ΕΛΛΗΝΙΚΟ ΜΕΣΟΓΕΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΜΟΥΣΙΚΗΣ ΚΑΙ ΟΠΤΟΑΚΟΥΣΤΙΚΩΝ ΤΕΧΝΟΛΟΓΙΩΝ

ΠΑΡΑΡΤΗΜΑ ΡΕΘΥΜΝΟΥ (ΚΡΗΤΗ)

ΤΜΗΜΑ ΜΟΥΣΙΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ

ΑΚΟΥΣΤΙΚΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«ΑΥΤΟΜΑΤΗ ΑΝΑΓΝΩΡΙΣΗ ΚΑΛΛΙΤΕΧΝΗ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ»

ΣΠΟΥΔΑΣΤΗΣ

ΕΥΤΥΧΙΟΣ ΒΑΒΑΓΙΑΚΗΣ

Α.Μ.: 1161

ΕΠΙΒΛΕΨΗ ΠΤΥΧΙΑΚΗΣ

ΔΡ. ΖΕΡΒΑΣ ΠΑΝΑΓΙΩΤΗΣ

ΡΕΘΥΜΝΟ 2020

Ευρετήριο περιεχομένων

Ευρετήριο πινάκων

Ευρετήριο σχημάτων

Πρόλογος

Πληροφορίες απαραίτητες για την κατανόηση του αντικειμένου της εργασίας, καθώς επίσης και για την εκπόνηση του πειραματικού και του θεωρητικού μέρους της, διατέθηκαν από το Δρ. Παναγιώτη Ζέρβα. Επίσης, παραπέμφθηκα σε βιβλιογραφικές αναφορές από τον ίδιο, μου προτάθηκαν εργαλεία για την περάτωση των πειραμάτων, μου δόθηκαν κατευθυντήριες οδοί και εναλλακτικές λύσεις στα προβλήματα που προέκυψαν κατά τη διεξαγωγή της εργασίας και μου δόθηκαν απαντήσεις σε απορίες που γεννήθηκαν κατά τη διάρκεια της ερευνητικής περιόδου. Τέλος, για την περισυλλογή πληροφοριών και τη διεξαγωγή πειραμάτων χρησιμοποιήθηκε ο εξοπλισμός της αίθουσας του εργαστηρίου Μουσικής Διάδρασης και Πολυφωνίας.

Περίληψη

Η διάδοση της ψηφιακής μουσικής πληροφορίας επέφερε τη δημιουργία τεράστιων μουσικών βάσεων δεδομένων οι οποίες χρησιμοποιούνται για προσωπικούς αλλά και επαγγελματικούς σκοπούς. Έτσι, δημιουργήθηκε η ανάγκη για γρήγορη και αποδοτική ανάκτηση πληροφορίας σε αυτές τις συλλογές. Αυτή την ανάγκη ήρθαν να καλύψουν τα συστήματα αυτόματης ανάκτησης μουσικής πληροφορίας. Ένα από τα πιο δημοφιλή στοιχεία προς αναζήτηση από τους χρήστες τέτοιων συλλογών είναι αυτό του καλλιτέχνη που ερμηνεύει το εκάστοτε μουσικό κομμάτι. Για αυτό το λόγο η αυτόματη αναγνώριση καλλιτέχνη αποτελεί ένα από τα πιο σημαντικά πεδία έρευνας στον τομέα της αυτόματης ανάκτησης μουσικής πληροφορίας. Σκοπός της συγκεκριμένης πτυχιακής εργασίας λοιπόν, είναι η ανάπτυξη ενός μοντέλου αυτόματης αναγνώρισης καλλιτέχνη με χρήση μεθόδων μηχανικής μάθησης. Μέσω αυτής της εργασίας επιδιώκεται η εύρεση και η ανάπτυξη του καλύτερου δυνατού μοντέλου για την κατά το δυνατόν ακριβέστερη και αποδοτικότερη κατηγοριοποίηση μουσικών έργων στους καλλιτέχνες που τα ερμηνεύουν ή τα εκτελούν.

Για την ανάπτυξη του μοντέλου και τις δοκιμές δημιουργήθηκαν και χρησιμοποιήθηκαν τέσσερις βάσεις δεδομένων με ηχητικά δείγματα από μουσικά έργα δεκαπέντε δημοφιλών καλλιτεχνών και συγκροτημάτων. Οι μέθοδοι και τα εργαλεία μηχανικής μάθησης (machine learning) μας δίνουν τη δυνατότητα να αναγνωρίζουμε καλλιτέχνες. Μέσω της μηχανικής μάθησης λοιπόν, και των μεθόδων της, αλλά κυρίως μέσω της διαδικασίας εξόρυξης δεδομένων (data mining), εξάγουμε συγκεκριμένα ηχητικά χαρακτηριστικά, τα οποία στη συνέχεια ταξινομούμε με χρήση αλγορίθμων αναγνώρισης διανυσμάτων και προτύπων. Με αυτόν τον τρόπο αναγνωρίζουμε αυτόματα τα χαρακτηριστικά αυτά, και κατ' επέκταση τους καλλιτέχνες που επιθυμούμε να κατηγοριοποιήσουμε.

Λέξεις-κλειδιά

Καλλιτέχνες, αυτόματη αναγνώριση καλλιτέχνη, εξαγωγή χαρακτηριστικών, MFCC, GFCC, LPC, LPCC, DMFCC, DGFCC, DLPC, DLPCC, ισχυρή φασματική κορυφή, φασματική ροή, φασματική αντίθεση, συντελεστές δέλτα, μηχανές υποστήριξης διανυσμάτων SVM, πολυστρωματικά νευρωνικά δίκτυα MLP, αλγόριθμος C4.5, Essentia, Weka, κατηγοριοποίηση με και χωρίς επιτήρηση, χρονικός διαχωρισμός

Abstract

The dissemination of digital musical information brought upon the creation of huge music databases which are used for both personal and professional reasons. Thus, the need was created for fast and reliable information retrieval in these collections. This need was fulfilled by the automatic music information retrieval systems. One of the most popular search indexes from the users of collections of this kind, is that of the artist who sings in each song. For this reason the automatic artist recognition is one of the most important fields of research into the field of automatic music information retrieval. So, the purpose of this thesis is the development of a model for automatic artist recognition by using machine learning methods and the - as much as possible - precise and reliable classification of musical pieces into the artists who either sing or play them.

Also for the development of the model and the testing was created and used four databases with sound samples from musical pieces of 15 artists and bands. The methods and the machine learning tools give us the opportunity to recognize artists. So, via machine learning and its methods, but mainly via the data mining process, we extract specific sound features which, consequently, we classify by using vector and pattern recognition algorithms. This way, we recognize automatically these features and additionally, the artists we want to classify.

Kew words

Artists, automatic artist recognition, feature extraction, MFCC, GFCC, LPC, LPCC, DMFCC, DGFCC, DLPC, DLPCC, strong spectral peak, spectral flux, spectral contrast, DCT, support vector machines SVM, multilayer perceptron MLP, C4.5 algorithm, Essentia, Weka, supervised and unsupervised learning, time decomposition and segmentation

Κεφάλαιο 1ο: Αυτόματη ανάκτηση μουσικής πληροφορίας

1.1 Εισαγωγικά στοιχεία

Η έλευση της ψηφιακής μουσικής τεχνολογίας είχε τεράστιο αντίκτυπο στην εξέλιξη της μουσικής βιομηχανίας όπως την ξέρουμε σήμερα. Η εξέλιξη των μέσων αποθήκευσης της μουσικής, η ραγδαία διάδοση της pop μουσικής, της ψηφιακής ηχογράφησης και των ψηφιακών μορφών της μουσικής, όπως για παράδειγμα των μουσικών συμπαγών δίσκων (CD), καθώς και των MP3, τα οποία είναι ανά πάσα στιγμή διαθέσιμα προς μεταφόρτωση στο Διαδίκτυο από υπηρεσίες παροχής μουσικής, έχουν επιφέρει την ανάπτυξη τεράστιων ψηφιακών μουσικών βιβλιοθηκών, για επαγγελματική, αλλά και για προσωπική χρήση. Αυτή η διάδοση οφείλεται, φυσικά, στην εξέλιξη της τεχνολογίας, όσον αφορά τους ηλεκτρονικούς υπολογιστές και στο Διαδίκτυο. Παρατηρείται επίσης, ότι όλο και περισσότεροι χρήστες διαθέτουν πλέον προσωπικά αρχεία ψηφιακής μουσικής στους προσωπικούς τους υπολογιστές, στις φορητές συσκευές αναπαραγωγής πολυμέσων τους, ή ακόμα και στα κινητά τους τηλέφωνα. Έτσι, τα μεγέθη αυτών των συλλογών ολοένα και αυξάνονται και η ανάγκη για την ανάπτυξη τεχνολογιών, όπως αυτόματων συστημάτων για την κατά το δυνατόν αποδοτικότερη κατηγοριοποίηση και ανάκτησή τέτοιων συλλογών είναι πια επιτακτική. Επίσης, αυτού του είδους τα συστήματα παρέχουν στο χρήστη κάποιες πανίσχυρες λειτουργίες, όπως την αναζήτηση μουσικού περιεχομένου, καθώς και την περιήγηση σε αυτό. Έτσι, τα συστήματα ανάκτησης μουσικής, βάσει περιεχομένου, έχουν μετατραπεί σε ένα εξαιρετικά ελκυστικό ζήτημα.

1.2 Ανάκτηση ηχητικής πληροφορίας

Ο όρος ανάκτηση ηχητικής πληροφορίας (AIR) αναφέρεται στην ανάλυση, εξαγωγή και σύγκριση χαρακτηριστικών σε ένα ηχητικό κομμάτι, το οποίο έχει διαχωριστεί σε ηχητικά αποσπάσματα μικρότερης διάρκειας. Η ανάκτηση μουσικής πληροφορίας (MIR) από την άλλη, αποτελεί υποκατηγορία της AIR που αφορά εφαρμογές αναγνώρισης τραγουδιστή και χροιάς, διαχωρισμού τραγουδιστής φωνής και εξαγωγής μελωδίας από μουσικά κομμάτια, εντοπισμού νότας και τονικού ύψους, μεταγραφής μελωδίας και άλλες (Deshmukh & Bhirud, 2012).

