την τελική περίληψη είναι τα εξής:

(1) **Γραμμική Σύμμειξη:** Αποτελεί το πιο διαισθητικό και απλό σενάριο σύμμειξης αλλά και βασική μέθοδο, που έχει χρησιμοποιηθεί σε προγενέστερες εργασίες μας. Τα κανονικοποιημένα χαρακτηριστικά συνδυάζονται βάσει ενός σταθμισμένου (*weighted*) γραμμικού σχήματος, όπου τα βάρη μπορεί να είναι ίσα, σταθερά ή προσαρμοζόμενα:

$$S_{\text{lin}} = w_1 S_1 + w_2 S_2 + w_3 S_3. \tag{5.4}$$

Σε αυτή την περίπτωση χρησιμοποιούμε το πιο απλό σχήμα με ίσα βάρη $w_i = 1/3$ και για τα τρία διανύσματα χαρακτηριστικών (LE).

(2) **Προσαρμοζόμενη Γραμμική Σύμμειξη (*Variance-based*):** Κάθε χαρακτηριστικό σταθμίζεται αντιστρόφως ανάλογα με τη διακύμανσή του:

$$S_{\text{va}} = \sum_i (S_i/\text{var}(S_i))/\sum_i (1/\text{var}(S_i)). \tag{5.5}$$

Το συγκεκριμένο σχήμα (VA) μπορεί να εφαρμοστεί είτε συνολικά σε όλο το διάνυσμα χαρακτηριστικών είτε δυναμικά σε μικρότερα τμήματα (π.χ., σκηνές ή πλάνα).

(3) **Μη-γραμμική σύμμειξη:** (i) min (MI) και (ii) max (MA), επιλέγοντας την ελάχιστη ή τη μέγιστη τιμή των τριών χαρακτηριστικών σε κάθε καρέ του βίντεο,

$$S_{\text{min}} = \min\{S_1, S_2, S_3\}, \quad S_{\text{max}} = \max\{S_1, S_2, S_3\}. \tag{5.6}$$

(iii) Επιπλέον, εξετάσαμε το σταθμισμένο min σχήμα (MIVA), το οποίο μπορεί να εφαρμοστεί είτε στο συνολικό σήμα είτε στις επιμέρους σκηνές ή πλάνα. Σε αυτή την περίπτωση κάθε διάνυσμα χαρακτηριστικών σταθμίζεται προσθετικά αντιστρόφως ανάλογα με τη $\log$ διακύμανση:

$$S_{\text{miva}} = \min(S_{\text{i}} - w_i) + \max(w_i) \tag{5.7}$$

$$S_{\text{miva}} = \min(S_1 - w_1, S_2 - w_2, S_3 - w_3) + \max(w_1, w_2, w_3), \tag{5.8}$$

όπου $w_i = \log(1/\text{var}(S_i))$. Η μέθοδος αυτή είναι αλγεβρικά ομομορφική (*homomorphic*) με την προσαρμοζόμενη γραμμική σύμμειξη (5.5).

Επιπρόσθετα, εξετάσαμε τρία διαφορετικά διαστήματα κανονικοποίησης των χαρακτηριστικών που συνοψίζονται στη συνέχεια:

(1) Γραμμική κανονικοποίηση στο συνολικό ηχητικό σήμα (GL).

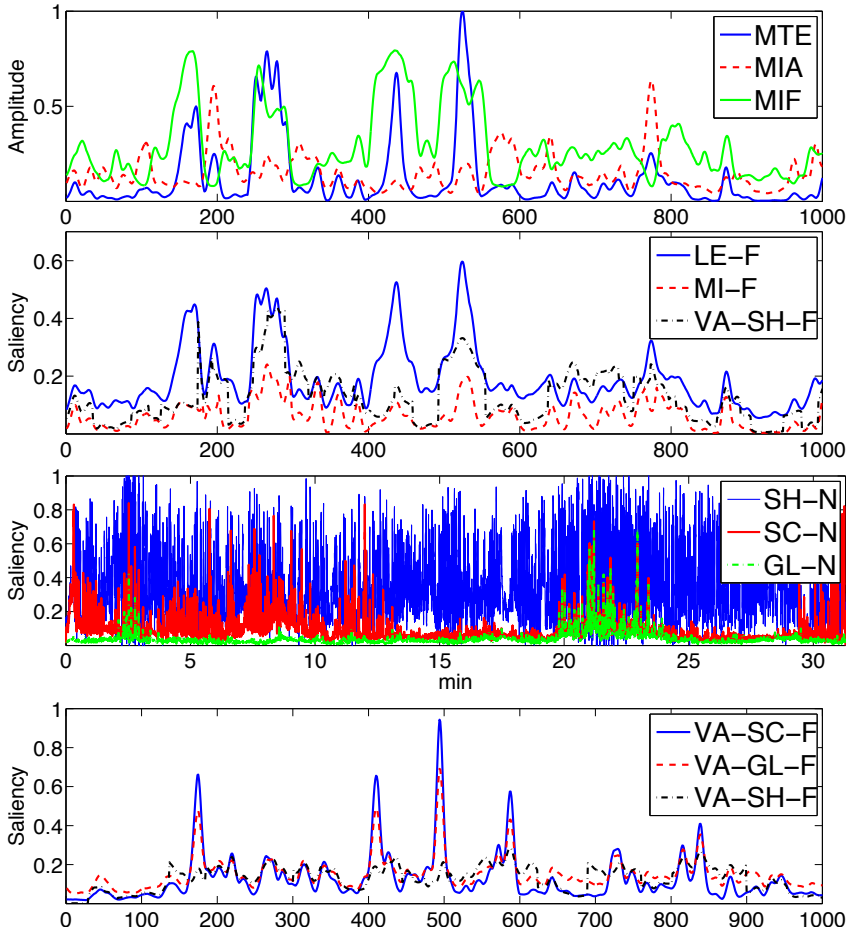
(2) Γραμμική κανονικοποίηση σε επιμέρους τμήματα του σήματος, όπου τα όρια ορίζονται από τις σκηνές (SC).

(3) Γραμμική κανονικοποίηση όπου τα όρια του σήματος ορίζονται από τα πλάνα (SH).

Τέλος, διερευνήσαμε τις δυνατότητες της δυναμικής αναπροσαρμογής των βαρών (*dynamic adaptation*), όπου δηλαδή τα βάρη αναπροσαρμόζονται χρονικά βάσει διαφορετικών χρονικών παραθύρων. Για παράδειγμα, για τη variance-based και τη μη-γραμμική σταθμισμένη min σύμμειξη τα βάρη μπορούν να υπολογιστούν είτε σε όλη τη χρονική διάρκεια του διανύσματος χαρακτηριστικών (VA-GL) είτε ανά πλάνο (VA-SH) ή ανά σκηνή (VA-SC). Στο Σχήμα 5.2 (από πάνω προς τα κάτω) βλέπουμε τα χαρακτηριστικά, τρεις ενδεικτικές καμπύλες σημαντικότητας, τη σύμμειξη των χαρακτηριστικών στα τρία διαφορετικά διαστήματα κανονικοποίησης και τα τρία διαφορετικά διαστήματα δυναμικής αναπροσαρμογής των βαρών.

### 5.1.3 Αλγόριθμος Δημιουργίας Περιλήψεων για την Ανίχνευση Ηχητικών Γεγονότων

Ο αλγόριθμος δημιουργίας περιλήψεων βασίζεται σε προηγούμενη ερευνητική μας εργασία (βλ. [42]) και ακολουθεί τα εξής βήματα: (i) φιλτράρισμα της καμπύλης σημαντικότητας με ένα median φίλτρο μήκους $2M + 1$ καρέ. (ii) Επιλογή του κατωφλιού σημαντικότητας (*saliency threshold*) $S_c$ το οποίο εξαρτάται από το ποσοστό της περίληψης $c$ και ορίζεται από τον χρήστη. Τα καρέ $m$ με τιμή σημαντικότητας μεγαλύτερη του κατωφλιού αυτού $S_{\text{A}}[m] > S_c$ επιλέγονται για να συμπεριληφθούν στην τελική περίληψη. Για παράδειγμα, για μια περίληψη που αντιστοιχεί στο 20% της συνολικής διάρκειας του βίντεο ($c = 0.2$), το κατώφλι $S_c$ επιλέγεται έτσι ώστε ο αριθμός (*cardinality*) των επιλεγμένων καρέ $D = \{m : S_{\text{A}}[m] > S_c\}$ να αποτελεί το 20% του συνολικού. Αποτέλεσμα αυτού είναι μια συνάρτηση $I_c$ για το επιθυμητό επίπεδο περίληψης $c$. (iii) Συνδυασμός των επιλεγμένων ακολουθιών καρέ σε τμήματα. Οι ακολουθίες των καρέ που είναι μικρότερες από έναν επιλεγμένο από τον χρήστη αριθμό $N$ διαγράφονται από τη περίληψη.




*Σχήμα 5.2:* Από πάνω προς τα κάτω: Χαρακτηριστικά MTE, MIA και MIF. Σύμμειξη των χαρακτηριστικών. Σύμμειξη των χαρακτηριστικών στα τρία διαφορετικά διαστήματα κανονικοποίησης. Σύμμειξη των χαρακτηριστικών με δυναμική αναπροσαρμογή των βαρών στα τρία διαφορετικά διαστήματα. Τα σχήματα παρουσιάζονται για 1000 καρέ, εκτός από το τρίτο όπου δείχνουμε την κανονικοποίηση για ολόκληρο το τριαντάλεπτο ηχητικό σήμα.

(iv) Ένωση των επιλεγμένων για την περίληψη γειτονικών τμημάτων, αν τα χωρίζουν λιγότερα από $K$ καρέ. (v) Συνένωση των τελικών τμημάτων για τη δημιουργία της περίληψης με την τεχνική overlap-add σε $L$ βίντεο καρέ. Για τις ανάγκες της συγκεκριμένης εφαρμογής και ύστερα από πειραματισμό για την εύρεση των καλύτερων εμπειρικά τιμών καταλήξαμε στις τιμές $M = N = 30$ και $K = L = 15$ καρέ για δεδομένα βίντεο στα 25 fps.

## 5.2 Ποσοτική Αξιολόγηση των Υπολογιστικών Μεθόδων Σύμμειξης

Αξιολογήσαμε τρεις διαφορετικές μεθόδους κανονικοποίησης των χαρακτηριστικών: στο συνολικό επίπεδο (GL), στο επίπεδο σκηνής (SC) και στο επίπεδο του πλάνου (SH)

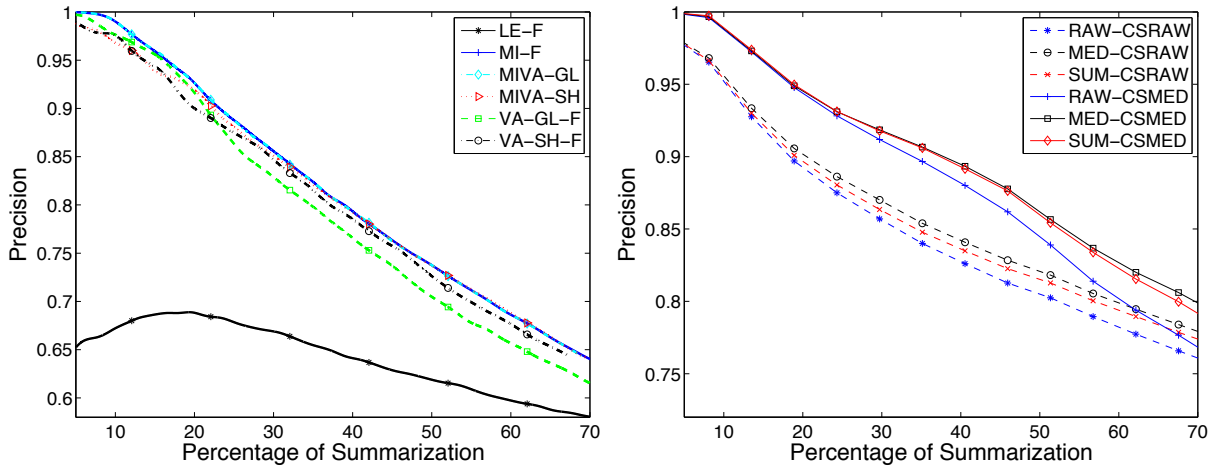
καθώς και εννέα μεθόδους σύμμειξης: γραμμική (LE), min (MI), max (MA), σταθμισμένη min σε τρία διαφορετικά δυναμικά επίπεδα (MIVA-GL, MIVA-SC, MIVA-SH), και το γραμμικό προσαρμοζόμενο σχήμα, όπου τα βάρη είναι αντιστρόφως ανάλογα με τη διακύμανση σε τρία διαφορετικά δυναμικά επίπεδα (VA-GL, VA-SC, SH-VA). Για τον σκοπό αυτό, χρησιμοποιήσαμε ηχητικά δεδομένα προερχόμενα από τη βάση δεδομένων *MovieSum* (βλ. Ενότητα 5.5), μια κοινή εργασία του ΕΜΠ και του Πολυτεχνείου Κρήτης, η οποία περιλαμβάνει μισάωρα τμήματα από τις ακόλουθες βραβευμένες με Όσκαρ ταινίες: *Σικάγο, Crash, Ο Πληροφοριοδότης, Ο Μονομάχος, Ο Άρχοντας των Δαχτυλιδιών - Η Επιστροφή του Βασιλιά* και *Ψάχνοντας τον Νέμο*. Πρόθεσή μας είναι να εξετάσουμε με συστηματικό τρόπο την απόδοση των διαφορετικών μεθόδων σύμμειξης στη δημιουργία περιλήψεων. Ιδεατά οι περιλήψεις αυτές βρίσκονται σε συμφωνία με τις επιλογές των χρηστών/επισημειωτών ως προς τα σημαντικά ακουστικά γεγονότα. Τα επισημειωμένα τμήματα της βάσης σχηματίζουν μια δυαδική συνάρτηση δείκτη, που μας δείχνει την ύπαρξη των ακουστικά σημαντικών καρέ του βίντεο.

Ο Πίνακας 5.1 δείχνει αποτελέσματα precision (όπου ως precision ορίζονται τα σωστά επιλεγμένα καρέ/του συνόλου των καρέ). Θεωρούμε πως το precision χαρακτηρίζει με τον καλύτερο τρόπο την απόδοση του συστήματος σε επίπεδο καρέ. Τα αποτελέσματα παρουσιάζονται για περιλήψεις που περιλαμβάνουν το 20%, το 33% και το 50% του αρχικού αριθμού των καρέ του μισάωρου βίντεο για όλους τους πιθανούς συνδυασμούς σύμμειξης και κανονικοποίησης (τα καλύτερα αποτελέσματα εμφανίζονται με έντονους χαρακτήρες). Παρατηρούμε πως για όλες τις μεθόδους αξιολόγησης η κανονικοποίηση στο συνολικό επίπεδο του σήματος GL υπερτερεί της κανονικοποίησης σε επίπεδο σκηνής και πλάνου. Για την κανονικοποίηση GL βλέπουμε πως (α) οι μη-γραμμικές MI-F και MIVA-F μέθοδοι υπερτερούν της γραμμικής και max (MA-F), ενώ (β) οι προσαρμοζόμενοι γραμμικοί συνδυασμοί και των τριών δυναμικών επιπέδων (GL-VA, VA-SC, VA-SH) ξεπερνούν τη γραμμική LE-F και τη MA-F σύμμειξη.

Το Σχήμα 5.3 (α) δείχνει τα αποτελέσματα του precision ως συνάρτηση του ποσοστού περίληψης (που κυμαίνεται από 5% έως 70%), για την GL κανονικοποίηση και τις πέντε καλύτερες μεθόδους σύμμειξης μαζί με τη γραμμική, η οποία παρατίθεται για λόγους αναφοράς. Παρατηρούμε πως η μέθοδος MI έχει εξίσου καλά αποτελέσματα με τη MIVA-GL, ενώ η MIVA-SH επιτυγχάνει καλά αποτελέσματα για περιλήψεις που αποτελούνται από περισσότερα από 40% των καρέ. Στη συνέχεια ακολουθούν οι VA-SH και VA-GL, ενώ η γραμμική μέθοδος LE-F, η οποία αποτελεί και μέθοδο αναφοράς, παρουσιάζει πολύ χαμηλότερα αποτελέσματα σε σύγκριση με τις υπόλοιπες μεθόδους σύμμειξης.

*Πίνακας 5.1:* Αποτελέσματα Precision για την αξιολόγηση των μεθόδων σύμμειξης. Τα χαρακτηριστικά αξιολογούνται βάσει της επισημείωσης σημαντικότητας στην ηχητική ροή του βίντεο. Λεπτομέρειες για τη βάση και τον τρόπο επισημείωσης βλ. Εν. 5.5.

| Features | | Audio Feature Fusion | | |
|---|---|---|---|---|
| Evaluated on: | | Audio (A) Labeling | | |
| | | Summarization Percent | | |
| Algorithm | | 20% | 33% | 50% |
| Norm | Fusion | Precision Scores | | |
| GL-N | LE-F | 68.8 | 66.1 | 61.9 |
| GL-N | MA-F | 48.8 | 51.2 | 52.6 |
| GL-N | MI-F | **92.6** | **83.6** | **73.8** |
| GL-N | MIVA-GL-F | **92.6** | **83.6** | **73.8** |
| GL-N | MIVA-SC-F | 91.1 | 81.9 | 72.8 |
| GL-N | MIVA-SH-F | 91.9 | 83.4 | 73.7 |
| GL-N | VA-GL-F | 91.6 | 81.0 | 70.5 |
| GL-N | VA-SC-F | 85.3 | 75.8 | 68.2 |
| GL-N | VA-SH-F | 90.0 | 82.8 | 72.6 |
| SC-N | LE-F | 66.1 | 64.3 | 62.0 |
| SC-N | MI-F | 77.8 | 73.2 | 69.1 |
| SC-N | MIVA-GL-F | 78.0 | 73.3 | 68.9 |
| SC-N | MIVA-SC-F | 77.6 | 72.3 | 67.6 |
| SC-N | VA-GL-F | 72.6 | 68.3 | 63.7 |
| SC-N | VA-SC-F | 72.6 | 65.4 | 61.6 |
| SH-N | LE-F | 73.2 | 68.8 | 64.2 |
| SH-N | MI-F | 68.9 | 67.6 | 64.7 |
| SH-N | MIVA-GL-F | 66.9 | 66.2 | 63.5 |
| SH-N | MIVA-SC-F | 68.4 | 66.9 | 64.4 |
| SH-N | MIVA-SH-F | 66.9 | 66.0 | 63.6 |
| SH-N | VA-GL-F | 73.2 | 68.9 | 64.2 |
| SH-N | VA-SC-F | 73.4 | 69.3 | 64.7 |




(α′) (β′)

*Σχήμα 5.3:* Αποτελέσματα Precision: (α′) για τις πέντε καλύτερες μεθόδους σύμμειξης και την baseline μεθόδο LE-F και (β′) για τον αλγόριθμο Κοντινότερων Γειτόνων.

## 5.3 Αξιολόγηση με Τεχνική Μηχανικής Μάθησης

Στη συνέχεια χρησιμοποιούμε έναν αλγόριθμο μηχανικής μάθησης για την εκπαίδευση των ακουστικών χαρακτηριστικών που παρουσιάστηκαν στην Ενότητα 5.1.1 [200]. Σκοπός μας είναι η αξιολόγηση της απόδοσης των προτεινόμενων μοντέλων σύμμειξης. Συγκεκριμένα, χρησιμοποιούμε το διάνυσμα χαρακτηριστικών $\vec{F}_{\mathrm{a}}[m] = [\mathrm{MTE}, \mathrm{MIA}, \mathrm{MIF}]\,[m]$ μαζί με την πρώτη και τη δεύτερη παράγωγό τους (υπολογισμένη σε τρία και πέντε πλαίσια αντίστοιχα). Ο ταξινομητής, ο οποίος βασίζεται στη μέθοδο $K$-Κοντινότερων Γειτόνων (*$K$-Nearest Neighbors, NNR-$k$*), εκπαιδεύεται στα επισημειωμένα γεγονότα χρησιμοποιώντας τη δυαδική συνάρτηση (όπου 1 καρέ με σημαντικά ηχητικά γεγονότα, 0 αλλού). Η έξοδος του ταξινομητή είναι μια δυαδική συνάρτηση μηδενικών (κανένα γεγονός) και μονάδων (ύπαρξη ηχητικού γεγονότος). Χρησιμοποιούμε 6 cross-validation, όπου τα μοντέλα NNR-$k$ εκπαιδεύονται σε πέντε ταινίες και δοκιμάζονται στην έκτη. Για να βελτιώσουμε την απόδοση των μοντέλων (και άρα να επιλέξουμε καρέ με μεγαλύτερη πιθανότητα αντιστοίχισης σε σημαντικό ακουστικό γεγονός) χρησιμοποιούμε τις παρακάτω παραμέτρους ομαλότητας (*smoothing*): (i) Καμία ομαλότητα, (RAW), (ii) Median-φιλτράρισμα της εξόδου του ταξινομητή με παράθυρο μήκους $2M+1$ (MED), (iii) Εφαρμογή του αλγορίθμου περίληψης που παρουσιάστηκε στην Ενότητα. 5.1.3, για να συμπεριλάβουμε την περίπτωση της κατωφλιόμενης καμπύλης σημαντικότητας (SUM).

Για να ληφθούν αποτελέσματα για μεταβλητά ποσοστά συμπίεσης καθορίζουμε ένα ποσοστό βαρύτητας (*confidence score*) σε κάθε αποτέλεσμα της ταξινόμησης, δηλαδή, σε κάθε καρέ. Επιλέγουμε ως ποσοστό βαρύτητας το τμήμα των $k$ κοντινότερων γειτόνων που έχουν την τιμή 1 (ηχητικά γεγονότα) και το εφαρμόζουμε τόσο στην αρχική έξοδο (CSRAW) όσο και στη median φιλτραρισμένη έξοδο (CSMED). Στο Σχήμα 5.3 (β) παρουσιάζουμε precision αποτελέσματα για όλους τους συνδυασμούς της εξόδου του ταξινομητή με τα ποσοστά βαρύτητας (των CSRAW και CSMED), π.χ., το «SUM-CSMED» αναφέρεται στη χρήση του αλγορίθμου περίληψης για ομαλοποίηση της εξόδου του ταξινομητή μετά το median φιλτράρισμα και τον καθορισμό του ποσοστού βαρύτητας. Τα αποτελέσματα έχουν βελτιστοποιηθεί ώστε να επιτευχθεί η καλύτερη δυνατή ακρίβεια ταξινόμησης για $k = 250$ γείτονες και $M = K = 50$ καρέ για το median φιλτράρισμα.

Συνολικά διαπιστώνουμε πως επιτυγχάνονται καλύτερα αποτελέσματα σε σχέση με τις μεθόδους σύμμειξης, με εξαίρεση την περιοχή περιλήψεων $5 - 20\%$, όπου μόνο η μέθοδος CSMED αποδίδει καλύτερα. Το φιλτράρισμα της εξόδου του ταξινομητή (MED, SUM) βελτιώνει κάπως την ακρίβεια σε σχέση με τα RAW αποτελέσματα για το ποσοστό περίληψης 30-70%. Τέλος, η ομαλοποίηση, αφού εφαρμοστούν τα ποσοστά βαρύτητας στις median φιλτραρισμένες εξόδους (CSMED), βελτιώνει σημαντικά την ακρίβεια στην περιοχή 5-50% σε σχέση με τα CSRAW αποτελέσματα.

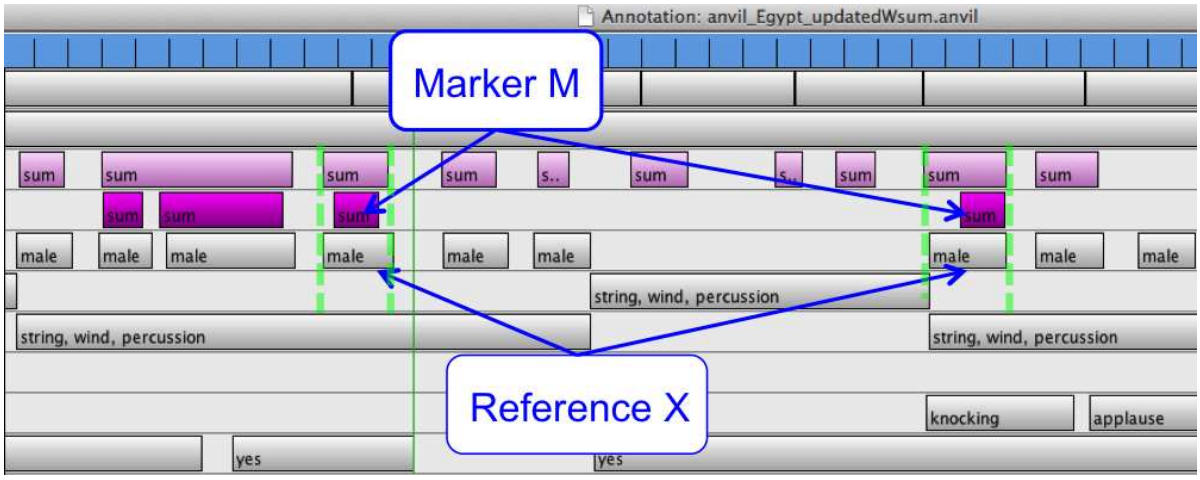
*Πίνακας 5.2:* Ποσοστά επί τοις % των τμημάτων που επιλέχθηκαν από τον αλγόριθμο δημιουργίας περίληψης και ανήκουν σε συγκεκριμένες κατηγορίες ήχου, για τη σύμμειξη των χαρακτηριστικών με το καλύτερο σχήμα MI-F και το baseline σχήμα LE-F, κανονικοποιημένα στο συνολικό διάνυσμα χαρακτηριστικών (GL).