Ο τομέας της ανάκτησης μουσικής πληροφορίας έχει εμπλουτιστεί πολύ κατά τις τελευταίες τρεις δεκαετίες με βελτιωμένους αλγόριθμους και καινοτόμες μεθοδολογίες, οι οποίες χρησιμοποιούνται σε διάφορες εργασίες σχετικές με μουσική και πληροφορία. Έπειτα

από εκτενή έρευνα πάνω στη φωνητική αναγνώριση έχουν επιτευχθεί κάποιες αξιόλογες εφαρμογές, οι οποίες δεν εξετάζουν όλες τη μουσική από την ίδια σκοπιά. Άλλες κάνουν χρήση αυτόματων ταξινομητών ή αλγόριθμων αναγνώρισης μελωδίας (Durey & Clements, 2002), (Akeroyd, et al., 2001), άλλες μουσικών οργάνων (Herrera, et al., 2000), (Eronen, 2003), άλλες είδους (Τζανετάκης & Cook, 2002), (Xu, et al., 2003), άλλες καλλιτέχνη ή τραγουδιστή (Berenzweig, et al., 2002), (Kim & Whitman, 2002), (Liu & Huang, 2002) και άλλες άλλων στοιχείων των μουσικών έργων (Byrd & Crawford, 2002), (Hsu, et al., 2001).

1.3 Ηχητικά χαρακτηριστικά

Οι ηχητικοί περιγραφείς αποτελούν ειδικά στοιχεία ή χαρακτηριστικά γνωρίσματα των ηχητικών αποσπασμάτων. Η αναγνώριση αυτών των ηχητικών περιγραφέων αποτελεί το πρώτο βήμα στην ανάλυση ενός ηχητικού δείγματος. Ένα σύνολο κατάλληλων ηχητικών περιγραφέων διευκολύνει τη διαδικασία εντοπισμού ενός κομματιού σε μία μεγάλη μουσική βάση δεδομένων.

Κατά τον Peeters (Peeters, 2004), τα ηχητικά χαρακτηριστικά μπορούν να εξεταστούν ως προς:

- τη δυναμική τους,
- τη χρονική έκταση του περιγραφέα (παγκόσμιος ή στιγμιαίος),
- την αφαιρετικότητά τους,
- τη διαδικασία εξαγωγής τους.

Ο Peeters επίσης, για την εργασία του CUICADO, δημιούργησε ένα σύστημα ταξινόμησης ως προς την είσοδο που δίνεται και την έξοδο που απαιτείται. Τα ηχητικά χαρακτηριστικά χωρίστηκαν σε:

- παγκόσμια προσωρινά χαρακτηριστικά, τα οποία είναι ο χρόνος λογαριθμικής ατάκας, η αύξηση, η μείωση, το κεντροειδές και η διάρκεια,
- στιγμιαία προσωρινά χαρακτηριστικά, τα οποία περιλαμβάνουν την αυτοσυσχέτιση του σήματος, το λόγος μηδενικής διέλευσης, τα χαρακτηριστικά που σχετίζονται με την ενέργεια και τα φασματικά χαρακτηριστικά
- και περιγραφείς παγκόσμιου φασματικού σχήματος, όπου εντάσσονται τα MFCC, DMFCC, DDMFCC, καθώς και τα αρμονικά και αντιληπτικά χαρακτηριστικά.

Για την εξαγωγή χαρακτηριστικών είναι απαραίτητο να προηγηθεί μία προεπεξεργασία προκειμένου να αναπαρασταθεί επαρκώς το σήμα για μετέπειτα επεξεργασία της εξαγωγής περιγραφέων.

Στη δημοσίευση (Tzanetakis, n.d.) παρουσιάζονται δύο τρόποι ταξινόμησης των χαρακτηριστικών, ως προς την υπολογιστική προσέγγιση και ως προς τα στοιχεία που περιγράφουν τον ήχο. Στην πρώτη κατηγορία ανήκουν χαρακτηριστικά όπως τα WT και τα STFT, ενώ στη δεύτερη ανήκουν χαρακτηριστικά όπως το ηχόχρωμα, ο ρυθμός και το τονικό ύψος. Έτσι, υπάρχουν δύο προσεγγίσεις κατηγοριοποίησης ηχητικών περιγραφέων, ως προς τον τρόπο εξαγωγής και υπολογισμού τους και ως προς τα κοινά τους στοιχεία. Κάποιοι περιγραφείς ωστόσο δεν είναι δυνατό να κατηγοριοποιηθούν και για αυτό το λόγο δεν υπάρχει ακόμα κάποιος συγκεκριμένος τρόπος ταξινόμησης ή κάποια διεθνής συμφωνία για την ταξινόμηση των χαρακτηριστικών.

Ασυμφωνίες παρουσιάζονται επίσης, μεταξύ ερευνητών, ως προς την ένταξη χαρακτηριστικών σε κάποια συγκεκριμένη ομάδα. Ένα χαρακτηριστικό παράδειγμα τέτοιας ασυμφωνίας αποτελεί ο ρυθμός διέλευσης μηδενικού σημείου (ZCR), το οποίο οι Esmaili, Krishnan και Raahemifar (Esmaili, et al., Μάιος 2004) το εντάσσουν στην κατηγορία των χρονικών χαρακτηριστικών, ενώ οι Lu, Zhang και Li (Lu, et al., Απρ. 2003) το εντάσσουν στην κατηγορία των αντιληπτικών χαρακτηριστικών. Αφού λοιπόν η κατηγοριοποίηση ηχητικών περιγραφέων δεν είναι δυνατό να πραγματοποιηθεί με ένα γενικευμένο τρόπο, είναι καλύτερο να πραγματοποιείται ανά περίπτωση.

Κεφάλαιο 2ο: Αναγνώριση καλλιτέχνη

2.1 Αυτόματη αναγνώριση καλλιτέχνη

Η αναγνώριση ανθρώπινης ομιλίας αποτελεί ένα σημαντικό ερευνητικό ζήτημα στον τομέα της ανάκτησης ηχητικής πληροφορίας. Εξίσου επίμαχα είναι τα ζητήματα της αναγνώρισης καλλιτέχνη και της ομοιότητας μεταξύ καλλιτεχνών. Η αυξανόμενη ανάγκη για ανάκτηση μουσικής πληροφορίας επέφερε τη διεύρυνση του ζητήματος της αναγνώρισης ομιλητή (SPID) στο ζήτημα της αναγνώρισης τραγουδιστή (SNID) (Tsai & Wang, 2006), (Nwe & Li, 2007). Η αυτόματη αναγνώριση του ερμηνευτή ενός μουσικού κομματιού είναι μία εκ των τεχνολογιών που αναπτύχθηκαν για την κατηγοριοποίηση και την ανάκτηση μουσικών συλλογών και αποτελεί μία εργασία παρεμφερή της ανθρώπινης ικανότητας αναγνώρισης ομιλίας ή τραγουδιστής φωνής. Αυτή περιγράφεται ως η αναγνώριση του ερμηνευτή ενός μουσικού έργου, μέσω της ανάλυσης των ηχητικών χαρακτηριστικών ενός μουσικού σήματος. Μάλιστα, αποτελεί την πιο σημαντική εφαρμογή ανάκτησης μουσικής πληροφορίας (MIR). Αυτή η ικανότητα ενός μουσικού συστήματος δίνει τη δυνατότητα στο χρήστη να λαμβάνει εύκολα πληροφορίες σχετικές με τον τραγουδιστή ενός οποιουδήποτε κομματιού, ή ακόμα και να ανακτά όλα τα κομμάτια τα οποία ερμηνεύει ένας συγκεκριμένος τραγουδιστής σε μία παρεχόμενη βάση δεδομένων. Επιπλέον, μέσω αυτής της τεχνολογίας δίνεται η δυνατότητα να ομαδοποιηθούν κομμάτια τα οποία ερμηνεύουν τραγουδιστές με παρόμοιες φωνές ή ακόμα και να αναζητηθούν κομμάτια, οι φωνές των ερμηνευτών των οποίων παρουσιάζουν ομοιότητες με τη φωνή του ερμηνευτή ενός συγκεκριμένου τραγούδιού.

Ο γενικός όρος «αναγνώριση τραγουδιστή» χρησιμοποιείται για να περιγράψει όλες τις επιμέρους διαδικασίες, οι οποίες χρησιμοποιούν χαρακτηριστικά της φωνής του ερμηνευτή, προκειμένου να διαχωρίσουν και να ομαδοποιήσουν μουσικά δεδομένα. Υπάρχουν πολλές τέτοιες προσεγγίσεις. Κάποιες από αυτές είναι: η αναγνώριση ερμηνευτή (SID), η αναγνώριση ερμηνευτή-στόχου (TSD) και η ανίχνευση ερμηνευτή-στόχου (TST). Ο όρος SID αναφέρεται στη διαδικασία αναγνώρισης του ερμηνευτή, ανάμεσα από άλλους, οι οποίοι ερμήνευσαν ένα συγκεκριμένο μέρος ενός τραγουδιού. Η διαδικασία αυτή περιλαμβάνει μία απόφαση νιοστής τάξης, όπου το N είναι το πλήθος των ερμηνευτών που συμμετείχαν στη διαδικασία. Βασική προϋπόθεση για την εκτέλεση αυτής της διαδικασίας, είναι να προηγηθεί η επισημείωση των αποσπασμάτων με χρήση ετικετών. Η TSD, από την άλλη, χρησιμοποιείται για τις περιπτώσεις που επιθυμούμε να εξακριβώσουμε εάν ένας επιλεγμένος τραγουδιστής ερμηνεύει ένα μουσικό

απόσπασμα. Στην ουσία αποτελεί μία ταξινόμηση δύο κλάσεων ή δυαδική, όπου η μία κλάση περιλαμβάνει μουσικά δεδομένα —συμπεριλαμβανομένων και αυτών που εμπεριέχουν τη φωνή του τραγουδιστή-στόχου— και η άλλη, μουσική η οποία εκτελείται αποκλειστικά από άλλους καλλιτέχνες (ο τραγουδιστής-στόχος δε συμπεριλαμβάνεται).