| Type of Audio | Movie (3 min) | | | | | Music Documentary (4 min) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Algorithm | MI-F | | LE-F | | % frames for each audio category | MI-F | | LE-F | | % frames for each audio category |
| Summarization Percent | 20% | 50% | 20% | 50% | | 20% | 50% | 20% | 50% | |
| | % of detected frames | | | | | % of detected frames | | | | |
| Audio Category | | | | | | | | | | |
| Speech | 28.0 | 82.0 | 18.8 | 70.9 | 36.6 | 0 | 25.9 | 1.5 | 29.2 | 61.0 |
| Music | 20.4 | 55.5 | 23.3 | 43.0 | 17.5 | 58.0 | 88.1 | 50.5 | 80.9 | 35.0 |
| Background Music | 36.8 | 68.6 | 27.2 | 51.2 | 35.9 | 0 | 25.1 | 1.3 | 28.8 | 52.1 |
| Song | - | - | - | - | - | 83.1 | 100 | 76.87 | 98.6 | 19.8 |
| Environmental | 38.6 | 52.5 | 71.9 | 95.0 | 8.3 | - | - | - | - | - |
| Machine | 33.5 | 81.2 | 23.4 | 62.9 | 4.6 | 50 | 100 | 59.5 | 59.5 | 0.7 |
| Foley | 18.0 | 47.6 | 17.5 | 41.8 | 10.0 | 0 | 10.6 | 19.5 | 100 | 1.9 |
| Impact | 100 | 100 | 86.1 | 100 | 2.3 | 0 | 100 | 0 | 38.5 | 0.2 |

## 5.4 Ανάλυση της Δομής των Γεγονότων και Δημιουργία Τελικής Περίληψης

Σε αυτή την ενότητα αναλύουμε και αξιολογούμε τα επιλεγμένα τμήματα ως προς τη δομή τους και την κατηγορία ήχου στην οποία ανήκουν, και προτείνουμε μια μέθοδο για τη διόρθωση των ορίων τους. Η αξιολόγηση γίνεται σε βάση δεδομένων που αποτελείται από τέσσερα μικρά βίντεο (3–4 λεπτά), τα οποία είναι επισημειωμένα τόσο ως προς τη σημαντικότητα όσο και ως προς την κατηγορία ήχου. Τα βίντεο αυτά ανήκουν στις εξής κατηγορίες: ντοκιμαντέρ, μουσικό ντοκιμαντέρ, σειρά δράσης και ταινία με έντονο μουσικό περιεχόμενο, και περιλαμβάνουν πληθώρα διαφορετικών ακουστικών γεγονότων. Η επισημείωση βασίζεται στις εξής ακουστικές κατηγορίες και υποκατηγορίες: ομιλία, μουσική, μουσικό φόντο, τραγούδι, περιβαλλοντικοί ήχοι (π.χ., άνεμος, κύματα, βροχή κ.λπ.), ήχοι μηχανών, ήχοι προερχόμενοι από τον άνθρωπο αλλά διαφορετικοί της ομιλίας (π.χ., γέλιο, χειροκροτήματα, βηματισμός) και διάφορα είδη ηχητικών εφέ (όπως εκρήξεις, πυροβολισμοί, χτυπήματα κ.ά.).

Στην παράγραφο που ακολουθεί γίνεται ανάλυση των κατηγοριών των αυτόματα επιλεγμένων τμημάτων. Αξιολογούμε την καλύτερη μέθοδο σύμμειξης βάσει της προηγούμενης ανάλυσης MI-F (GL-N) σε σύγκριση με τη γραμμική σύμμειξη LE-F (GL-N). Σημειώνουμε ότι το σχήμα MI-F περιλαμβάνει σχεδόν όλα τα τμήματα ομιλίας στις μεγαλύτερες περιλήψεις (50%), αλλά μόνο τα πιο σημαντικά κατά τη γνώμη μας υψηλής έντασης τμήματα ομιλίας για τις μικρότερες περιλήψεις (20%). Επιπλέον, περιλαμβάνει τμήματα με έντονη και δυνατή μουσική (τα οποία δεν αποτελούν μουσική υπόκρουση), όλα τα ηχητικά εφέ, ήχους μηχανών αλλά και άλλους που ξεχωρίζουν σε τμήματα σιωπής. Στο μουσικό ντοκιμαντέρ παρατηρούμε ότι εξάγονται περισσότερα τμήματα μουσικής, πιθανότατα λόγω της υψηλής τους έντασης σε σύγκριση με την ομιλία (τμήματα




*Σχήμα 5.4:* Η συνεκτική συνιστώσα «φωνής» $X$ και ο σημαδευτής $M$ για τον αλγόριθμο reconstruction opening.

συνέντευξης). Ο Πίνακας 5.2 δείχνει το ποσοστό επί τοις % των εξαχθέντων καρέ για κάθε ηχητική κατηγορία. Το συνολικό ποσοστό των καρέ για κάθε κατηγορία ήχου παρουσιάζεται στην στήλη «% *of frames for each audio category*».

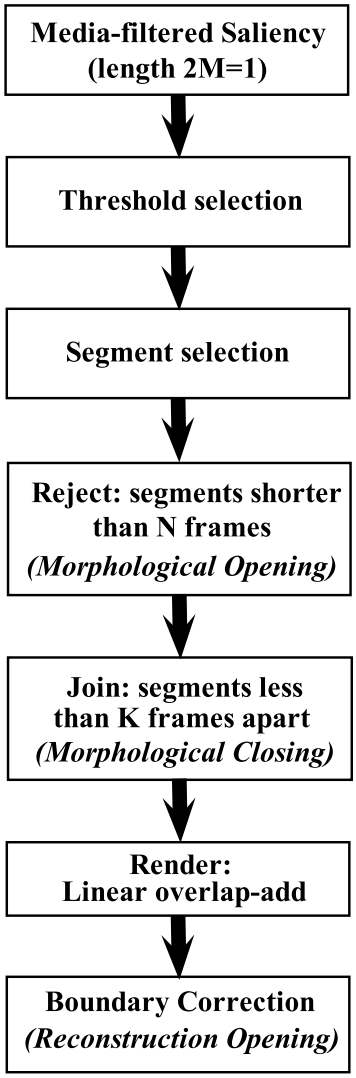
**Τελική Περίληψη**

Τόσο σε περιλήψεις πολυμεσικών δεδομένων όσο και σε περιλήψεις ηχητικών δεδομένων είναι μείζονος σημασίας τα περιλαμβανόμενα τμήματα εκτός από σημαντικά να είναι και σημασιολογικά σωστά. Ο έλεγχος των αυτόματων περιλήψεων με υποκειμενικά πειράματα ο οποίος έχει διεξαχθεί για τις αυτόματες περιλήψεις πολυμεσικών δεδομένων (περιλήψεις βίντεο) [42], (βλ. επίσης Εν. 5.5) έδειξε πως ο θεατής δίνει μεγάλη βαρύτητα στο ηχητικό αποτέλεσμα. Άρα οι περιλαμβανόμενοι ήχοι, και ειδικά η ομιλία, πρέπει να αποτελούν ολοκληρωμένες φράσεις, διαφορετικά ο ακροατής/θεατής αξιολογεί την περίληψη ως αδύναμη σε σχέση με την αντιληπτική της ποιότητα και αισθητική.

Γι' αυτό τον λόγο, προτείνουμε αλγόριθμο διόρθωσης των ορίων των αυτόματα επιλεγμένων τμημάτων περίληψης, ο οποίος βασίζεται σε ιδέες της μαθηματικής μορφολογίας και συγκεκριμένα στο reconstruction opening: $\rho^{-}(M|X) \triangleq$ η συνεκτική συνιστώσα του $X$ που περιέχει τον σημαδευτή $M$.

Χρησιμοποιούμε την κατωφλιόμενη καμπύλη σημαντικότητας (*thresholded saliency curve*) (βλ. Εν. 5.1.3) ως δείκτη $M$ που σηματοδοτεί τα πιο σημαντικά ηχητικά τμήματα για να εξαγάγουμε μεγαλύτερης κλίμακας τμήματα, γνωρίζοντας μόνο αυτούς τους μικρότερους δείκτες στο εσωτερικό τους. Τα μεγαλύτερης κλίμακας τμήματα αναφοράς $X$ ορίζονται από την επισημείωση της βάσης. Το Σχήμα 5.4 δείχνει τη συνεκτική συνιστώσα «φωνής» $X$ και τον σημαδευτή $M$.

Σημειώνουμε πως για ηχητικά τμήματα που περιλαμβάνουν ομιλία, η ίδια διαδικασία




*Σχήμα 5.5:* Αλγόριθμος δημιουργίας ηχητικών περιλήψεων μετά τη σύμμειξη.

θα μπορούσε να γίνει αυτόματα με κάποιο αλγόριθμο αυτόματης κατάτμησης VAD [40]. Η διόρθωση αυτή των ορίων αναμένεται να βελτιώσει την ακρίβεια της μεθόδου σε σχέση με το επισημειωμένο ground-truth, δεδομένου ότι οι επισημειωτές τείνουν να επιλέγουν ενιαία τμήματα ιδίως όσον αφορά την ομιλία. Το Σχήμα 5.5 δείχνει τον συνολικό αλγόριθμο δημιουργίας περιλήψεων (μετά τη σύμμειξη των χαρακτηριστικών) ξεκινώντας από το median φιλτράρισμα και καταλήγοντας στη διόρθωση των ορίων των επιλεγμένων από τον αλγόριθμο τμημάτων.

**Συμπεράσματα**

Σε αυτό το κεφάλαιο προτείναμε γραμμικές και μη-γραμμικές μεθόδους σύμμειξης για τη δημιουργία μιας μονοτροπικής καμπύλης σημαντικότητας, με εφαρμογή στην ανίχνευση ακουστικών γεγονότων. Ανάμεσα στις διάφορες μεθόδους κανονικοποίησης

και σύμμειξης που διερευνήθηκαν, είδαμε πως η κανονικοποίηση στο συνολικό επίπεδο σήματος (GL) και η μη-γραμμική min (MI), η σταθμισμένη min (MIVA) καθώς και η shot-variance (VA-SH) σύμμειξη λειτουργούν ικανοποιητικά, με αποτέλεσμα περιλήψεις αποτελούμενες από σημαντικά γεγονότα επιλεγμένα από τους επισημειωτές. Η αξιολόγηση της MI-F σε σχέση με το είδος των ηχητικών κατηγοριών έδειξε πως είναι κατάλληλη τόσο για γενικά ηχητικά σήματα όσο και για σήματα με έντονο το μουσικό περιεχόμενο. Τέλος, προτείναμε μέθοδο για τη διόρθωση των ορίων των επιλεγμένων τμημάτων, κάτι που έχει ως αποτέλεσμα αντιληπτικά ποιοτικές περιλήψεις.

## 5.5 Βάση Δεδομένων και Ανίχνευση Σημαντικών Γεγονότων σε Πολυμεσικά Δεδομένα

Κατά τη διάρκεια της διδακτορικής αυτής διατριβής ασχοληθήκαμε επίσης με το θέμα της ανίχνευσης σημαντικών γεγονότων σε πολυμεσικά δεδομένα και συγκεκριμένα σε ταινίες, για τη δημιουργία περιλήψεων. Κίνητρο μας η συνεχής αύξηση του ενδιαφέροντος για την ψηφιακή επεξεργασία ακουστικών σημάτων τα οποία και αποτελούν βασικό μέρος των πολυμεσικών δεδομένων. Το συγκεκριμένο θέμα είναι ένα συνεχιζόμενο πεδίο μελέτης της ερευνητικής μας ομάδας, ενώ στην έρευνα αυτή ασχοληθήκαμε συγκεκριμένα με το κομμάτι του ήχου.

Οι κατευθύνσεις με τις οποίες ασχοληθήκαμε είναι οι εξής:

- Δημιουργία βάσης δεδομένων από ταινίες «***MovieSum Database***», η οποία επισημειώνεται με βάση τα εξής κριτήρια: τη μονοτροπική και την πολυτροπική σημαντικότητα ή προκαλούμενη προσοχή (*monomodal & multimodal saliency*) του βίντεο, το συναίσθημα (προκαλούμενο ή επιδιωκόμενο), το σημασιολογικό περιεχόμενο και τη σημασιολογία σχετικά με τη δομή του (κατάτμηση σκηνών και πλάνων).

- Διεξοδικές ποσοτικές αξιολογήσεις του αλγορίθμου δημιουργίας περιλήψεων, όπου ως «ground-true» χρησιμοποιείται η επισημείωση της βάσης ταινιών.

- Διεξοδική αξιολόγηση και έλεγχος της αποδοτικότητας των αυτόματων περιλήψεων με υποκειμενικά πειράματα από ανθρώπους.

- Διερεύνηση καινούριων μεθόδων σύμμειξης των εξαχθέντων χαρακτηριστικών.

- Διερεύνηση και αξιολόγηση των χαρακτηριστικών, των μεθόδων σύμμειξης και των αυτόματα εξαχθέντων ακουστικών γεγονότων στα πλαίσια της εργασίας μας όπως περιγράφηκε παραπάνω.

**Βάση Δεδομένων: *MovieSum Database***

Αλγόριθμοι ανίχνευσης γεγονότων και παραγωγής περιλήψεων μπορούν να βελτιωθούν σημαντικά όταν υπάρχουν οι κατάλληλες συλλογές δεδομένων για την εκπαίδευση, την προσαρμογή και την αξιολόγηση των παραμέτρων τους. Η ανάγκη καθώς και η έλλειψη αξιόπιστης βάσης επισημειωμένης με οπτικοακουστικά γεγονότα μάς ώθησε να δημιουργήσουμε τη βάση ταινιών «*MovieSum*». Σκοπός της βάσης είναι να χρησιμοποιήσουμε γεγονότα επισημειωμένα από χρήστες καθώς και περιλήψεις ταινιών από χρήστες, για την αξιολόγηση και την προσαρμογή των αλγορίθμων ανίχνευσης γεγονότων και παραγωγής περιλήψεων.

**Συλλογή Δεδομένων:** Η διαδικασία δημιουργίας της βάσης αυτής περιλάμβανε τη συλλογή δεδομένων, τη μετατροπή σε κατάλληλο format και, το βασικότερο, τη διαδικασία επισημείωσης. Συγκεκριμένα, η βάση ταινιών αποτελείται από συνεχόμενα μισάωρα τμήματα 8 ταινιών (4 ώρες συνολικά). Και οι 8 ταινίες έχουν κερδίσει κινηματογραφικό βραβείο «Όσκαρ» Καλύτερης Ταινίας (*Academy Award for Best Motion Picture*) τις χρονιές 1998–2007 και ανήκουν σε διαφορετικά κινηματογραφικά είδη (π.χ., ταινίες δράσης, κωμωδίες, δραματικές, επικές, κινουμένων σχεδίων κ.ά.). Οι τίτλοι των ταινιών αυτών είναι: *Σικάγο, Crash, Ο Πληροφοριοδότης, Ο Μονομάχος, Ο Άρχοντας των Δαχτυλιδιών - Η Επιστροφή του Βασιλιά*, και η ταινία κινουμένων σχεδίων *Ψάχνοντας του Νέμο*[1].

Τα τμήματα των ταινιών της βάσης, τα οποία έχουν ληφθεί από την επίσημη, εμπορική διανομή DVD έχουν επιλεχθεί ως συνεχόμενα κομμάτια, διάρκειας μισής ώρας περίπου, τα οποία συμπεριλαμβάνουν ολόκληρη την τελευταία σκηνή και πλάνο. Τα βίντεο έχουν κωδικοποιηθεί και είναι διαθέσιμα σε μορφή avi (Xvid κωδικοποίηση) με συγκεκριμένες τεχνικές προδιαγραφές, σε δύο επίπεδα ανάλυσης: υψηλή, για την οπτικοποίηση περιλήψεων και αποτελεσμάτων, και χαμηλή για επεξεργασία και επισημείωση. Η βάση περιλαμβάνει επιπλέον ολόκληρες τις ταινίες αλλά και τα αρχεία υποτίτλων τους στη γλώσσα πηγής. Περίληψη των προδιαγραφών και των περιεχομένων της βάσης ταινιών μπορεί να βρεθεί στον Πίνακα 5.3.

Οι συγκεκριμένες ταινίες επιλέχθηκαν αφενός λόγω της αντικειμενικής δημοτικότητάς τους και της ποικιλίας που παρουσιάζουν ως κινηματογραφικά είδη, σκηνοθετική άποψη και έτος παραγωγής, αφετέρου λόγω ποιοτικών στοιχείων της πλοκής τους, δημιουργώντας έτσι μια βάση συστηματική και ανεξάρτητη από το είδος. Χαρακτηριστικά στοιχεία των ταινιών αυτών, όπως οι βασικοί χαρακτήρες (οι οποίοι αντιμετωπίζουν ανυπέρβλητες δοκιμασίες καθώς προσπαθούν να εκπληρώσουν τον στόχο τους), η αλληλουχία των γεγονότων που οδηγούν τους πρωταγωνιστές προς κάποιο βασικό στόχο,

[1] Ο αγγλικός τίτλος, το έτος παραγωγής και το στούντιο παραγωγής καθεμιάς από αυτές τις ταινίες είναι: *Chicago* 2002 (Miramax), *Crash* 2004 (Lions Gate), *The Departed* 2006 (Warner Bros.), *Gladiator* 2000 (Universal & DreamWorks), *Lord of the Rings - The Return of the King* 2003 (New Line), *Finding Nemo* 2003 (Walt Disney Pictures, Pixar Animation Studios).

*Πίνακας 5.3:* Βάση δεδομένων Ταινιών.

| **MovieSum Database** | |
|---|---|
| Είδος Αρχείων | Οπτικοακουστικά |
| Media & Πηγές Πληροφορίας | Ομιλία<br>Κείμενο (υπότιτλοι, κείμενο γραφικών, κείμενο σκηνών)<br>Ήχος: Μουσική, Ομιλία, Ηχητικά Εφέ, Περιβαλλοντικοί ήχοι (φυσικοί ή τεχνητοί) |
| Μορφή Αρχείων | AVI<br>Υψηλή ποιότητα: 25fps, 720x572, Xvid, MPEG4 codec, aspect ratio: 16:9, BitRate: ca. 2000kbps, 44100Hz audio sampling, 2-stereo, PCM<br>Χαμηλή ποιότητα: 25fps, 352x240, Xvid, MPEG4 codec, BitRate: ca. 1200kbps, 44100Hz audio sampling, 2-stereo, PCM |
| Είδος Ταινίας | Ταινίες βραβευμένες με Όσκαρ Καλύτερης Ταινίας (δράσης, δραματικές, κωμωδίες, επικές, κινουμένων σχεδίων) |
| Μέγεθος | 4 ώρες |
| Γλώσσα | Αγγλικά |

το συναισθηματικό περιεχόμενο, η επιθυμία ή η αντιπαράθεση δημιουργούν τη δομή της πλοκής. Επιπλέον, γενικά οι ταινίες αμερικανικής παραγωγής διαθέτουν χαρακτηριστικά όπως οι έντονες χρωματικές εναλλαγές, τα οπτικά εφέ, η συνοδευτική μουσική και τα ηχητικά εφέ, η ταχύτητα της δράσης αλλά και η συνεχής ροή γεγονότων, τα οποία θεωρούνται τα κύρια εργαλεία για την ανάπτυξη και την εξέλιξη της πλοκής. Το είδος αυτό της δομής που περιλαμβάνεται στις ταινίες μπορεί να οδηγήσει σε αποτελεσματικές περιλήψεις.

**Επισημείωση:** Η επισημείωση των ταινιών έγινε από τρεις έμπειρους χρήστες, βάσει του οπτικοακουστικού περιεχομένου τους, ύστερα από εκπαίδευση, με τη χρήση κατά τη διάρκεια της διαδικασίας εγχειριδίου οδηγιών και «κανόνων». Αρχικά, το κλιπ κάθε ταινίας επισημειώθηκε σε σχέση με τη σημασιολογική δομή, δηλ. κατάτμηση του μισάωρου τμήματος σε πλάνα και σκηνές, όπου ως πλάνο ορίζεται το διάστημα συνεχόμενης λήψης της κάμερας ανάμεσα σε δύο *cuts* (με μέση διάρκεια 2.5 δευτερόλεπτα) ενώ μια σκηνή, αποτελεί μία πλήρη, μεγαλύτερη αφηγηματική ενότητα με συνεχή ροή γεγονότων, που εμφανίζονται στον ίδιο τόπο και χρόνο (με μέση διάρκεια 3.5 λεπτά). Στη συνέχεια, εκτελέστηκε η επισημείωση με βάση τα εξής κριτήρια:

i. Τη **μονοτροπική** (*monomodal*) και την **πολυτροπική σημαντικότητα ή προκαλούμενη προσοχή** (*multimodal saliency*) του βίντεο (*sensory information*). Εδώ επισημειώνονται τμήματα του βίντεο τα οποία θεωρούνται ενδιαφέροντα από ακουστική, οπτική και οπτικοακουστική άποψη.

ii. Το **σημασιολογικό περιεχόμενο** (*cognitive information*) το οποίο περιλαμβάνει την επισημείωση σημαντικών γεγονότων βάσει της αισθητηριακής και σημασιολογικής πληροφορίας.

iii. Το **συναίσθημα** (*affective information*) – προκαλούμενο (*experienced*) ή επιδιωκόμενο (*intended*). Για περισσότερες λεπτομέρειες σχετικά με την επισημειώση όσον αφορά το συναίσθημα βλ. [99].

Η βάση αποτελείται επίσης από περιλήψεις, περίπου 5 λεπτών, που έχουν δημιουργηθεί από έμπειρο χρήστη (ο οποίος σχετίζεται επαγγελματικά με την παραγωγή ταινιών/διαφημιστικών σποτ, μοντάζ κ.λπ.). Οι οδηγίες που δόθηκαν ήταν να δημιουργηθεί μια ουσιώδης περίληψη σε σχέση με την πλοκή του τριαντάλεπτου τμήματος. Η διάρκεια της μπορούσε να ποικίλλει μεταξύ 1–10 λεπτών, σύμφωνα με την εκτίμηση και τις προτιμήσεις του χρήστη. Επιπλέον, σημεία και γεγονότα με έντονα ηχητικά/οπτικά εφέ, που συνήθως ελκύουν τον θεατή, δεν είναι απαραίτητα εκτός και αν εμπεριέχουν θεμελιώδη στοιχεία για την εξέλιξη της πλοκής.

Για τη διαδικασία της επισημείωσης δημιουργήθηκε ειδικό εγχειρίδιο που περιλαμβάνει την περιγραφή του εργαλείου επισημείωσης Anvil (http://www.anvil-software.de), της διαδικασίας επισημείωσης με λεπτομέρειες και παραδείγματα αλλά και με τεχνικά χαρακτηριστικά των ταινιών, τον τρόπο επεξεργασίας των DVD για τη δημιουργία των μισάωρων τμημάτων καθώς και πληροφορίες σχετικές για τα βασικά βήματα που δημιουργούν τη δομή του είδους των ταινιών που εμπεριέχονται στη βάση. Η συστηματική αυτή καταγραφή όλων των σταδίων δημιουργίας της βάσης (από τη συλλογή έως την επισημείωση) καθιστά δυνατή τη μελλοντική εξέλιξη και ανάπτυξη της βάσης και τη διατήρηση όλων των τεχνικών χαρακτηριστικών της. Επιπλέον, οδηγεί σε πιο έγκυρη επισημείωση καθώς και σε μεγαλύτερη συμφωνία μεταξύ των επισημειωτών.

Η επισημείωση της βάσης αποτελεί «*ground truth*» για την αξιολόγηση του αλγορίθμου αυτόματης δημιουργίας περιλήψεων. Πρέπει να αναφέρουμε πως η επισημείωση γίνεται βάσει δυαδικής λογικής, όπου 1 σημαίνει πως υπάρχει κάποιο σημαντικό γεγονός ενώ 0 πως δεν υπάρχει. Έτσι για την τελική αξιολόγηση δημιουργείται μια δυαδική συνάρτηση όπου τα σημεία (1) που θεωρούνται σημαντικά έχουν επισημειωθεί τουλάχιστον από τους 2 στους 3 επισημειωτές.

Η βάση τη δεδομένη χρονική στιγμή βρίσκεται σε στάδιο εξέλιξης. Συγκεκριμένα έχει επισημειωθεί ακόμα μία ολόκληρη ταινία (Gone with the Wind) καθώς και πέντε ταξιδιωτικά ντοκιμαντέρ με τα κριτήρια που ήδη παρουσιάστηκαν.

Λεπτομέρειες για το σύνολο της ερευνητικής αυτής προσπάθειας μπορούν να βρεθούν στις αντίστοιχες δημοσιεύσεις (Παράρτημα Α', δημοσιεύσεις J2, C1, C2, C3, C4).

# Κεφάλαιο 6

# Σύνοψη Προόδου και Κατευθύνσεις Μελλοντικής Έρευνας

## 6.1 Ερευνητική Συνεισφορά

Η διδακτορική αυτή έρευνα κινείται στην περιοχή της ψηφιακής επεξεργασίας μουσικών σημάτων και αφορά την ανάλυσή τους με υπολογιστικές μεθόδους για την εξαγωγή χρήσιμης πληροφορίας για την αναγνώρισή τους. Συγκεκριμένα, διερευνούμε και επεκτείνουμε μεθόδους που βασίζονται σε μη-γραμμικά μοντέλα για την κατανόηση της δομής των μουσικών ήχων και των σχέσεων των μουσικών οργάνων και την εξέταση των γνωρισμάτων των διαφορετικών ειδών μουσικής. Παράλληλα διερευνούμε τις μεθόδους αυτές για εφαρμογές όπως η δημιουργία ηχητικών συνόψεων. Η έρευνα αυτή συνεισφέρει στην τεχνολογία αιχμής που σχετίζεται με την αυτόματη κατηγοριοποίηση της μουσικής μέσω των διαφορετικών αυτών πλαισίων αλλά και τη γρήγορη αναζήτηση πληροφοριών του περιεχομένου των δεδομένων.