Η τραγουδιστή φωνή αποτελεί το κυρίαρχο στοιχείο κάθε τραγουδιού και προσελκύει αμέσως το ενδιαφέρον του ακροατή. Τα φωνητικά μεταφέρουν πληροφορίες ταυτοποίησης του τραγουδιστή και γλωσσικές πληροφορίες. Έτσι, οι ακροατές είναι ικανοί να αναγνωρίσουν τη φωνή ενός τραγουδιστή που γνωρίζουν και να τη διακρίνουν από παρόμοιες φωνές, ακούγοντας μόλις ένα σύντομο απόσπασμα ενός κομματιού, καθώς και να καταγράψουν στίχους τραγουδιών. Πιστεύεται ότι με κατάλληλη εξαγωγή και ανάλυση ηχητικών χαρακτηριστικών ένα αυτοματοποιημένο σύστημα είναι ικανό να επιτύχει επίσης, σε ένα βαθμό, αναγνώριση τραγουδιστή. Ωστόσο, παρόλο που είναι αρκετά εύκολο για τους ανθρώπους να αναγνωρίσουν έναν τραγουδιστή, η εκτέλεση μίας αξιόπιστης εργασίας αναγνώρισης τραγουδιστή από ένα σύστημα μηχανής ακρόασης, παρουσιάζει ιδιαίτερες δυσκολίες. Ακόμα και σε σήματα καθαρά, απαλλαγμένα από συνοδεία οργάνων και περιβάλλοντα θόρυβο, τα απλά συχνοτικά χαρακτηριστικά, αλλά και τα χαρακτηριστικά πεδίου χρόνου, δεν είναι ικανά να αποδώσουν εύκολα ένα μοναδικό ηχητικό αποτύπωμα. Το πρόβλημα γίνεται μάλιστα ακόμα πιο περίπλοκο εάν αναλογιστούμε ότι στις περισσότερες μουσικές εκτελέσεις ή ηχογραφήσεις, η φωνή περιπλέκεται με ένα μείγμα από άλλους ήχους. Αυτοί οι ήχοι είναι συνήθως μία εκκωφαντική μουσική συνοδεία, η οποία είναι ιδιαίτερα έντονη στην ποπ μουσική και καθιστά την εξαγωγή χαρακτηριστικών από κομμάτια που παρουσιάζουν αυτή την ιδιαιτερότητα, μία πολύ επίπονη διαδικασία. Σε κάποιες περιπτώσεις μάλιστα καθίσταται ανέφικτη η αλίευση σόλο φωνητικών (χωρίς συνοδεία μουσικής) για άμεση εξαγωγή χαρακτηριστικών, κάτι το οποίο είναι απόλυτα εφικτό στην αναγνώριση ομιλητή. Επίσης, η πολυπλοκότητα των μίξεων δυσχεραίνει την αναγνώριση φασματικών χαρακτηριστικών που σχετίζονται με την τραγουδιστή φωνή. Για όλους τους παραπάνω λόγους λοιπόν, οι αυτοματοποιημένες μέθοδοι που ασχολούνται με εργασίες σαν αυτές που περιγράφηκαν δεν έχουν καταφέρει να αγγίξουν την απόδοση της αντιληπτικής ικανότητας του ανθρώπου.

Η ανθρώπινη φωνή είναι ένα εξαιρετικά πολύπλοκο μουσικό όργανο και αδιαμφισβήτητα ένα από τα εκφραστικότερα. Η φωνή αποτελεί, επίσης, το παλαιότερο μουσικό όργανο. Ως εκ τούτου, η ανθρώπινη φυσιολογία και τα ακουστικά μας όργανα έχουν

εξελιχθεί με τέτοιο τρόπο, ώστε να διαθέτουν υψηλή ευαισθησία στην ανθρώπινη φωνή. Ο μουσικός (ερμηνευτής) δημιουργεί έναν εξαιρετικά ισχυρό δεσμό με το όργανό του (φωνή). Το συγκεκριμένο μουσικό όργανο (φωνή) αποτελείται από σάρκα και οστά (ανθρώπινο σώμα), εν αντιθέσει με τα υπόλοιπα μουσικά όργανα, τα οποία συνήθως αποτελούνται από ξύλο ή μέταλλο. Αφού το κάθε ανθρώπινο σώμα είναι μοναδικό, τα φυσικά χαρακτηριστικά κάθε φωνής είναι κι αυτά μοναδικά με τη σειρά τους και συντελούν στη φωνητική ταυτότητα κάθε ερμηνευτή. Οι παράγοντες οι οποίοι συμβάλλουν σε αυτή είναι τόσο βιολογικοί, όσο και τεχνικοί. Η φωνή είναι δυνατό να μεταβληθεί, εν μέρει, μέσω της εξάσκησης και των αλλαγών που επιφέρει το γήρας. Παρόλα αυτά, υπάρχουν κάποια χαρακτηριστικά τα οποία παραμένουν αμετάβλητα.

Στην πραγματικότητα, η τραγουδιστή φωνή είναι συνεχόμενη ομιλία και ο στόχος ενός συστήματος αναγνώρισης ερμηνευτή είναι ανάλογος ενός συστήματος αναγνώρισης ομιλητή (Reynolds & Rose, 1995), το οποίο έχει ως σκοπό να προσδιορίσει την ταυτότητα ενός ομιλητή, αφού η αναγνώριση καλλιτέχνη περιλαμβάνει και την αναγνώριση της ανθρώπινης φωνής. Επιπλέον, η λύση και των δύο προβλημάτων κρύβεται στον εντοπισμό και στη διαχείριση των ιδιαίτερων χαρακτηριστικών που διακρίνουν φωνές μεταξύ τους. Έχουν πραγματοποιηθεί προσπάθειες δημιουργίας συστημάτων αυτόματης αναγνώρισης ομιλητή στο παρελθόν. Αυτά τα συστήματα είχαν τη δυνατότητα να αναγνωρίζουν την ταυτότητα του ομιλητή σε ένα δοσμένο φωνητικό απόσπασμα. Παρόλα αυτά οι τεχνικές που χρησιμοποιούνται για την ανάλυση και τη σύνθεση ομιλίας δεν είναι ίδιες με αυτές της τραγουδιστής φωνής και δεν υπάρχει αξιόπιστος αλγόριθμος που να αποδίδει το ίδιο καλά και σε τραγουδιστή φωνή και σε ομιλία. Αυτό συμβαίνει επειδή η ομιλία διαφέρει σημαντικά από την τραγουδιστή φωνή, ως προς την παραγωγή και την αντίληψή της καθώς και ως προς τα χαρακτηριστικά χρόνου-συχνότητας. Η αξιοπιστία ενός αλγόριθμου ανάλυσης και σύνθεσης ομιλίας εξαρτάται από τον τύπο και τα χαρακτηριστικά γνωρίσματα της εισόδου, τη μεθοδολογία που ακολουθείται για την εξαγωγή χαρακτηριστικών και την τεχνική κατηγοριοποίησης που χρησιμοποιείται.

2.1.1 Έρευνες που σχετίζονται με την αναγνώριση τραγουδιστή

Υπάρχουν διάφορες προσεγγίσεις αναγνώρισης τραγουδιστή. Μία προσέγγιση είναι η μοντελοποίηση του μιξαρισμένου σήματος χωρίς να ακολουθείται κάποια διαδικασία διαχωρισμού των τμημάτων που περιλαμβάνουν μόνο μουσική από τα αποσπάσματα που περιλαμβάνουν φωνή. Ωστόσο, τα συστήματα τα οποία έχουν αποπειραθεί να αναγνωρίσουν

ερμηνευτές εμπορικών ηχογραφήσεων, χωρίς να προηγηθεί διαχωρισμός των φωνητικών του ερμηνευτή από το λοιπό μουσικό περιεχόμενο, δηλαδή από τα μουσικά όργανα (Kim & Whitman, 2002), (Berenzweig, et al., 2002), (Liu & Huang, 2002), έχουν γενικά κακή απόδοση. Έτσι, καταλήγουμε στη διαπίστωση ότι η λύση για τη βελτίωση των συστημάτων αναγνώρισης τραγουδιστή βρίσκεται στο διαχωρισμό των φωνητικών από τη μουσική. Άλλες προσεγγίσεις επιλέγουν πληροφορίες σχετιζόμενες με την τραγουδιστή φωνή, χρησιμοποιώντας τις μεθόδους, του διαχωρισμού τμημάτων με φωνητικά από τμήματα χωρίς (Berenzweig & Ellis, 2001), (Rao, et al., 2009), της στατιστικής εκτίμησης (Ozerov, et al., Οκτ. 2005), του μετασχηματισμού χαρακτηριστικού (Tsai & Lin, Ιούλιος 2010), του διαχωρισμού ηχητικών πηγών (Burred & Sikora, 2007), καθώς και της απομόνωσης φωνητικών και επιλογής αξιόπιστων καρέ (Fujihara, et al., 2005). Η απόδοση των συστημάτων αναγνώρισης τραγουδιστή ποικίλει, αφού στην πλειοψηφία των περιπτώσεων οι βάσεις δεδομένων επαλήθευσης διαφέρουν και αποτελούνται κυρίως από μικρό αριθμό τραγουδιστών. Οι έρευνες που αποτελούνται από βάσεις δεδομένων 8-20 τραγουδιστών, επιτυγχάνουν ακρίβεια κατηγοριοποίησης πάνω από 80%.