Οι ερευνητικές μας συνεισφορές μπορούν να συνοψισθούν στα ακόλουθα σημεία:

- Ανάπτυξη και επέκταση μεθόδων για την ανάλυση μουσικών σημάτων και την αναγνώριση των διαφορετικών μουσικών οργάνων:

  - Προτείναμε τη μη-γραμμική επεξεργασία των μουσικών σημάτων, με ιδέες της Φράκταλ θεωρίας, όπου και εξετάσαμε τη φράκταλ διάσταση των ήχων των μουσικών οργάνων σε πολλαπλές κλίμακες. Διεξάγαμε εκτενή ανάλυση της δομής των μουσικών σημάτων για τις διαφορετικές μεταβατικές καταστάσεις και προσδιορίσαμε το MFD (*Multiscale Fractal Dimension*) προφίλ των διαφορετικών οργάνων με τη χρήση του αλγορίθμου μορφολογικής κάλυψης. Επίσης εφαρμόσαμε τις ιδέες μας σε συνθετικά σήματα προκειμένου να αξιολογήσουμε τις διάφορες παρατηρήσεις μας. Απόρροια της ανάλυσης

αυτής ήταν η εξαγωγή περιγραφικών και σύντομων αναπαραστάσεων των μουσικών σημάτων, τις οποίες χρησιμοποιήσαμε σε πειράματα αναγνώρισης για να ενισχύσουμε καθιερωμένα σύνολα χαρακτηριστικών, όπως τα MFCC, με αποτέλεσμα τη μείωση του λάθους αναγνώρισης των μουσικών οργάνων ως και 32%.

- Διερευνήσαμε και επεκτείναμε το μοντέλο διαμόρφωσης πλάτους και συχνότητας (AM-FM), εξάγοντας χαρακτηριστικά όπως το μέσο στιγμιαίο πλάτος και η μέση στιγμιαία συχνότητα για τη μοντελοποίηση των μικροδομών των μουσικών ήχων και την κατηγοριοποίηση των μουσικών οργάνων. Διαπιστώνοντας την ακρίβεια του μοντέλου και τη δυνατότητα των χαρακτηριστικών για τη συγκεκριμένη εφαρμογή συνεχίσαμε με την ανάλυση και την εφαρμογή του Επαναληπτικού-ESA (*Iterative-ESA*), όπου και παρατηρήσαμε τη δυνατότητα του αλγορίθμου για τη δημιουργία καλύτερων εκτιμήσεων των χαρακτηριστικών και πιο εύστοχο προσδιορισμό των δομών της μουσικής. Επιπρόσθετα διαπιστώσαμε πως θα μπορούσαμε πιθανώς να εκτιμήσουμε το αρμονικό περιεχόμενο των μουσικών ήχων. Η ανάλυση με τα μοντέλα διαμορφώσεων οδήγησε σε πολύ καλά αποτελέσματα αναγνώρισης των μουσικών οργάνων με μείωση του λάθους αναγνώρισης ως και 56%–60% (μέση ακρίβεια αναγνώρισης 95.89%–98.64% αντίστοιχα για 12 και 7 μουσικά όργανα) σε συνδυασμό με τα MFCC και με τη χρήση των HMM.

- Διερεύνηση μεθόδων επεξεργασίας σήματος για την ανάλυση και την κατηγοριοποίηση των διαφορετικών ειδών της μουσικής:

  - Διερευνήσαμε το μη-γραμμικό μοντέλο διαμόρφωσης πλάτους και συχνότητας για την ανάλυση και την κατηγοριοποίηση των μικροδομών και των μακροδομών των μουσικών σημάτων. Επεκτείναμε τον αλγόριθμο διαχωρισμού ενέργειας ενώ παράλληλα προτείναμε τη δημιουργία συστοιχίας φίλτρων εστιασμένων στη δομή των μουσικών σημάτων. Επιπρόσθετα, εξετάσαμε διαφορετικές μορφές αναπαραστάσεων των χαρακτηριστικών, για παράδειγμα περιγραφείς βασισμένους σε ανάλυση βραχέος χρόνου ή στις μακροδομές της μουσικής. Η ανάλυση με τα μοντέλα διαμορφώσεων οδήγησε σε ενθαρρυντικά αποτελέσματα αναγνώρισης των διαφορετικών ειδών μουσικής με μείωση του σφάλματος ως και 28%. Συμπεραίνουμε άρα πως οι διαμορφώσεις μπορούν να περιγράψουν σημαντικά φαινόμενα των σημάτων μουσικής, όπως τις μικρο-μεταβολές που συμβαίνουν λόγω των δομών της. Επιπλέον, η χρήση αναπαραστάσεων οι οποίες βασίζονται στις μακροδομές επιφέρουν μείωση της πολυπλοκότητας του συστήματος κατηγοριοποίησης, εφόσον

επιτυγχάνουν ικανοποιητικά αποτελέσματα χρησιμοποιώντας απλούστερα στατιστικά μοντέλα, τύπου GMM. Τέλος, η εισαγωγή της «μουσικής» συστοιχίας φίλτρων δημιουργεί σύνολα χαρακτηριστικών με αξιόλογη διακριτική ικανότητα.

- Ακολουθώντας μία άλλη προσέγγιση, προτείναμε την ιδέα των Bag-of-Words καθιστώντας έτσι δυνατή την εισαγωγή εναλλακτικών αναπαραστάσεων των μουσικών σημάτων. Εφαρμόζοντας τη μέθοδο αυτή δημιουργούμε ένα «μουσικό λεξικό» και περιγράφουμε το κάθε μουσικό κομμάτι, βάσει της συχνότητας των «μουσικών λέξεων» που περιλαμβάνει. Λόγω των νέων συμπαγών αναπαραστάσεων, οι οποίες θεωρούμε πως περιγράφουν φράσεις, μοτίβα καθώς και άλλα δομικά στοιχεία των μουσικών κομματιών, αντιμετωπίζουμε διάφορα προβλήματα πολυπλοκότητας κατά την κατηγοριοποίηση. Η αξιολόγηση των πειραμάτων διενεργήθηκε με τη χρήση των SVMs και παρουσίασε μείωση λάθους ως και 11% για τον συνδυασμό των μη-γραμμικών AM-FM και φράκταλ χαρακτηριστικών σε σχέση με τα MFCC, και 16% σε συνδυασμό με τα MFCC.

- Διερεύνηση μεθόδων επεξεργασίας σήματος για τη δημιουργία ηχητικών συνόψεων:

  - Μελετήσαμε και αξιολογήσαμε το μη-γραμμικό μοντέλο διαμόρφωσης πλάτους και συχνότητας (AM-FM) ως προς τη χρησιμότητα και την καταλληλότητά του σε θέματα ανίχνευσης σημαντικών μουσικών και ακουστικών γεγονότων. Βασιστήκαμε στη χαμηλού επιπέδου πληροφορία του σήματος για να προσεγγίσουμε το πρόβλημα, χρησιμοποιώντας αναπαραστάσεις της ηχητικής κυματομορφής. Επιπλέον εξετάσαμε νέα υπολογιστικά μοντέλα σύμμειξης των AM-FM χαρακτηριστικών για τη δημιουργία περιγραφικών καμπυλών σημαντικότητας οι οποίες και αποτελούν το κριτήριο δημιουργίας ηχητικών συνόψεων. Τέλος προτείναμε την επέκταση υπάρχοντος αλγορίθμου για τη δημιουργία των περιλήψεων, που σαν αποτέλεσμα έχει ηχητικές συνόψεις με καλύτερη αντιληπτική ποιότητα και αισθητική.
  - Επεκτείναμε τις παραπάνω ιδέες σε πολυμεσικά δεδομένα και συγκεκριμένα ταινίες. Επιπλέον δημιουργήσαμε συστηματική βάση ταινιών («*MovieSum Database*») για τη διεξοδική αξιολόγηση και έλεγχο της αποδοτικότητας των αλγορίθμων και των αποτελεσμάτων μας στην ανίχνευση σημαντικών ακουστικών και οπτικοακουστικών γεγονότων αλλά και την παραγωγή περιλήψεων.

Κατάλογος με τις σχετικές μας δημοσιεύσεις παρατίθεται στο Παράρτημα Α΄.

## 6.2 Κατευθύνσεις Μελλοντικής Έρευνας

Με αφορμή την παρούσα διδακτορική έρευνα, διαγράφονται διάφορες κατευθύνσεις για μελλοντική έρευνα. Συγκεκριμένα οι κατευθύνσεις αυτές είναι οι εξής:

- Οι μεθοδολογίες οι οποίες ακολουθήθηκαν για τις εφαρμογές της διατριβής αυτής, αναδεικνύουν τη δυναμική των μοντέλων που μελετήθηκαν. Για το λόγο αυτό θα μπορούσαν να φανούν σημαντικές για τη μοντελοποίηση και την κατηγοριοποίηση των σημάτων μουσικής ως προς τις εκφραστικές τους ιδιότητες (*Mood Classification*).

- Η δημιουργία ζωνοπερατών μουσικών αναπαραστάσεων με τη χρήση εστιασμένων συστοιχιών φίλτρων, όπως η μουσική συστοιχία Gabor φίλτρων επέδειξε ιδιαίτερη διακριτική ικανότητα. Θεωρούμε πως μπορεί να χρησιμοποιηθεί για την επεξεργασία και τη μοντελοποίηση των μουσικών σημάτων σε εφαρμογές διαφορετικές της κατηγοριοποίησης. Επιπλέον πιστεύουμε πως ο συνδυασμός των συγκεκριμένων χαρακτηριστικών με άλλα βασικά χαρακτηριστικά θα μπορούσε να επιφέρει περαιτέρω βελτίωση των αποτελεσμάτων αναγνώρισης.

- Η μέθοδος δημιουργίας συνόλων χαρακτηριστικών που βασίζονται στις μακροδομές των σημάτων μουσικής είναι εφικτό να γίνει πιο αποδοτική χρησιμοποιώντας διαφορετικές μεθόδους για τη μείωση του χώρου των χαρακτηριστικών, αντί της ανάλυσης σε κύριες συνιστώσες (PCA). Για παράδειγμα η ανάλυση με Heteroscedastic Linear Discriminant Analysis (HLDA) έχει παρουσιάσει ιδιαίτερα ικανοποιητικά αποτελέσματα σε εφαρμογές αναγνώρισης φωνής. Η επέκταση αυτή του αλγορίθμου, για τη δημιουργία αναπαραστάσεων των μουσικών σημάτων, θεωρούμε πως θα μπορούσε να οδηγήσει σε επιπλέον βελτιώσεις των αποτελεσμάτων κατηγοριοποίησης.

# Βιβλιογραφία


[1] G. Agostini, M. Longari, and E. Pollastri. Musical instrument timbres classification with spectral features. *Journal on Applied Signal Process.*, 1:5–14, 2003.

[2] American Standard Association. *Acoustical Terminology*, New York, 1960.

[3] E. Amid and S. R. Aghdam. Musical instrument classification using embedded hidden Markov models. In *World Academy of Science, Engineering and Technology*, volume 67, 2012.

[4] I. Antonopoulos, A. Pikrakis, and S. Theodoridis. Self-similarity analysis on tempo induction from music recordings. *Journal of New Music Research*, 36(1):27–38, 2007.

[5] J. J. Aucouturier and F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

[6] R. Bader. Characterization of guitars through fractal correlation dimensions of initial transients. *Journal of New Music Research*, 25(4):323–332, 2006.

[7] P. Ball. *The Music Instinct, How Music Works and Why We Can't Do Without It.* Oxford Univ. Press, 2010.

[8] J. Barbedo and G. Tzanetakis. Musical instrument classification using individual partials. *IEEE Trans. on Audio, Speech and Language Process.*, 19(1), 2010.

[9] J. Barbedo, G. Tzanetakis, N. Ono, and S. Sagayma. Beyond timbral statistics: Improving music classification using percussive patterns and bass lines. *IEEE Trans. on Audio, Speech, and Language Process.*, 19(4), 2011.

[10] L. Barreira, S. Cavaco, and J. F. da Silva. *Unsupervised Music Genre Classification with a Model-Based Approach*, volume 7026, pages 268–281. Springer Berlin Heidelberg, Progress in Artificial Intelligence, Lecture Notes in Computer Science, 2011.

[11] E. Benetos and C. Kotropoulos. Non-negative tensor factorization applied to music genre classification. *IEEE Trans. on Audio, Speech, and Language Process.*, 18(8):1955–1967, 2010.

[12] E. Benetos, M. Kotti, and C. Kotropoulos. Musical instrument classification using non-negative matrix factorization algorithms. In *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2006.

[13] J. Bicknell. *Why Music Moves Us*. Palgrave MacMillan, 2009.

[14] M. Bigerelle and A. Iost. Fractal dimension and classification of music. *Chaos, Solitons, & Fractals*, 11:2179 – 2192, 2000.

[15] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science & Business Media, LLC, 2006.

[16] A. Bosch, X. Munoz, and R. Martí. A review: Which is the best way to organize/classify images by content? *Image and Video Computing, IVC 01*, 2006.

[17] A. C. Bovik, P. Maragos, and T. F. Quatieri. AM-FM energy detection and separation in noise using multiband energy operators. *IEEE Trans. on Signal Processing*, 41(12):3245–3265, 1993.

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

[19] J. C. Brown, O. Houix, and S. McAdams. Feature dependence in the automatic identification of musical woodwind instruments. *J. Acoust. Soc. Amer.*, 109(3):1064 –1072, Mar. 2001.

[20] S. Brown. The "musilanguage" model of music evolution. In N. L. Wallin, B. Merker, and S. Brown, editors, *The Origins of Music*, pages 271–300. MIT Press, 2000.

[21] S. Brown. Are music and language homologues. *The Biological Foundations of Music*, 930:372–374, 2001.

[22] J.J. Burred and T. Sikora. Monaural source separation from musical mixtures based on time-frequency timbre models. In *Proc. Int'l. Conf. on Music Information Retrieval (ISMIR-07)*, 2007.

[23] L. Cardoso, R. Panda, and R. P. Paiva. MOODetector: A prototype software tool for mood-based playlist generation. In *Simposio de Informatica -INForum 2011*, 2011.

[24] R. P. Carlyon. How the brain separates sounds. *Trends in Cognitive Sciences*, 8(10):465–471, Oct. 2004.

[25] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE Trans. on Neural Networks*, 10(5):1055–1064, 1999.

[26] M. Cooper and J. Foote. Automatic music summarization via similarity analysis. In *Proc. Int'l. Conf. Music Information Retrieval (ISMIR-02)*, pages 81–85, 2002.

[27] C. V. Cotton and D. P. W. Ellis. Spectral vs. spectro-temporal features for acoustic event detection. In *IEEE Workshop on Applications of Signal Process. to Audio and Acoustics (WASPAA-11)*, New Paltz, NY, 2011.

[28] A. Das and P. Das. Fractal analysis of songs: Performer's preference. *Nonlinear analysis: Real World Applications (Elsevier)*, 11(3):1790–1793, 2010.

[29] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, and Signal Process.*, 28(4):357–366, 1980.

[30] J. D. Deng, C. Simmermacher, and S. Cranefield. A study on feature analysis for musical instrument classification. *IEEE Trans. on Systems, Man, and Cybernetics - Part B: Cybernetics*, 38(2):429–438, 2008.

[31] A. G. Dimakis and P. Maragos. Phase modulated resonances modeled as self-similar processes with application to turbulent sounds. *IEEE Trans. on Signal Process.*, 53(11):4261–4272, Nov. 2005.

[32] D. Dimitriadis and P. Maragos. Continuous energy demodulation methods and application to speech analysis. *Speech Communication*, 48:819–837, 2006.

[33] D. Dimitriadis, P. Maragos, and A. Potamianos. Robust AM-FM features for speech recognition. *IEEE Signal Processing Letters*, 12(9):621–624, Sep. 2005.

[34] M. Eichner, M. Wolff, and R. Hoffmann. Instrument classification using hidden Markov models. In *Int'l. Symp. on Music Information Retrieval (ISMIR-06)*, 2006.

[35] A. Eronen. Comparison of features for musical instrument recognition. In *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001.

[36] A. Eronen. Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In *Proc. Signal Process. & Its Applications*, volume 2, pages 133–136, 2003.

[37] S. Essid, G. Richard, and B. David. Musical instrument recogtition on solo performances. In *Proc. Int'l Conf European Signal Processing (EUSIPCO-04)*, 2004.

[38] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Trans. on Audio, Speech, and Language Process.*, 14(1):68–80, 2006.

[39] S. Essid, G. Richard, and B. David. Musical instrument recognition by pairwise classification strategies. *IEEE Trans. on Audio, Speech, and Language Process.*, 14(4):1401–1312, 2006.

[40] G. Evangelopoulos and P. Maragos. Multiband modulation energy tracking for noisy speech detection. *IEEE Trans. on Audio, Speech, and Language Process.*, 14:2024–2038, Nov. 2006.

[41] 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. on Multimedia*, 15(7):1553–1568, Nov. 2013.

[42] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis. Video event detection & summarization using audio, visual & text saliency. In *Proc. Int'l Conf. Acoustics, Speech and Signal Process. (ICASSP-2009)*, Taipei, Taiwan, Apr. 2009.

[43] K. Falconer. *Fractal Geometry, Mathematical Foundations and Applications*. John Wiley & Sons, 2nd edition, 2003.

[44] R. R. Alves Faria, R. A. Rushioni, and J. A. Zuffo. Wavelets in music analysis and synthesis: Timbre analysis and perspectives. In *Proc. SPIE 2825, Wavelet Applications in Signal and Image Processing IV, 950*, 1996.

[45] Kitty Ferguson. *Η μουσική του Πυθαγόρα (The music of Pythagoras)*. ΕΚδοτικός Οίκος ΤΡΑΥΛΟΣ, 2008.

[46] H. Fletcher. Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *J. Acoust. Soc. Amer.*, 6(2):59–69, Oct. 1934.

[47] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer, 2nd edition, 1998.

[48] A. Friberg. Digital audio emotions - An overview of computer analysis and synthesis of emotional expression in music. In *Proc. Int'l. Conf. on Digital Audio Effects (DAFx-08)*, 2008.

[49] A. Friberg and A. Hedblad. A comparison of perceptual ratings and computed audio features. In *Proc. Conf. Sound and Music Computing*, 2011.

[50] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma. Auditory attention–focusing the searchlight on sound. *Current Opinion in Neurobiology*, 17(4):437–455, Aug. 2007.

[51] A. Gabrielsson and E. Lindström. The influece of musical structure on emotinal expression. In P. N. Juslin and J. A. Sloboda, editors, *Music and Emotion: Theory and Research*, chapter 10, pages 223–248. Oxford University Press, 2001.

[52] R. O. Gjerdingen and D. Perrott. Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, 37(2):93–100, 2008.

[53] M. Goto, H. Hashiguish, T. Nishimura, and R. Oka. RWC music database: popular, classical, and jazz music databases. In *Proc. Int'l. Symp. Music Information Retrieval (ISMIR-02)*, 2002.

[54] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti. Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR. In *Proc. Int'l Conf. Acoust., Speech, Signal Process*, pages 853–856, 2002.

[55] J. M Grey. Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.*, 61(5):1270–1277, 1977.

[56] E. Guaus and P. Herrera. A basic system for music genre classification. In *Music Information Retrieval Evaluation eXchange (MIREX-07)*, 2007.

[57] G. Gündüz and U. Gündüz. The mathematical analysis of the structure of some songs. *Physica A*, 357:565 – 592, 2005.

[58] D. E. Hall. *Musical Acoustics*. Brooks/Cole, 3rd edition, 2002.

[59] H. M. Hanson, P. Maragos, and A. Potamianos. A system for finding speech formants and modulations via energy separation. *IEEE Trans. on Speech and Audio Process.*, 2(2):436–443, July 1994.

[60] P. Herrera-Boyer, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21, 2003.

[61] P. Holonowicz, P. Herrera, and H. Purwins. The role of loudness in detection of surprising events in music recordings. In *7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM-09)*, 2009.

[62] A. Holzapfel and Y. Stylianou. Musical genre classification using non-negative matrix factorization-based features. *IEEE Trans. on Audio, Speech, and Language Process.*, 16(2):424–434, 2008.

[63] K. J. Hsu and A. J. Hsu. Fractal geometry of music. In *Proc. Natl. Acad. Sci. USA*, volume 87, pages 938 – 941, Feb. 1990.

[64] D. Huron. Music in advertising: An analytic paradigm. *Musical Quartely*, 73(4):557–574, 1989.

[65] D. Huron and B. Aarden. Cognitive issues and approaches in music information retrieval, 2002 [Online]. Available: http://www.musicog.ohio-state.edu/Huron/publications.html, unpublished.

[66] P. Iverson and C. L. Krumhansl. Isolating the dynamic attributes of musical timbre. *J. Acoust. Soc. Amer.*, 95(5):2595–2603, Nov. 1993.

[67] N. Jakovljevic, D. Miskovic, M. Janev, M. Secujski, and V. Delic. Comparison of linear discriminant analysis approaches in automatic speech recognition. *Electronics and Electrical Engineering, ISSN 1392-1215*, 19(7):76–79, 2013.

[68] D. Jang, M. Jin, and C. D. Yoo. Music genre classification using novel features and a weighted voting method. In *IEEE Int'l. Conf. on Multimedia and Expo*, 2008.

[69] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal Bag-of-Features for object categorization and semantic video retrieval. In *Proc. ACM Int'l Conf. on Image and Video Retrieval (CIVR-07)*, 2007.

[70] F. Jing, M. Li, H.-J. Zhang, and B. Zhang. Support vector machines for region-based image retrieval. In *in Proc. Int'l. Conf. on Multimedia and Expo (ICM-03)*, 2003.

[71] C. Joder, S. Essid, and G. Richard. Temporal integration for audio classification with application to musical instrument classification. *IEEE Trans. on Audio, Speech, and Language Process.*, 17(1):174–186, 2009.

[72] P. N. Juslin. Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Perormance*, 26(6):1797–1813, 2000.

[73] P. N. Juslin and P. Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33:217–238, 2004.

[74] P. N. Juslin and D. Västfjäll. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, 31(5):559–621, 2008.

[75] S. Karnebäck. Discrimination between speech and music based on a low frequency modulation feature. In *Proc. European Conf. on Speech Comm. and Technology*, 2001.

[76] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, 15(21):1943–1947, 2005.

[77] P. Kivy. *Introduction to a Philosophy of Music*. Oxford Univ. Press, 2002.

[78] S. Koelsch, T. Fritz, D. Yves, V. Cramon, K. Müller, and A. D. Friederici. Investigating emotion with music: An fMRI study. *Human Brain Mapping*, 27:239–250, 2006.

[79] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Trans. on Speech and Audio Process.*, 13(6):1098–1109, 2005.

[80] S. K. Kopparapu, M. A. Pandharipande, and G. Sita. Music and vocal separation using multiband modulation based features. In *IEEE Symposium on Industrual Electronics and Applications*, pages 733–737, 2010.

[81] B. Kostek and P. Zwan. Wavelet-based automatic recognition of musical instruments. *J. Acoust. Soc. Amer.*, 2001.

[82] R. Kronland-Martinet, J. Morlet, and A. Grossmann. Analysis of sound patterns through wavelet transforms. *Int'l. Jour. Pattern Recognition and Artificial Intelligence*, 1(2):273–302, 1987.

[83] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. Computer Vision and Pattern Recognition (CVPR-08)*, 2008.

[84] C.-H. Lee, J.-L. Shioh, K.-M. Yu, and J.-M. Su. Automatic music genre classification using modulation spectral contrast feature. In *Proc. Int'l. Conf. on Multimedia and Expo (ICME-2007)*, 2007.

[85] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, 1983.

[86] C. Lèvi-Strauss. *Myth and Meaning*. Schocken Books New York, 1979.

[87] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proc. Int'l. ACM Conf. Research and Development in Information Retrieval*, 2003.

[88] T. Lidy and A. Rauber. Combined fluctuation features for music genre classification. In *Proc. Int'l. Symp. Music Information Retrieval (ISMIR-05)*, 2005.

[89] T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. Int'l Symp. Music Information Retrieval (ISMIR-05)*, 2005.

[90] T. Lidy and E. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. Int'l. Symp. Music Information Retrieval (ISMIR-05)*, pages 34–41, London, UK, 2005.

[91] A. Livshin, G. Peeters, and X. Rodet. Studies and improvements in automatic classification of musical sound samples. In *Proc. Computer Music Conference, (ICMC-03)*, 2003.

[92] A. Livshin and X. Rodet. Musical instrument identification in continuous recordings. In *Proc. Int'l Conf. Digital Audio Effects (DAFx-04)*, 2004.

[93] B. Logan and S. Chu. Music summarization using key phrases. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP-00)*, 2000.

[94] L. Lue and H.-J. Zhang. Automated extraction of music snippets. In *Proc. Int'l Conf. ACM*, 2003.

[95] L.-O. Lundqvist, F. Carlsson, P. Hilmersson, and P. N. Juslin. Emotional responses to music: Experience, expression, and physiology. *Psychology of Music*, 37(1):61–90, 2009.

[96] S. A. Shamma M. Elhilali, J. Xiang and J. Z. Simon. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS biology*, 7(6), 2009.