Υπάρχουν πολλά πεδία στην επεξεργασία σήματος που αφορούν στην τραγουδιστή φωνή, τα οποία καλύπτουν ένα εύρος θεμάτων. Αυτά τα θέματα είναι ο εντοπισμός και η αναγνώριση τραγουδιστής φωνής, η ταυτοποίηση τραγουδιστή, ο συγχρονισμός τραγουδιστής φωνής με στίχους ή παρτιτούρα, η αναγνώριση στιχουργικού περιεχομένου, ο φωνητικός διαχωρισμός, η ανάλυση και σύνθεση τραγουδιστής φωνής και οι κατηγοριοποιήσεις φωνητικής ποιότητας και χροιάς. Η επεξεργασία των τραγουδιστικών πληροφοριών αποτελεί αυτόνομο πεδίο έρευνας (Goto, et al., 2010) και σχετίζεται άμεσα με την επεξεργασία της τραγουδιστής φωνής μουσικών σημάτων. Υποκατηγορίες της επεξεργασίας τραγουδιστικών πληροφοριών αποτελούν:

- Η ακρόαση τραγουδιστής φωνής, η οποία χρησιμοποιείται για την εξαγωγή πληροφοριών που αφορούν στους στίχους, στην ταυτότητα του τραγουδιστή, στις ικανότητες του τραγουδιστή και στην οπτικοποίηση.
- Η ανάκτηση μουσικής πληροφορίας που βασίζεται στην τραγουδιστή φωνή και κάνει χρήση της χροιάς και απομαγνητοφωνήσεων τραγουδιστής φωνής και

- η σύνθεση τραγουδιστής φωνής που χρησιμοποιείται για τη σύνθεση τραγουδιστής φωνής από ομιλία ή για μετασχηματισμούς τραγουδιστής φωνής από τραγουδιστή φωνή.

Από κάθε μουσικό έργο συλλέγονται πληροφορίες, οι οποίες καθίστανται απαραίτητες για την οργάνωση, αναζήτηση και ανάκτηση μουσικών συλλογών. Κάποιες από αυτές είναι και εκείνες που αφορούν στον ερμηνευτή του εκάστοτε μουσικού έργου. Υπάρχουν ωστόσο, στοιχεία ενός μουσικού κομματιού, τα οποία δε δύνανται να προσαρτηθούν σε αυτές τις πληροφορίες. Ένα χαρακτηριστικό παράδειγμα τέτοιου στοιχείου είναι η χροιά της φωνής του ερμηνευτή. Παρόλο που δεν είναι ξεκάθαρο ποια αντιληπτικά χαρακτηριστικά προσδιορίζουν τη φωνητική ταυτότητα ενός ερμηνευτή, η χροιά εικάζεται ότι αποτελεί ένα εκ των κυριότερων. Ωστόσο, αυτή η θεωρία χρήζει περαιτέρω διερεύνησης (Erickson, et al., 2001). Το τονικό ύψος αποτελεί ένα εκ των χαρακτηριστικών που μεταβάλλονται με τις αλλαγές στο τονικό ύψος (Handel & Erickson, 2001). Έτσι, καταλήγουμε στο συμπέρασμα ότι ένα σύστημα αναγνώρισης καλλιτέχνη θα πρέπει να λαμβάνει υπόψη του τις τονικές εξαρτήσεις. Οι μεταβολές της χροιάς συναρτήσει του τονικού ύψους, αποτελούν αναπόσπαστο κομμάτι της θεωρίας της φωνητικής έκτασης (Titze, 1994), (Sundberg, 1987). Η τελευταία προτείνει διάφορους μηχανισμούς παραγωγής φωνής, εντός της φωνητικής έκτασης ενός τραγουδιστή. Ο τραγουδιστής είναι ικανός να αναπαράγει μία εξαιρετικά μεγάλη ποικιλία ήχων. Αυτό αποτελεί ένα επιπλέον εμπόδιο στη διαδικασία αναγνώρισης καλλιτέχνη, το οποίο πρέπει να ξεπεραστεί. Εν κατακλείδι, η φωνητική ταυτοποίηση αποτελεί συνάρτηση των φωνηέντων, του τονικού ύψους, της δυναμικής και της εκπαίδευσης του ερμηνευτή.

Τα συστήματα αναγνώρισης τραγουδιστή (Kim & Whitman, 2002), (Liu & Huang, 2002), (Zhang, 2003), (Bartsch & Wakefield, 2004), (Tsai, et al., 2003) ή καλλιτέχνη (Whitman, et al., 2001), (Berenzweig, et al., 2002), δημιουργήθηκαν επίσης, προκειμένου να ελαττώσουν, ή ακόμα και να αντικαταστήσουν, τις εξαιρετικά χρονοβόρες, χειροκίνητες διαδικασίες, εγγραφής δεδομένων σε αρχεία, οι οποίες σε κάποιες περιπτώσεις, παρέχονται ελλιπείς, λανθασμένες, ή σε κάποιες περιπτώσεις είναι ακόμα και ανύπαρκτες. Παραδείγματα τέτοιων περιπτώσεων βρίσκονται σε μουσική η οποία είναι κατεβασμένη από ιστοσελίδες μεταφόρτωσης μουσικής ή από ψηφιακά μουσικά κανάλια. Υπάρχουν μάλιστα περιπτώσεις ροκ ή ποπ συγκροτημάτων, όπου ο βασικός τραγουδιστής ερμηνεύει μεν την πλειονότητα των κομματιών του συγκροτήματος, αλλά υπάρχει δε και μία μειονότητα κομματιών, τα οποία ερμηνεύονται από άλλα μέλη του συγκροτήματος. Αφού η πλειοψηφία των μουσικών αρχείων,

που περιέχουν καταγεγραμμένα δεδομένα, περιορίζονται σε στοιχεία του τύπου, τίτλος, καλλιτέχνης (όνομα συγκροτήματος), ή βασικός τραγουδιστής, αποκλείουν την καταγραφή μεμονωμένων περιπτώσεων, όπου τμήματα ενός κομματιού μπορεί να έχουν ερμηνευτεί από άλλον —εκτός του βασικού— τραγουδιστή.

Κάθε σύστημα αναγνώρισης τραγουδιστή ακολουθεί μία διαδικασία που χωρίζεται σε τρία στάδια, αυτό της εισαγωγής, αυτό της διερεύνησης και αυτό της ανάκτησης. Κατά το πρώτο στάδιο ηχητικά χαρακτηριστικά του τραγουδιστή από αρχεία ήχου εξάγονται και αποθηκεύονται σε μία βάση δεδομένων χαρακτηριστικών. Στο επόμενο στάδιο πραγματοποιείται μία διερεύνηση για το εάν τα δεδομένα του τραγουδιστή είναι γνωστά ή διευρευνάται το σύνολο χαρακτηριστικών του τραγουδιστή, για την αναγνώριση κατά το στάδιο ανάκτησης. Το στάδιο ανάκτησης, τέλος, χρησιμοποιεί μεθόδους σύγκρισης ή ένα σύνολο τέτοιων μεθόδων και ενημερώνει για το εάν υπάρχει κάποια αντιστοίχιση με τη βάση δεδομένων. Η διαδικασία αυτή παρουσιάζεται στο σχήμα 1.

Σχήμα 1: Γενική δομή διαδικασίας αναγνώρισης τραγουδιστή

Τα μοντέλα που λαμβάνονται με τη χρήση χαρακτηριστικών που εξάγονται από μίξεις φωνής και μουσικών οργάνων, αναπαριστούν την ταυτότητα του καλλιτέχνη ή του συγκροτήματος και όχι του τραγουδιστή. Οι Kim και Whitman (Kim & Whitman, 2002) και οι Berenzweig, Ellis και Laurence (Berenzweig, et al., 2002) αναφέρουν ότι για τη λήψη αντιπροσωπευτικών μοντέλων για την ταυτοποίηση του τραγουδιστή, η κατηγοριοποίηση πρέπει να πραγματοποιηθεί με τη χρήση τμημάτων κομματιών, στα οποία μόνο η φωνή είναι παρούσα ή να απορριφθούν από την ανάλυση τα τμήματα στα οποία εμπεριέχονται μόνο μουσικά όργανα. Ωστόσο, τα χαρακτηριστικά σε αυτά τα δύο συστήματα εξάγονται από μίξεις φωνητικών-μουσικής και δεν είναι δυνατό να μελετηθεί το κατά πόσο τα αποτελέσματα αλλοιώθηκαν από την παρουσία της μουσικής. Έτσι, οι Maddage, Xu και Wang (Maddage, et al., 2004) και οι Tsai και Wang (Tsai & Wang, 2006) πρότειναν συστήματα τα οποία εξάγουν ένα μοντέλο σόλο τραγουδιστή από ένα μοντέλο που εξάχθηκε από αυστηρώς ορχηστρικά

μέρη κομματιών και από ένα μοντέλο που εξάχθηκε από αποσπάσματα που εμπεριείχαν μίξη φωνητικών και μουσικής. Η πλειοψηφία αυτών των ερευνών δεν ανέδειξε κάποια βελτίωση στην απόδοσή των συστημάτων από τη διαφορετική διαχείριση αποσπασμάτων με φωνητικά και αποσπασμάτων μόνο με μουσική. Σε κάποιες περιπτώσεις, η απόδοση των συστημάτων που βασίζονται σε φασματικά χαρακτηριστικά από μίξεις φωνητικών-μουσικής επηρεάζονται πολύ από το λεγόμενο "album effect" ή "produced effect" (Y.E. Kim, 2006). Αυτό το φαινόμενο εμφανίζεται επειδή όλα τα κομμάτια από το ίδιο άλμπουμ ή παραγωγό μοιράζονται τα ίδια γενικά φασματικά χαρακτηριστικά. Μία άλλη προσέγγιση είναι η εκτέλεση της κατηγοριοποίησης σε απομονωμένα φωνητικά. Παραδείγματα αυτής της προσέγγισης αποτελούν τα συστήματα που παρουσιάζονται στις δημοσιεύσεις (Mesaros, et al., 2007) και (Fujihara, et al., 2005). Συγκεκριμένα, στη (Fujihara, et al., 2005) η φωνή απομονώνεται μέσω της ελάττωσης της έντασης της μουσικής συνοδείας. Στη (Mesaros, et al., 2007), η φωνή επανασυντίθεται με τη χρήση αρμονικά σχετιζόμενων συνιστωσών με τη θεμελιώδη συχνότητα της μελωδίας της φωνής. Πραγματοποιήθηκαν συγκρίσεις αποτελεσμάτων (Mesaros, et al., 2007), (Bartsch & Wakefield, 2004) τα οποία εξάχθηκαν από ακαπέλα ηχογραφήσεις με αποτελέσματα από φωνή απομονωμένη από μίξεις φωνητικών μουσικής. Οι μίξεις αυτές δημιουργήθηκαν μέσω της χρήσης των ίδιων ακαπέλα ηχογραφήσεων μιξαρισμένων με άλλα ορχηστρικά κομμάτια. Από αυτές τις συγκρίσεις βγήκε το συμπέρασμα ότι η απόδοση των συστημάτων που χρησιμοποίησαν απομονωμένα φωνητικά είναι πολύ χειρότερη από την απόδοση των συστημάτων που χρησιμοποίησαν ακαπέλα ηχογραφήσεις. Αυτή η απώλεια στην απόδοση των δεύτερων εικάζεται ότι οφείλεται σε ανεπιθύμητους ήχους που δημιουργήθηκαν κατά τη διαδικασία απομόνωσης της φωνής.