[97] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *Proc. Int'l. Conf. on ACM Multimedia*, pages 533–542, 2003.

[98] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of 5th Symp. on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[99] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi. A supervised approach to movie emotion tracking. In *Proc. Int'l. Conf. on Acoustics, Speech and Signal Process., (ICASSP-11)*, pages 2376–2379, 2011.

[100] M. Mandel and D. Ellis. Song-level features and Support Vector Machines for music classification. In *Proc. Int'l. Symp. Music Information Retrieval (ISMIR-05)*, pages 594–599, London, UK, 2005.

[101] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W.H. Freeman, San Francisco, 1982.

[102] P. Maragos. Fractal aspects of speech signals: Dimension and interpolation. In *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 1991.

[103] P. Maragos. Fractal signal analysis using mathematical morphology. *Advances in electronics and electron physics, Academic Press*, 88:199 – 246, 1994.

[104] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. on Signal Process.*, 41:3024–3051, Oct. 1993.

[105] P. Maragos and A. Potamianos. Fractal dimension of speech sounds: Computation and application to automatic speech recognition. *J. Acoust. Soc. Amer.*, 105(3):1925 – 1932, 1999.

[106] M. Markaki and Y. Stylianou. Discrimination of speech from nonspeech in broadcast news based on modulation frequency features. *Speech Communication*, 53(5):726–735, 2010.

[107] M. Marolt. Non-negative matrox factorization with selective sparsity constraints for transcription of bell chiming recordings. In *Proc. 6th Sound and Music Computing Conf. (SMC-09)*, 2009.

[108] C. Marques, I. R. Guilherme, R. Y. M. Nakamura, and J. P. Papa. New trends in musical genre classification using optimum-path forest. In *Proc Int'l Conf. Music Information Retrieval (ISMIR-2011)*, 2011.

[109] J. Marques and P. J. Moreno. A study of musical instrument classification using Gaussian mixture models and Support Vector Machines. Technical report, Camprige Research Laboratory, 1999.

[110] K. D. Martins. *Sound-source recognition: A theory and computational model*. PhD thesis, Maccachusets Institute of Technology, 1999.

[111] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on Acoustics, Speech and Signal Process.*, 34(4):744–754, 1986.

[112] G. Monro. Fractal interpolation waveforms. *Computer Music Journal*, 19(1):88–98, 1995.

[113] B.C.J. Moore. *Psychology of Hearing*. Acad. Press, 5th ed., 2003.

[114] O. M. Mubarak, E. Ambikairajah, J. Epps, and T. S. Gunawan. Modulation features for speech and music classification. In *Proc. Int'l Conf. Communication systems (ICCS-2006)*, pages 1–5, Oct. 2006.

[115] A. B. Nielsen, S. Sigurdsson, L. K. Hansen, and J. Arenas-Garcia. On the relevance of spectral features for instrument classification. In *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2007.

[116] H. F. Olson. *Music, Physics and Engineering*. Dover, 1967.

[117] F. Opolko and J. Wapnick. McGill University Master Samples, 1987. McGill Univ.

[118] F. Pachet and D. Cazaly. A taxonomy of musical genres. In *Proc. Content-Based Multimedia Information Access Conf. (RIAO), Paris*, Apr. 2000.

[119] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio based music similarity and genre classification? In *Proc. Int'l. Symp. Music Information Retrieval (ISMIR-05)*, pages 628–633, London, UK, 2005.

[120] Y. Panagakis and C. Kotropoulos. Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In *Proc. Int'l. Acoustics, Speech and Signal Process. (ICASSP-10)*, 2010.

[121] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music genre classification via sparse representations of auditory temporal modulations. In *Proc. European Signal Processing Conf. (EUSIPCO-2009)*, 2009.

[122] A. D. Patel. Language, music, syntax and the brain. *REVIEW nature Neuroscience*, 6(7):674–681, July 2003.

[123] A. D. Patel. *Music, Language, and the Brain*. Oxford Univ. Press, 2008.

[124] A. D. Patel. Language, music, and the brain: a resource-sharing framework. In P. Rebuschat, M. Rohrmeier, J. Hawkins, and I. Cross, editors, *Language and Music as Cognitive Systems*, pages 204–223. Osford Univ. Press, 2012.

[125] G. Peeters. A large set of features for sound description (similarity and classification), in the CUIDADO project. Technical report, CUIDADO I.S.T Project Rep., IRCAM, 2004.

[126] I. Peretz. Music, language and modularity framed in action. *Psychologica Belgica*, 49:157–175, 2009.

[127] I. Peretz and M. Coltheart. Modularity of music processing. *Nature Neuroscience*, 6(7):688–691, July 2003.

[128] I. Peretz and K. L. Hyde. What is specific to music processing? Insights from congenital amusia. *Trends in Cognitive Sciences*, 7(8):362–367, 2003.

[129] A. Pikrakis. Audio thumbnailing in video sharing sites. In *Proc. 20th European Signal Processing Conf. (EUSIPCO-2012)*, 2012.

[130] A. Pikrakis, S. Theodoridis, and D. Kamarotos. Recognitions of isolated musical patterns in the context of greek traditional music. In *IEEE Int'l. Conf. on Electronics, Circuits and Systems (ICECS-96)*, 1996.

[131] A. Pikrakis, S. Theodoridis, and D. Kamarotos. Recognition of isolated musical patterns using hidden Markov models. In *Lecture Notes in Computer Science, Music and Artificial Intelligence*, pages 55–58. Springer, 2002.

[132] A. Pikrakis, S. Theodoridis, and D. Kamarotos. Classification of musical patterns using variable duration hidden Markov models. *IEEE Trans. on Audio, Speech and Language Process.*, 14(5):1795–1807, Sep. 2006.

[133] V. Pitsikalis and P. Maragos. Filtered dynamics and fractal dimensions for noisy speech recognition. *IEEE Signal Processing Letters*, 13(11):711–713, 2006.

[134] Reinier Plomp. *The Intelligent Ear: On the Nature of Sound Perception*. Psychology Press, 1st ed., 2001.

[135] G. De Poli and P. Prandoni. Sonological models for timbre characterization. *Journal of New Music Research*, 26(2):170–197, 1997.

[136] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gomez, S. Streich, and O. Beesuan. Melody transcription from music audio: Approaches and evaluation. *IEEE Trans. Audio, Speech, and Language Process.*, 15(4):1247–1256, 2007.

[137] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *J. Acoust. Soc. Amer.*, 9(6):196–200, Jun. 1996.

[138] A. Potamianos and P. Maragos. Speech processing applications using an AM-FM modulation models. *Speech Communication*, 28:195–209, 1999.

[139] G. Potamianos and H. P. Graf. Discriminative training of HMM stream exponents for audio-visual speech recognition. In *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, volume 6, pages 3733–3736, 1998.

[140] T. F. Quatieri, T. E. Hanna, and G. C. O'Leary. AM-FM separation using auditory-motivated filters. *IEEE Trans. on Speech and Audio Process.*, 5(5):465–480, Sep. 1997.

[141] L. Rabiner and B.-H. Juang. *Fundamantals of Speech Recognition*. Prentice Hall, 1993.

[142] L.R. Rabiner and R. W. Schafer. *Introduction to Digital Speech Processing*. Foundations and Trends in Signal Processing, 2007.

[143] S. K. Rankin and E. W. Large. Fractal tempo fluctuation and pulse prediction. *Music Perception*, 26(5):401–413, 2009.

[144] W. Ro and Y. Kwon. 1/f noise analysis of songs in various genres of music. *Chaos, Solitons, & Fractals*, 42:2305–2311, 2009.

[145] O. Sacks. *Musicophilia, Tales of Music and the Brain*. First Vintage Books edition, 2008.

[146] N. Scaringella and G. Zoia. On the modeling of time information for automatic genre recognition systems in audio signals. In *Proc. Int'l. Symp. Music Information Retrieval (ISMIR-05)*, 2005.

[147] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content a survey. *EEE Signal Processing Magazine*, 133, 2006.

[148] K. R. Scherer. Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them. *Journal of New Music Research*, 33(3), 2004.

[149] K. R. Scherer and M. R. Zentner. Emotional effects of music: Production rules. In P. N. Juslin and J. A. Sloboda, editors, *Music and Emotion: Theory and Research*, chapter 16, pages 361–392. Oxford University Press, 2001.

[150] K. R. Scherer, M. R. Zentner, and A. Schacht. Emotional states generated by music: and exploratory stydy of music experts. *Music Scientiae*, pages 149–171, 2002.

[151] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *Proc. 4th European Conf. Computer Vision (ECCV-96)*, volume 1, pages 610–619, 1996.

[152] S. C. Sekhar and T. V. Sreenivas. Novel approach to AM-FM decomposition with applications to speech and music analysis. In *Int'l Conf. Acoustics, Speech, and Signal Processing*, volume 2, pages 753–756, 2004.

[153] X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S. Pope, A. Picialli, and G. De Poli, editors, *Musical Signal Processing*. Swets & Zeitlinger, 1997.

[154] K. Seyerlehner. *Content-Based Music Recommender Systems: Beyond simple Frame-Level Audio Similarity*. PhD thesis, Johannes Kelpler University Linz, 2010.

[155] K. Seyerlehner, G. Widmer, and P. Knees. Frame level audio similarity - a codebook approach. In *Proc. 11th Int'l. Conf. on DIgital Audio Effects (DAFx-08)*, 2008.

[156] K. Seyerlehner, G. Widmer, and T. Pohle. Fusing block-level features for music similarity estimation. In *Proc. Int'l Conf. on Digital Audio Effects (DAFx-10), Graz, Austria*, 2010.

[157] X. Shao, X. Changsheng, and M. S. Kankanhalli. Unsupervised classification of music genre using hidden Markov models. In *IEEE Int. Conf. Multimedia and Expo, (ICME-04)*, pages 2023–2026, 2004.

[158] X. Shao, N. C. Maddage, X. Changsheng, and M. S. Kankanhalli. Automatic music summarization based on music structure analysis. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP-05)*, 2005.

[159] C. Simmermacher, D. Deng, and S. Cranefield. Feature analysis and classification of classical musical instruments: an empirical study. In P. Perner, editor, *Advances in Data Mining, Lecture Notes in Computer Science*, volume 4065, pages 333–358. Springer, 2006.

[160] N. Singh-Miller, M. Collins, and T. J. Hazen. Dimensionality reduction for speech recognition using neighborhood components analysis. In *Proc. Int'l. Speech Communication Conf. (InterSpeech-07)*, 2007.

[161] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. Int'l. Conf. on Computer Vision (ICCV-03)*, 2003.

[162] L. R. Slevc, J. C. Rosenberg, and A. D. Patel. Making psycholinguistics musical: Self-paced reading time evidence for shared processing of linguistic and musical syntax. *Psychonomic Bulletin & Review*, 16(2):374–381, 2009.

[163] J. A. Sloboda. Music structure and emotional response: Some empirical findings. *Psychology of Music*, 19:110–120, 1991.

[164] J. A. Sloboda and S. A. O'Neill. Emotions in everyday listening to music. In P. N. Juslin and J. A. Sloboda, editors, *Music and Emotion: Theory and Research*, pages 415–430. Oxford University Press, 2001.

[165] P. Smaragdis and J. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. IEEE WASPAA*, 2003.

[166] H. Soltau, T. Schultz, M. Westphal, and A. Waibel. Recognition of music types. In *Proc. Int'l. Conf. Acoustics, Speech and Signal Process. (ICASSP-98)*, 1998.

[167] D. Spachos, A. Zlatintsi, V. Moschou, P. Antonopoulou, E. Benetos, M. Kotti, K. Tzimouli, C. Kotropoulos, N. Nikolaidis, P. Maragos, and I. Pitas. MUSCLE Movie Database: A multimodal corpus with rich annotation for dialogue and saliency detection. In *Proc. Int'l Conf. Language Resources and Evaluation (LREC-2008)*, Marrakech, Marocco, May 2008.

[168] B. L. Sturm. An analysis of the GTZAN music genre dataset. In *MIRUM-2012*, 2012.

[169] Z. Y. Su and T. Wu. Multifractal analyses of music sequences. *Physica D*, 221:188 – 194, 2006.

[170] Z. Y. Su and T. Wu. Music walk, fractal geometry in music. *Physica A*, 380:418 – 428, 2007.

[171] S. R. Subramanya and A. Youssef. Wavelet-based indexing of audio data in audio/multimedia databases. In *Proc. MultiMedia Database Management Systems*, 1998.

[172] S. Sukittanon, L. E. Atlas, and J. W. Pitton. Modulation scale analysis for content identification. Technical Report 3, Dept. Electr. Engineer. Univ. of Washington, 2003.

[173] R. B. Sussman and M. Kahrs. Analysis and resynthesis of musical instrument sounds using energy separation. In *Proc. Int'l. Conf. Acoustics, Speech and Signal Processing (ICASSP-96)*, 1996.