Το σύστημα των Regnier και Peeters (Regnier & Peeters, 2011) εξετάζει την ιδανική περίπτωση που η βάση δεδομένων αποτελείται από ακαπέλα ηχογραφήσεις και η αναγνώριση πραγματοποιείται μέσω του συνδυασμού τοπικών και παγκόσμιων περιγραφέων για τη φωνή. Χρησιμοποιήθηκαν δύο βάσεις δεδομένων, η LYR και η POP. Από τα αποτελέσματα των πειραμάτων τους βγαίνει το συμπέρασμα ότι η LYR βγάζει καλύτερα αποτελέσματα από την POP, μάλλον γιατί η POP επηρεάζεται από το "album-effect". Επίσης, η απόδοση της κατηγοριοποίησης στην POP επαληθεύτηκε μέσω της εκμάθησης των μοντέλων του τραγουδιστή στα 2/3 κάθε κομματιού και η επαλήθευση πραγματοποιήθηκε στα

εναπομείνοντα δεδομένα. Με χρήση της μεθόδου διασταυρωμένης επικύρωσης σε τρία μέρη, η μέση ακρίβεια που επετεύχθη ανέρχεται στο 96%.

Έχει παρουσιαστεί ένας εξαιρετικά μεγάλος όγκος δημοσιεύσεων που εξετάζει το ζήτημα της αυτόματης αναγνώρισης τραγουδιστή (SID), καθώς η φωνή αποτελεί το στοιχείο των κομματιών στο οποίο οι ακροατές εστιάζουν περισσότερο την προσοχή τους. Ένα από τα πιο πρώιμα τέτοια συστήματα είναι το Minnowmatch (Whitman, et al., 2001). Το συγκεκριμένο σύστημα προσεγγίζει το ζήτημα από τη σκοπιά της αναγνώρισης καλλιτέχνη και όχι από εκείνη της αναγνώρισης τραγουδιστή και δεν πραγματοποιεί διαχωρισμό της τραγουδιστής φωνής από τη μουσική. Τα συστήματα που περιγράφονται στις δημοσιεύσεις (Kim & Whitman, 2002) και (Berenzweig, et al., 2002) από την άλλη, προσεγγίζουν το ζήτημα από τη σκοπιά της αναγνώρισης τραγουδιστή. Τα συστήματα αυτά, σε πρώτη φάση ταυτοποιούν τα τμήματα των κομματιών τα οποία εμπεριέχουν φωνητικά. Συγκεκριμένα, στην περίπτωση του συστήματος των Berenzweig, Ellis και Laurence (Berenzweig, et al., 2002), σε κάθε τραγούδι, ο εντοπισμός των αποσπασμάτων που εμπεριέχουν φωνητικά πραγματοποιείται μέσω της εκπαίδευσης ενός MLP δικτύου από LPCC. Το σύστημα αυτό πέτυχε υψηλότερο ποσοστό ακρίβειας κατηγοριοποίησης σε σχέση με το σύστημα που περιγράφεται στη δημοσίευση (Whitman, et al., 2001), 65% στην ίδια βάση δεδομένων των 21 καλλιτεχνών και 269 κομματιών που αναφέρθηκε παραπάνω. Ο λόγος αυτής της βελτίωσης είναι ότι διαχωρίστηκε η μουσική από τα φωνητικά και χρησιμοποιήθηκαν μόνο τα φωνητικά για την εκπαίδευση του μοντέλου του τραγουδιστή. Αντιθέτως, οι Kim και Whitman (Kim & Whitman, 2002) χρησιμοποίησαν ανάστροφα συνδυασμένες μπάντες φίλτρων για να αναλύσουν την αρμονικότητα. Οι περιοχές των φωνητικών εντοπίστηκαν θέτοντας ένα κατώφλι (threshold) σε μία σταθερή τιμή, στην αρμονικότητα. Αυτό το σύστημα είχε ποσοστό ακρίβειας κατηγοριοποίησης 45,3 % σε ένα υποσύνολο 17 καλλιτεχνών και 200 τραγουδιών. Τα ταυτοποιημένα αποσπάσματα χρησιμοποιούνται προκειμένου να υπολογίσουν τα χαρακτηριστικά της κατηγοριοποίησης.

Το σύστημα των Liu και Huang (Liu & Huang, 2002) χρησιμοποιεί τη μέθοδο της αποσύνθεσης υποζώνης (subband decomposition) προκειμένου να αναγνωρίζει τραγουδιστές σε αρχεία κωδικοποιημένα σε μορφή MP3. Για την απομόνωση των MP3 φωνημάτων χρησιμοποιήθηκε μία μέθοδος εντοπισμού βασισμένη στις περιβάλλουσες. Το συγκεκριμένο σύστημα χρησιμοποιεί διαφορετική βάση δεδομένων επαλήθευσης από τα παραπάνω

συστήματα και η ακρίβεια κατηγοριοποίησης ανέρχεται περίπου στο 65% σε ένα σύνολο 10 αντρών και 10 γυναικών ερμηνευτών και ερμηνευτριών.

Άλλη μία προσέγγιση στο ζήτημα της αναγνώρισης τραγουδιστή είναι αυτή του Zhang (Zhang, 2003). Η μέθοδος που πρότεινε περιλαμβάνει τον εντοπισμό των σημείων έναρξης των τμημάτων που εμπεριέχουν φωνητικά, χρησιμοποιώντας απλές ρυθμίσεις κατωφλιού (threshold settings). Αυτές οι ρυθμίσεις υπολογίστηκαν από εξαγμένα χαρακτηριστικά, όπως η ενέργεια, ο ρυθμός μηδενικής διέλευσης (ZCR), η φασματική ροή (spectral flux) και οι αρμονικοί συντελεστές (harmonic coefficients). Το ποσοστό ακρίβειας των συστημάτων που παρουσιάζονται στις δημοσιεύσεις (Kim & Whitman, 2002), (Berenzweig, et al., 2002) και (Zhang, 2003) έφτασε έως 80% σε επίπεδο καρέ. Παρόλα αυτά, οι αποδόσεις τους είναι ανεπαρκείς για τους εξής λόγους:

- τα πειράματα εκτελέστηκαν σε καθαρά φωνητικά, ηχογραφημένα σε στούντιο και όχι σε μίξη φωνητικών-μουσικής.
- ο εντοπισμός των σημείων που ξεκινούν και τελειώνουν τα φωνητικά υπήρξε ανακριβής.
- δεν εκμεταλλεύθηκε η μουσική γνώση για τη μοντελοποίηση των τραγουδιστών στις συγκεκριμένες μεθόδους που είναι κυρίως τύπου bottom-up.

Οι Maddage, Xu και Wang (Maddage, et al., 2004), υποστήριξαν ότι ένας συνδυασμός bottom-up και top-down προσεγγίσεων θα μπορούσε να προσδώσει καλύτερες προοπτικές σε αυτού του τύπου τα συστήματα, καθώς έτσι θα μπορούσε να συνδυαστεί η δύναμη των χαρακτηριστικών χαμηλού επιπέδου με τη μουσική γνώση υψηλού επιπέδου.

Τα συστήματα (Whitman, et al., 2001), (Kim & Whitman, 2002), (Berenzweig, et al., 2002), (Liu & Huang, 2002) και (Zhang, 2003) αναζητούν λύσεις σε διαφορετικού είδους πρόβλημα από το σύστημα των Bartsch και Wakefield (Bartsch & Wakefield, 2004). Τα πρώτα δοκιμάστηκαν σε μουσικές μίξεις πραγματικών εκτελέσεων δημοφιλούς μουσικής, ενώ το δεύτερο εξετάζει την ιδανική συνθήκη στην οποία τα ηχητικά δείγματα είναι ηχογραφήσεις εκπαιδευμένων ερμηνευτών κλασικής μουσικής, οι οποίοι εκτέλεσαν φωνητικές ασκήσεις χωρίς συνοδεία μουσικής. Επειδή λοιπόν το σύστημα των Bartsch και Wakefield (Bartsch & Wakefield, 2004) είναι τόσο διαφορετικής φύσης από τα πρώτα πέντε, δεν μπορούν τα αποτελέσματά του να συγκριθούν άμεσα με εκείνων.

Άλλες προσεγγίσεις εντοπισμού φωνητικών παρουσιάζονται στις δημοσιεύσεις (Nwe & Wang, 2004) και (Maddage, et al., 2004), στις οποίες έγινε χρήση και μουσικών εννοιών εκτός από ακουστικών χαρακτηριστικών. Στο σύστημα που περιγράφεται στη δημοσίευση (Nwe & Wang, 2004), πληροφορίες για τη δομή του τραγουδιού και φωνητικά χαρακτηριστικά για κάθε τραγούδι ενσωματώνονται στη μοντελοποίηση του συστήματος διαχωρισμού τμημάτων με φωνητικά από τμήματα χωρίς φωνητικά. Στο (Maddage, et al., 2004), η κλίμακα οκταβικής συχνότητας χρησιμοποιείται για το χαρακτηρισμό της αρμονικής δομής των τμημάτων με φωνητικά και των τμημάτων χωρίς φωνητικά.