[174] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In W. J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*, volume 15. NATO Advanced Study Institute, Series D, Boston, MA: Kluwer, Jult 1989.

[175] Φ. Τερζάκης. *Σημειώσεις για μιαν Ανθρωπολογία της Μουσικής*. Πρίσμα, 1990.

[176] W. Torres and T. Quatieri. Estimation of modulation based on FM-to-AM Transduction: Two-sinusoid case. *IEEE Trans. on Signal Process.*, 47(11):3084–3097, 1999.

[177] C.-F. Tsai. Bag-of-Words representation in image annotation: A review. *International Scholarly Research Network ISRN Artificial Intelligence*, 2012, 2012.

[178] R. Typke, F. Wiering, and R. C. Veltkamp. A survey of music information retrieval systems. In *Proc. Int'l. Symp. Music Information Retrieval (ISMIR-05)*, 2005.

[179] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Process.*, 10(5):293–302, July 2002.

[180] G. Tzanetakis, G. Essl, and P. Cook. Audio analysis using the Discrete Wavelet Transform. *IEEE Trans. on Speech and Audio Process.*, 2002.

[181] University of Iowa Musical Instrument Sample Database. [ONLINE], Available: http://theremin.music.uiowa.edu/.

[182] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[183] D. Västfjäll. Emotion induction through music: A review of the musical mood induction procedure. *Musicae Scientae. Special Issue: Current trends in the study of music and emotion*, pages 173–212, 2002.

[184] R. F. Voss and J. Clarke. 1/f noise in music and speech. *Nature*, 258:317 – 318, Nov. 1975.

[185] R. F. Voss and J. Clarke. "1/f noise" in music: Music from 1/f noise. *J. Acoust. Soc. Am.*, 63(1):258 – 263, 1978.

[186] P. M. Warren. *Auditory Perception*. Cambridge University Press, 3rd edition, 2008.

[187] J. Wei, C. Courtenay, and A. C. Loui. Automatic consumer video summarization by audio and visual analysis. In *Multimedia and Expo (ICME-11)*, pages 1–6, Jul. 2011.

[188] P. White. *Basic MIDI*. SMT, Technology Sanctuary Publishing, 2003.

[189] C. Xu, N. C. Maddage, and X. Shao. Automatic music classification and summarization. *IEEE Trans. on Speech and Audio Process.*, 13(3):441–450, 2005.

[190] G. Xu, N. C. Maddage, X. Shao, F. Cao, and Q. Tian. Musical genre classification using Support Vector Machines. In *Proc. Int'l. Conf. Acoustics, Speech and Signal Process. (ICASSP-03)*, 2003.

[191] Univ. of INRIA YAEL Library. [ONLINE], https://gforge.inria.fr/projects/yael/.

[192] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating Bag-of-Visual-Words representations in scene classification. In *ACM SIGMM Workshop on Multimedia Information Retrieval (MIR), in conjunction with ACM Multimedia*, 2007.

[193] S. Young et al. *The HTK Book (for HTK Version 3.4)*. Copyright © 2001-2009 Cambridge Univ. Engineer. Department, 2009.

[194] R. J. Zatorre, P. Belin, and V. B. Penhune. Structure and function of auditory cortex: music and speech. *TRENDS in Cognitive Sciences*, 6(1):37–46, Jan. 2002.

[195] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: An in-depth study. Technical report, INRIA RR-5737, 2005.

[196] Y. Zhang, R. Jin, and Z.-H. Zhou. Understanding Bag-of-Words model: a statistical framework. *Int'l. Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, Dec. 2010.

[197] A. Zlatintsi and P. Maragos. Musical instruments signal analysis and recognition using fractal features. In *Proc. 19th European Signal Processing Conf. (EUSIPCO-2011)*, Barcelona, Spain, Aug.-Sep. 2011.

[198] A. Zlatintsi and P. Maragos. AM-FM modulation features for music instrument signal analysis and recogntion. In *Proc. 20th European Signal Processing Conf. (EUSIPCO-2012)*, Bucharest, Romania, Aug. 2012.

[199] A. Zlatintsi and P. Maragos. Multiscale fractal analysis of musical instrument signals with application to recognition. *IEEE Trans. on Audio, Speech and Language Process.*, 21(4):737–748, 2013.

[200] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos. A saliency-based approach to audio event detection and summarization. In *Proc. 20th European Signal Processing Conf. (EUSIPCO-2012)*, Bucharest, Romania, Aug. 2012.

# Παράρτημα Α΄

# Δημοσιεύσεις Διατριβής

Αποτελέσματα της διδακτορικής μας έρευνας έχουν δημοσιευθεί σε διεθνώς αναγνωρισμένα περιοδικά και συνέδρια με κριτές. Ακολουθεί πλήρης κατάλογος των σχετικών δημοσιεύσεων. Ηλεκτρονικά αντίτυπα είναι διαθέσιμα από την ιστοσελίδα: http://cvsp.cs.ntua.gr/nancy/.

## Δημοσιεύσεις σε Διεθνή Περιοδικά

J1. **A. Zlatintsi** and P. Maragos. Multiscale Fractal Analysis of Musical Instrument Signals with Application to Recognition. *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 21, No. 4, pp. 737-748, Apr. 2013.

J2. 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. on Multimedia*, Vol. 15, No. 7, pp.1553-1568, Nov. 2013.

## Δημοσιεύσεις σε Διεθνή Συνέδρια με Κρίση στο Πλήρες Κείμενο

C1. G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, **A. Zlatintsi** and Y. Avrithis. Movie Summarization Based on Audiovisual Saliency Detection. In *Proc. of IEEE International Conference on Image Processing (ICIP-08)*, San Diego, CA, U.S.A., pages 2528-2531, Oct. 12-15, 2008.

C2. D. Spachos, **A. Zlatintsi**, V. Moschou, P. Antonopoulos, E. Benetos, M. Kotti, K. Tzimouli, C. Kotropoulos, N. Nikolaidis, P. Maragos and I. Pitas. MUSCLE

Movie Database: A Multimodal Corpus With Rich Annotation For Dialogue And Saliency Detection. In *Proc. of International Conference on Language Resources and Evaluation, (LREC-08)*, Marrakech, Morocco, May 2008.

C3. G. Evangelopoulos, **A. Zlatintsi**, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos and Y. Avrithis. Video Event Detection and Summarization Using Audio, Visual and Text Saliency. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-09)*, Taipei, Taiwan, Apr. 19-24, 2009.

C4. N. Malandrakis and A. Potamianos and G. Evangelopoulos and **A. Zlatintsi**. A Supervised Approach to Movie Emotion Tracking. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-11)*, pp. 2376-2379, Prague, Czech Republic, May 2011.

C5. **A. Zlatintsi** and P. Maragos. Musical Instruments Signal Analysis and Recognition Using Fractal Features. In *Proc. 19th European Signal Processing Conference (EUSIPCO-11)*, Barcelona, Spain, Aug.-Sep. 2011.

C6. **A. Zlatintsi** and P. Maragos. AM-FM Modulation Features for Music Instrument Signal Analysis and Recognition. In *Proc. 20th European Signal Processing Conference (EUSIPCO-12)*, Bucharest, Romania, Aug. 2012.

C7. **A. Zlatintsi**, P. Maragos, A. Potamianos and G. Evangelopoulos. A Saliency-Based Approach to Audio Event Detection and Summarization. In *Proc. 20th European Signal Processing Conference (EUSIPCO-12)*, Bucharest, Romania, Aug. 2012.

# Αθανασία Ζλατίντση

## *Βιογραφικό Σημείωμα*

*Σχ. Ηλεκτρ. Μηχ. & Μηχ. Υπολ.,*
*Εθνικό Μετσόβιο Πολυτεχνείο*
*email:nzlat@cs.ntua.gr*
*cvsp.cs.ntua.gr/nancy*

## Εκπαίδευση

2007–2013 **Δίπλωμα Διδάκτορα Μηχανικού**, *Εθνικό Μετσόβιο Πολυτεχνείο*, Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών.

* Ερευνητικό πεδίο: Ανάλυση και Επεξεργασία Μουσικών Σημάτων και Εφαρμογές Αναγνώρισης.
* Θέμα Διατριβής: Επεξεργασία Σημάτων Μουσικής και Εφαρμογές Αναγνώρισης (Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος ΙΙ).
* Επιβλέπων: Καθ. Πέτρος Μαραγκός (http:cvsp.cs.ntua.gr/maragos).

2004–2006 **Master of Science in Media Engineering**, *Royal Institute of Technology (KTH), Stockholm, Sweden*, School of Media Technology (Computer Science and Communication).

* Ερευνητικό πεδίο: Music Acoustics.
* Θέμα Διπλωματικής:When the Clarinet Sounds Bad. Identification study. (Unofficial part of the European project VEMUS, financed by the European Commission under the Information Society Technologies Programme.)
* Κατεύθυνση σπουδών: Sound and Moving Picture.
* Επιβλέπων: Kjetil Falkenberg Hansen, Anders Askenfelt.

2001–2004 **Bachelor of Science in Media Engineering**, *Royal Institute of Technology (KTH), Stockholm, Sweden*, School of Media Technology.

* Ερευνητικό πεδίο: Αλληλεπίδραση Ανθρώπου-Υπολογιστή.
* Θέμα Διπλωματικής: I want to say something. Digitalbook as textbook and teaching aid in school, evaluation. (Τίτλος στα σουηδικά: "Jag vill säga något". Digitalbok som läromedel I skolan, utvärdering.)

2000–2001 **Σπουδές Μουσικολογίας**, *Stockholm University, Stockholm, Sweden*, Department of Musicology and Performance Studies.

* Μαθήματα: Musicology I,II, Film Music I, II

1986–1996 **Λοιπές Μουσικές Σπουδές**.

## Ερευνητική Εμπειρία

2007–2013 **Διδακτορική Ερευνήτρια**, *Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο.*

Μέλος της ερευνητικής ομάδας Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σήματος (CVSP). Συμμετοχή σε Ελληνικά και Ευρωπαϊκά ερευνητικά προγράμματα στις περιοχές της ψηφιακής ανάλυσης σήματος και πολυτροπικής επεξεργασίας.

* Συμμετοχή στο Πρόγραμμα Βασικής Έρευνας ΗΡΑΚΛΕΙΤΟΣ ΙΙ.
* Συμμετοχή στο Ερευνητικό Πρόγραμμα Αριστείας MUSCLE.

2006 **Προπτυχιακή Ερευνήτρια**, *Royal Institute of Technology (KTH), Stockholm, Sweden, Department of Speech, Music and Hearing, School of Computer Science and Communication.*
Συμμετοχή στο Ευρωπαϊκό ερευνητικό πρόγραμμα VEMUS κατά τη διάρκεια της Πτυχιακής Εργασίας, "When the Clarinet Sounds Bad. identification study".

## Τρέχοντα Ερευνητικά Ενδιαφέροντα

* Επεξεργασία σημάτων μουσικής και αναγνώριση των δομών, του είδους και των εκφραστικών ιδιοτήτων τους.
* Επεξεργασία σημάτων και εξαγωγή χαρακτηριστικών για εφαρμογές ανάλυσης και αναγνώρισης του περιεχομένου τους, ανίχνευσης σημαντικών γεγονότων και δημιουργίας περιλήψεων.

## Υποτροφίες

2011–2013 **Ερευνητικό Χρηματοδοτούμενο Έργο: ΗΡΑΚΛΕΙΤΟΣ ΙΙ**, Ερευνητική υποτροφία συγχρηματοδοτούμενη από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινοτικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ).

2009–2011 **Εθνικό Μετσόβιο Πολυτεχνείο**, Μονοετής ερευνητική υποτροφία για διδακτορικές σπουδές.

## Διδακτική Εμπειρία

2010–2013 **Μεταπτυχιακή βοηθός μαθήματος**, *Όραση Υπολογιστών (μεταπτυχιακό και προπτυχιακό μάθημα 8ου εξαμήνου)*, Σχ. Ηλεκτρ. Μηχ. & Μηχ. Υπολ., ΕΜΠ. Επικουρικό έργο, εργαστήρια, προετοιμασία υλικού.

2009–2012 **Μεταπτυχιακή βοηθός μαθήματος**, *Ψηφιακή Επεξεργασία Σήματος (προπτυχιακό μάθημα 6ου εξαμήνου)*, Σχ. Ηλεκτρ. Μηχ. & Μηχ. Υπολ., ΕΜΠ, Επικουρικό έργο, εργαστήρια, ασκήσεις, προετοιμασία υλικού, παρουσιάσεις.

## Μέλος Επιστημονικών Οργανώσεων

2012–τώρα **Institute of Electrical and Electronics Engineers (IEEE)**.

2010–τώρα **Τεχνικό Επιμελητήριο Ελλάδος (ΤΕΕ)**.

## Γλώσσες

| | | |
|---|---|---|
| Αγγλικά | **Άριστα** | *First Certificate in English - Univ. Of Cambridge* |
| Σουηδικά | **Άριστα** | *TISUS* |
| Ελληνικά | **Μητρική** | |