Οι Fujihara, Kitahara, Goto, Komatani, Ogata και Okuno (Fujihara, et al., 2005) παρουσίασαν ένα σύστημα το οποίο βασίστηκε σε τεχνικές απομάκρυνσης συνοδείας μουσικής από μουσικά σήματα και πέτυχαν ακρίβεια κατηγοριοποίησης 70-95 % σε μία βάση δεδομένων 10 καλλιτεχνών και 40 τραγουδιών.

Η μέθοδος SID που προτάθηκε από τους Bartsch και Wakefield (Bartsch & Wakefield, 2004) βασίζεται στη σύνθετη συνάρτηση μεταφοράς (CTF) για εκτίμηση φασματική περιβάλλουσας. Η CTF εξάγεται από το στιγμιαίο πλάτος και τη στιγμιαία συχνότητα των αρμονικών μερικών του σήματος. Αποδίδει καλά στο χαρακτηρισμό πολύπλοκων φωνητικών μεταβολών. Ένα παράδειγμα μίας τέτοιας μεταβολής είναι το βιμπράτο. Από το σύστημα αυτό επετεύχθη ακρίβεια κατηγοριοποίησης 82 % σε δεδομένα 12 ερμηνευτριών που ηχογραφήθηκαν από τους ίδιους τους ερευνητές. Μία πιο πρώιμη έκδοση της CTF παρουσιάστηκε στη δημοσίευση (Mellody, et al., 2001). Στην (Mellody & Wakefield, 2000) χρησιμοποιήθηκε ένα μοντέλο χαμηλής τάξης (low-order model) βασισμένο στα υπόλοιπα των CTF για την αναγνώριση της φωνητικής ταυτότητας του τραγουδιστή. Και πράγματι, μία ανάλυση κατά συστάδες (clustering analysis) αυτής της αναπαράστασης χαμηλής τάξης αποδεικνύει ότι όντως επιτυγχάνει να καταγράφει κάποια χαρακτηριστικά της φωνητικής ταυτότητας του τραγουδιστή.

Οι Shen, Shepherd, Cui και Tan (Shen, et al., 2006) πρότειναν ένα σύστημα για αποδοτική και αποτελεσματική αυτόματη αναγνώριση τραγουδιστή σε μεγάλες μουσικές βάσεις δεδομένων. Το σύστημα HSI χρησιμοποιεί μία πολυεπίπεδη δομή η οποία αποτελείται από τρία μείζονα δομικά επίπεδα, τη μονάδα προεπεξεργασίας, τη μονάδα μοντελοποίησης τραγουδιστή και τη μονάδα κατηγοριοποίησης συστήματος. Το διαχωρισμό της μουσικής σε τμήματα με φωνητικά και τμήματα χωρίς φωνητικά, τον αναλαμβάνει μία μηχανή υποστήριξης

διανυσμάτων (SVM). Η διαφορά του συγκεκριμένου συστήματος σε σχέση με προηγούμενα είναι ότι τα προηγούμενα χρησιμοποιούσαν μόνο ακουστικά χαρακτηριστικά χαμηλού επιπέδου από εξαγμένα από ανεπεξέργαστο σήμα. Το συγκεκριμένο σύστημα είναι ικανό να περισυλλέξει πληροφορίες για τον τραγουδιστή προκειμένου να βελτιωθεί η αποτελεσματικότητα της αναγνώρισης μέσω μίας υβριδικής αρχιτεκτονικής.

Το σύστημα HSI συγκεκριμένα, βελτιώνει την ακρίβεια ανάκτησης κατά 17,5% για γυναίκες τραγουδίστριες, 22,5% για άντρες και 20,0% κατά μέσο όρο σε σχέση με το σύστημα του Tsai. Επίσης, το σύστημα HSI έχει πολύ καλύτερη επεκτασιμότητα σε σχέση με τα συστήματα των Tsai (Tsai & Wang, 2006) και Liu (Liu & Huang, 2002), αφού δεν παρουσιάζει ιδιαίτερη πτώση στην ακρίβεια όταν αναφέρεται σε μεγαλύτερο όγκο δεδομένων. Επίσης, το HSI είναι αποδοτικό αφού ο χρόνος απόκρισής του σε διαφορετικούς όγκους δεδομένων είναι συγκρίσιμος με αυτούς των συστημάτων των Tsai και Liu. Τέλος, το σύστημα HSI παρουσιάζει ευρωστία, αφού η ακρίβειά του δεν επηρεάζεται ιδιαίτερα από θορύβους ή ηχητικές παρεμβολές. Μάλιστα, δοκιμάστηκε από τους Shen, Shepherd, Cui και Tan (Shen, et al., 2006) σε διαφορετικές συνθήκες παραμόρφωσης. Τέτοιες είναι ο περιβάλλων θόρυβος, η ηχώ με συγκεκριμένο χρόνο καθυστέρησης, η περικοπή, η μείωση και η ενίσχυση της έντασης.

Η μέθοδος ανταγνώρισης τραγουδιστή του Nikam (Nikam, 5 Δεκ. 2013) βασίζεται κυρίως στα φωνητικά περιεχόμενα του ηχητικού σήματος. Οι Holzapfel και Στυλιανού (Holzapfel & Στυλιανού, n.d.) εξέτασαν το πρόβλημα της αυτόματης αναγνώρισης τραγουδιστή χρησιμοποιώντας μεθόδους αναγνώρισης ομιλητή. Παρουσίασαν τρόπους χρήσης παγκόσμιων μοντέλων και αξιολογήθηκε η χρήση της αφαίρεσης φασματικού μέσου (CMS). Προκειμένου να ελαττωθούν οι διαφορές που οφείλονται στο μουσικό είδος χρησιμοποίησαν μία βάση δεδομένων που αποτελείται από δείγματα ρεμπέτικης μουσικής τα οποία παρουσίαζαν πολλές ομοιότητες. Μέσω της χρήσης αυτής της βάσης δεδομένων εξετάζεται για πρώτη φορά η επίδραση της ποιότητας ηχογράφησης, λόγω του ότι εμπεριέχονται πολλές ιστορικές ηχογραφήσεις, οι οποίες πραγματοποιήθηκαν σε γραμμόφωνο. Πειραματικές επαληθεύσεις ανέδειξαν τα προτερήματα των παγκόσμιων μοντέλων για επιλογή καρέ και CMS, δίνοντας μέση ακρίβεια κατηγοριοποίησης 81% σε βάση δεδομένων 21 καλλιτεχνών.

Στη δημοσίευση (Mellody, et al., 2001) παρουσιάζεται ένα παράδειγμα έρευνας πάνω στον τρόπο αντίληψης της ταυτότητας του τραγουδιστή, για την περίπτωση που εξετάζονται

εκπαιδευμένοι ερμηνευτές κλασικής μουσικής. Συγκεκριμένα, εξετάζεται η αναγνώριση τραγουδιστή σε δείγματα φωνητικών που δημιουργήθηκαν από τους ερευνητές με τεχνητό τρόπο, βάσει μεταβολών στα χαρακτηριστικά που αφορούν στο βιμπράτο. Ένα σύνολο από πειράματα αντιληπτικής αναγνώρισης τραγουδιστή παρουσιάζονται στη δημοσίευση (Mellody, 2001). Στη συγκεκριμένη δημοσίευση παρουσιάζεται ένα πείραμα στο οποίο ένα σύνολο τεσσάρων ακροατών, έπειτα από 10 με 12 ώρες εκπαίδευσης, κατάφεραν να κατηγοριοποιήσουν δώδεκα τραγουδιστές με μέση ακρίβεια κατηγοριοποίησης 82%. Η βάση δεδομένων που χρησιμοποιήθηκε για τις ανάγκες του συγκεκριμένου πειράματος χρησιμοποιήθηκε και στην έρευνα των Bartsch και Wakefield (Bartsch & Wakefield, 2004). Για αυτό το λόγο τα αποτελέσματα της συγκεκριμένης έρευνας είναι συγκρίσιμα με αυτά της έρευνας των Bartsch και Wakefield (Bartsch & Wakefield, 2004).

Το σύστημα που παρουσιάζεται στη δημοσίευση (Zhang, Ιούλιος 2003) είναι ένα σύστημα αναγνώρισης τραγουδιστή που προσεγγίζεται ως αναγνώριση ομιλητή μη εξαρτώμενη από κείμενο και επιτυγχάνει ακρίβεια κατηγοριοποίησης 82 % σε βάση δεδομένων 8 καλλιτεχνών και 45 τραγουδιών. Ωστόσο, η ακρίβεια κατηγοριοποίησης μειώθηκε σε 75 % όταν πραγματοποιήθηκαν τα πειράματα σε βάση δεδομένων 16 καλλιτεχνών.

Η εργασία των Tsai και Lee (Tsai & Lee, 2012) εξετάζει την πιθανότητα χρήσης δεδομένων ομιλίας αντί για δεδομένα τραγουδιστής φωνής. Παρόλα αυτά, επειδή δεν ήταν εύκολη η πλήρης αντικατάσταση δεδομένων τραγουδιστής φωνής από δεδομένα ομιλίας, στη μοντελοποίηση της φωνής του τραγουδιστή. Αυτό που δυσχέρανε τη διαδικασία της αντικατάστασης ήταν οι σημαντικές διαφορές μεταξύ της ομιλίας και της τραγουδιστής φωνής πολλών ανθρώπων. Έτσι, προτάθηκε μία εναλλακτική λύση βασισμένη στη χρήση των ολιγάριθμων διαθέσιμων δεδομένων τραγουδιστής φωνής. Χρησιμοποιήθηκε η μέθοδος της προσαρμογής MAP σε μερικά δεδομένα τραγουδιστής φωνής για την τροποποίηση φωνητικών μοντέλων που εξάχθηκαν από ομιλία. Με αυτό τον τρόπο τα προσαρμοσμένα φωνητικά μοντέλα είναι ικανά να καλύψουν τα χαρακτηριστικά τραγουδιστής φωνής των ερμηνευτών. Τα πειράματα έδειξαν ότι η πλειοψηφία των αποσπασμάτων τραγουδιστής φωνής μπορεί να αναγνωριστεί σωστά με τη χρήση προσαρμοσμένων φωνητικών μοντέλων.

Τα συστήματα που περιγράφονται στις δημοσιεύσεις (Kim & Whitman, 2002), (Berenzweig, et al., 2002), (Liu & Huang, 2002), (Zhang, 2003), (Tsai & Wang, 2006), (Nwe

& Li, 2007) και (Bartsch & Wakefield, 2004) αποτελούν παραδείγματα συστημάτων που είτε αγνόησαν την επίδραση της μουσικής συνοδείας στο χαρακτηρισμό της τραγουδιστής φωνής, είτε αντιμετώπισαν το ζήτημα της αναγνώρισης τραγουδιστή ως εργασία αναγνώρισης ομιλητή. Έτσι, στην περίπτωση των συγκεκριμένων συστημάτων δεν επιχειρήθηκε η απομάκρυνση της μουσικής συνοδείας από τα φωνητικά χαρακτηριστικά. Τα συστήματα που προσεγγίζουν το ζήτημα της αναγνώρισης τραγουδιστή ως ζήτημα αναγνώρισης ομιλίας (Tsai & Wang, 2006), (Nwe & Li, 2007) μοντελοποιούν τα χαρακτηριστικά κάθε ατόμου στα δικά του φωνητικά δεδομένα. Το μειονέκτημα συστημάτων τέτοιου είδους είναι ότι δεν είναι πάντα το ίδιο εφικτό να συλλέγονται σόλο ακαπέλα αποσπάσματα για κάθε τραγουδιστή, όσο είναι να συλλέγονται δεδομένα ομιλίας από κάθε ομιλητή, για τη χρήση σε SPID εφαρμογές.

Πολύ λίγα συστήματα εφαρμόζουν μεθόδους διαχωρισμού της τραγουδιστής φωνής από τη μουσική συνοδεία. Ένα εξ αυτών είναι αυτό του Wang (Wang, 1994), στο οποίο εκτελείται διαχωρισμός της τραγουδιστής φωνής με τη χρήση μίας τεχνικής βρόγχου κλειδωμένης αρμονικής (harmonic-locked loop technique) προκειμένου να παρακολουθηθεί ένα σύνολο από αρμονικά σχετιζόμενες μερικές. Στο συγκεκριμένο σύστημα, η θεμελιώδης συχνότητα της τραγουδιστής φωνής πρέπει να είναι γνωστή από την αρχή. Μία ακόμα ιδιότητα του συγκεκριμένου συστήματος είναι ότι δεν είναι ικανό να ξεχωρίσει της τραγουδιστή φωνή από τους υπόλοιπους μουσικούς ήχους. Αυτή η αδυναμία του συστήματος το οδηγεί στο να παρακολουθεί λανθασμένα μερικές που ανήκουν σε άλλη αρμονική πηγή, όταν η τραγουδιστή φωνή απουσιάζει. Ο βρόγχος κλειδωμένης αρμονικής απαιτεί τον υπολογισμό της στιγμιαίας συχνότητας μίας μερικής. Ο υπολογισμός αυτός όμως δεν είναι αξιόπιστος στην περίπτωση της παρουσίας μερικών και από άλλες ηχητικές πηγές. Έτσι, το σύστημα λειτουργεί σωστά μόνο στην περίπτωση που ο λόγος της ενέργειας της τραγουδιστής φωνής ως προς τη συνοδεία μουσικής είναι αρκετά υψηλός. Το σύστημα των Meron και Hirose (Meron & Hirose, 1998) από την άλλη, έχει ως σκοπό το διαχωρισμό της τραγουδιστής φωνής από συνοδεία πιάνου. Για τη λειτουργία του συστήματος απαραίτητη προϋπόθεση αποτελεί η πρότερη ύπαρξη μία σημαντικής ποσότητας πληροφοριών. Παραδείγματα τέτοιων πληροφοριών αποτελούν τα κανάλια των μερικών της προκαταρκτικής μίξης της τραγουδιστής φωνής και του πιάνου, καθώς και η παρτιτούρα του πιάνου. Ωστόσο, αυτού του είδους η γνώση στην πλειοψηφία των περιπτώσεων δε διατίθεται. Το γεγονός αυτό οδηγεί στην αδυναμία εφαρμογής του στην πλειονότητα των κομματιών.

Οι Li και Wang (Li & Wang, Μάιος 2007) προσπάθησαν να βρουν μία λύση ηχητικού διαχωρισμού για μονοφωνικές ηχογραφήσεις, αφού οι προσεγγίσεις διαχωρισμού ομιλίας βασισμένες σε σειρές μικροφώνων δεν είναι εφαρμόσιμες. Για το διαχωρισμό των μονοφωνικών ηχογραφήσεων είναι δυνατό να χρησιμοποιηθούν μέθοδοι βελτίωσης ομιλίας. Οι μέθοδοι βελτίωσης ομιλίας τείνουν να αναζητούν την ύπαρξη παρεμβολών, όπως στατικότητας. Οι παρεμβολές δεν αποτελούν ευχάριστο χαρακτηριστικό μίας μουσικής συνοδείας. Μία προσέγγιση γενικού ηχητικού διαχωρισμού αξιοποιεί τις γνώσεις μας πάνω στο ανθρώπινο ακουστικό σύστημα. Ο Bregman (Bregman, 1990) στο βιβλίο του ανέφερε ότι το ακουστικό σύστημα εφαρμόζει μία διαδικασία η οποία ονομάζεται ανάλυση ακουστικής σκηνής (ASA). Σκοπός αυτής της ανάλυσης είναι η οργάνωση μίας ακουστικής μίξης σε ξεχωριστές αντιληπτικές ροές, η καθεμιά ροή εκ των οποίων αντιστοιχεί σε διαφορετικές ηχητικές πηγές. Η διαδικασία περιλαμβάνει δύο κύρια στάδια. Τα στάδια αυτά είναι αυτό του διαχωρισμού και αυτό της ομαδοποίησης. Στο πρώτο στάδιο, η ακουστική είσοδος χωρίζεται σε αποσπάσματα χρόνου-συχνότητας (T-F). Το κάθε απόσπασμα προέρχεται από μία συγκεκριμένη πηγή. Στο δεύτερο στάδιο, τα τμήματα της ίδιας πηγής ομαδοποιούνται σύμφωνα με τις θεμελιώδεις αρχές ASA. Παραδείγματα τέτοιων αρχών αποτελεί το onset, το offset και η αρμονικότητα. Η ASA οδήγησε στη δημιουργία των συστημάτων υπολογιστικής ακουστικής ανάλυσης σκηνής (CASA). Αυτά τα συστήματα χρησιμοποιήθηκαν για ηχητικό διαχωρισμό (Brown & Wang, n.d.), (Divenyi, 2005), (Rosenthal & Okuno, 1998). Τα συστήματα CASA πραγματοποιούν πολύ λιγότερες υποθέσεις για τους ήχους που περιβάλλοντες ήχους, σε σχέση με τα υπόλοιπα συστήματα ηχητικού διαχωρισμού. Αντιθέτως, εστιάζει στα εγγενή χαρακτηριστικά των ήχων και για αυτό το λόγο παρουσιάζει καλύτερες προοπτικές για το διαχωρισμό της φωνής σε μονοφωνικές ηχογραφήσεις.

Ο Mellinger (Mellinger, 1991) έκανε την πρώτη απόπειρα να χρησιμοποιήσει το σύστημα CASA για το διαχωρισμό μουσικών ήχων. Το σύστημά του εξάγει onset και κοινές συχνοτικές μεταβολές και χρησιμοποιεί τα εξαγμένα στοιχεία προκειμένου να ομαδοποιήσει τις συχνοτικές μερικές του ίδιου μουσικού οργάνου. Ωστόσο, αυτά τα στοιχεία δεν επαρκούν για το διαχωρισμό ήχων. Γι’ αυτό το λόγο, προτάθηκαν από το Mellinger άλλα στοιχεία όπως το τονικό ύψος για τον ηχητικό διαχωρισμό. Αυτό και η αρχή της αρμονικότητας χρησιμοποιούνται ευρέως σε συστήματα CASA. Παραδείγματα εφαρμογών που χρησιμοποιούν αυτά τα στοιχεία είναι αυτά των Godsmark και Brown (D. Godsmark & Brown,

1999) και του Goto (Goto, 2004). Στο πρώτο χρησιμοποιήθηκε η αρμονικότητα και άλλες αρχές, σε μία αρχιτεκτονική μαυροπίνακα (blackboard architecture) για τη διαδικασία ομαδοποίησης. Ενώ στο δεύτερο χρησιμοποιήθηκε η αρχή της αρμονικότητας από ένα σύστημα περιγραφής μουσικής σκηνής (music-scene-description) για εντοπισμό μελωδίας.

Επίσης, οι Hu και Liu (Hu & Liu, 2013) πρότειναν μία μέθοδο αναγνώρισης τραγουδιστή σε μονοφωνική δημοφιλή μουσική, η οποία αποτελείται από δύο στάδια. Στο πρώτο εξ αυτών εφαρμόζεται η CASA ως εξής. Η εκτιμώμενη δυαδική μάσκα T-F εντοπίζει για κάθε καρέ, τα τμήματα T-F τα οποία εμπεριέχουν κυρίως τραγουδιστή φωνή και θεωρούνται αξιόπιστα. Άλλα τμήματα όμως είναι αναξιόπιστα ή απόντα. Για αυτό το λόγο, το φάσμα παραμένει ελλιπές. Στο δεύτερο στάδιο, χρησιμοποιούνται δύο μέθοδοι ελλιπές χαρακτηριστικού για την αναγνώριση του τραγουδιστή. Αυτές οι μέθοδοι ονομάζονται ανακατασκευή (reconstruction) και περιθωριοποίηση (marginalization). Στην περίπτωση της ανακατασκευής, ανακατασκευάζεται ολόκληρο το φάσμα στην αρχή και έπειτα μετατρέπεται σε GFCC, ενώ στην περίπτωση της περιθωριοποίησης, υπολογίζονται οι πιθανότητες της φωνής του τραγουδιστή μόνο για τα αξιόπιστα στοιχεία. Η μέθοδος της ανακατασκευής επιφέρει καλύτερα αποτελέσματα από την περιθωριοποίηση. Παρόλα αυτά και οι δύο παρουσιάζουν πολύ καλά αποτελέσματα σε σχέση με άλλα συστήματα, ειδικά στην περίπτωση που ο SAR είναι 0 dB και -3 dB. Συγκεκριμένα, η καλύτερη δυνατή ακρίβεια που επιτυγχάνει η μέθοδος ανακατασκευής είναι 90,28% με SAR 0 dB και σε φάσμα 128 διαστάσεων, ενώ η καλύτερη της μεθόδου περιθωριοποίησης είναι 83,25% με SAR 0 dB και σε φάσμα 64 διαστάσεων. Και τα δύο ποσοστά αναφέρονται σε αναγνώριση τραγουδιστή υπό συνθήκες IBM. Οι μέθοδοι αυτοί συγκρίθηκαν με το σύστημα των Tsai και Lin (Tsai & Lin, 2011). Το σύστημα των Tsai και Lin έδωσε ποσοστό ακρίβειας κατά 39 % μικρότερο από τη μέθοδο ανακατασκευής με 0 dB SAR, ενώ κατά 33 % με -3 dB SAR. Ωστόσο, η μέθοδος των Tsai και Lin έδωσε ποσοστό ακρίβειας ελάχιστα μεγαλύτερο από τη μέθοδο ανακατασκευής με SAR -9 dB, ενώ κατά 7 % και 20 % μεγαλύτερο από τη μέθοδο περιθωριοποίησης με -6 dB και -9 dB SAR αντίστοιχα.

Το σύστημα των Tsai και Lin (Tsai & Lin, 2011) ερευνά τις σχέσεις μεταξύ της σόλο τραγουδιστής φωνής και της μουσικής συνοδείας σε ένα φάσμα. Τα φάσματα των τραγουδιστών φωνών πάρθηκαν από το μετασχηματισμό των φασμάτων της μίξης τραγουδιστής φωνής-μουσικών οργάνων. Τέλος, τα φάσματα της τραγουδιστής φωνής

χρησιμοποιήθηκαν για τη μοντελοποίηση κάθε τραγουδιστή και για την τροφοδότηση του ταξινομητή για την αναγνώριση τραγουδιστή.

Οι Hu και Wang (Hu & Wang, Σεπ. 2004) ανέπτυξαν ένα σύστημα ηχητικού διαχωρισμού το οποίο διαχωρίζει με επιτυχία ομιλία από ακουστικές παρεμβολές και βασίζεται στην παρακολούθηση του τονικού ύψους και στη διαμόρφωση του πλάτους. Επίσης, χρησιμοποιεί διαφορετικές μεθόδους διαχωρισμού για επιλυμμένες (resolved) και άλυτες (unresolved) αρμονικές. Η συστηματική επαλήθευσή του σε μία συχνά χρησιμοποιούμενη βάση δεδομένων απέδειξε την υπεροχή του έναντι στα παλαιότερα συστήματα.

Επίσης, το σύστημα στηρίζεται στο τονικό ύψος για την ομαδοποίηση των τμημάτων. Για το λόγο αυτό, είναι πολύ σημαντικό ο εντοπισμός του τονικού ύψους να πραγματοποιείται με ακρίβεια. Παρόλα αυτά, διατηρεί την αρχική εκτίμηση του τονικού του ύψους από το χρόνο υστέρησης που ανταποκρίνεται στο μέγιστο μίας σύνοψης της συνάρτησης αυτοσυσχέτισης. Αυτή η εκτίμηση του τονικού ύψους είναι αναξιόπιστη για τραγουδιστή φωνή όπως φαίνεται στο και περιορίζει την απόδοση του διαχωρισμού στο σύστημα (Li & Wang, 2005). Στο (Li & Wang, 2005) παρουσιάστηκε ένας επικρατών αλγόριθμος εντοπισμού του τονικού ύψους, ο οποίος έχει τη δυνατότητα να εντοπίσει το τονικό ύψος τραγουδιστής φωνής για διαφορετικά μουσικά είδη ακόμα και στην περίπτωση που η συνοδεία μουσικής είναι πολύ έντονη. Το σύστημα των Hu και Wang (Li & Wang, 2005) λαμβάνει ως συνθήκη ότι η ομιλία είναι πάντα παρούσα, κάτι που φυσικά δε συμβαίνει στην περίπτωση του διαχωρισμού τραγουδιστής φωνής από ένα σήμα. Έτσι, είναι αναγκαία η δημιουργία ενός μηχανισμού διάκρισης των τμημάτων με τραγουδιστή φωνή από εκείνα χωρίς. Παρόλο που το σύστημα δεν είναι ικανό να διαχωρίσει ομιλία χαμηλής έντασης από ένα σήμα, αυτό δεν αποτελεί μεγάλο πρόβλημα για το διαχωρισμό τραγουδιστής φωνής από ένα σήμα, αφού η τραγουδιστή φωνή χαμηλής έντασης καταλαμβάνει μικρό ποσοστό των τραγουδιών και η συνεισφορά της στην αναγνώριση τραγουδιστή είναι πολύ μικρότερη από εκείνη στην αναγνώριση ομιλίας.

Οι Deshmukh και Bhirud (Deshmukh & Bhirud, 2012) προτείνουν μία υβριδική μέθοδο επιλογής σωστών περιγραφέων ήχου για την εργασία της αναγνώρισης ερμηνευτών κλασικής μουσική της Βόρειας Ινδίας. Η μεθοδολογία που ακολουθείται είναι η εξής. Πρώτα απελευθερώνονται στο σύστημα μόνο οι πρωταρχικοί ηχητικοί περιγραφείς κατά το εμπρόσθιο πέρασμα και καταγράφεται η επίδραση της κατηγοριοποίησης. Έπειτα επιλέγονται μόνο οι ηχητικοί περιγραφείς με την καλύτερη απόδοση στην αναγνώριση τραγουδιστή και οι

υπόλοιποι απομακρύνονται κατά το οπίσθιο πέρασμα. Η διαδικασία της επιλογής και απομάκρυνσης όλων των λιγότερο σημαντικών ηχητικών περιγραφέων από τις ομάδες που παρουσιάζουν τη μεγαλύτερη επιρροή στην αναγνώριση καλλιτέχνη ενισχύει την απόδοση των συστημάτων αναγνώρισης τραγουδιστή. Επίσης, η μέθοδος αυτή ελαττώνει σημαντικά τον αριθμό των ηχητικών περιγραφέων, διατηρώντας μόνο τους σημαντικούς. Τέλος, οι επιλεγμένοι ηχητικοί περιγραφείς τροφοδοτούνται σε ταξινομητές.

Η μέθοδος που προτάθηκε από το Nikam (Nikam, 5 Δεκ. 2013) επαληθεύεται σε ηχοχρωματικά χαρακτηριστικά οποιουδήποτε ηχητικού δείγματος. Η συγκεκριμένη εργασία εστιάζει σε διάφορους ηχητικούς περιγραφείς της χροιάς, προκειμένου να εντοπίσει ποια χαρακτηριστικά του ήχου είναι τα πλέον χρήσιμα για τη διαδικασία της αναγνώρισης τραγουδιστή. Εξετάστηκαν βασικοί και φασματικοί ηχητικοί περιγραφείς όπως το φασματικό κεντροειδές (SC), το φασματικό Roll-Off, η φασματική κύρτωση μηδενικής διέλευσης, η RMS ενέργεια, η εντροπία Shannon, η φωτεινότητα (brightness), οι MFCC και η θεμελιώδης συχνότητα (f0). Από τα πειράματα βγήκε το συμπέρασμα ότι οι βασικοί ηχητικοί περιγραφείς και τα φασματικά χαρακτηριστικά μπορούν να χρησιμοποιηθούν συνδυαστικά για να δημιουργήσουν ένα αποδοτικό διάνυσμα χαρακτηριστικών, το οποίο να είναι ικανό να φέρει εις πέρας εργασίες ανάκτησης μουσικής πληροφορίας. Ένας τέτοιος συνδυασμός είναι ικανός να βελτιώσει την απόδοση μεθόδων βασισμένων σε MFCC χαρακτηριστικά και η εφαρμογή αυτής της συνδυαστικής μεθόδου είναι πολλά υποσχόμενη για χρήση σε συστήματα αναγνώρισης τραγουδιστή. Ο καλύτερος συνδυασμός ηχητικών περιγραφέων φαίνεται ότι περιλαμβάνει τους εξής: SC, φασματικό Roll-Off, φασματική κύρτωση μηδενικής διέλευσης, χαμηλή ενέργεια (low energy), πλήθος αρμονικών και ανωμαλία (irregularity).

Η αναγνώριση καλλιτέχνη στην περίπτωση που αφορά την αναγνώριση σόλο φωνητικών, έχει πάρα πολλά κοινά με την αναγνώριση ομιλητή και πολλοί ερευνητές έχουν ασχοληθεί με την εργασία της αναγνώρισης ομιλητή στο παρελθόν (Campbell Jr, 1997), (Mammone, et al., 1996), (Furui, 1997). Τα καλύτερα συστήματα αναγνώρισης ομιλητή έχουν αγγίξει ποσοστά ακρίβειας της τάξης του 95% για καθαρή ομιλία και 80% για τηλεφωνική. Ωστόσο υπάρχουν σημαντικές διαφορές ανάμεσα στην εργασία της αναγνώρισης τραγουδιστή και σε αυτή της αναγνώρισης ομιλίας. Η πρώτη βασική διαφορά είναι ότι η τραγουδιστή φωνή είναι πολύ πιο ασταθής από την ομιλία, σε φυσιολογικές συνθήκες, όσον αφορά τη θεμελιώδη συχνότητα και το πλάτος των φωνημάτων. Επιπλέον, η πλειοψηφία των μοτίβων της

τραγουδιστής φωνής ακολουθεί μία παρτιτούρα. Επομένως δεν είναι δυνατό να θεωρηθεί ότι τα μοτίβα αυτά αποτελούν χαρακτηριστικά μοτίβα ενός συγκεκριμένου ατόμου. Η ομιλία και η τραγουδιστή φωνή επίσης, αλλοιώνονται συνήθως με αρκετά διαφορετικούς τρόπους. Συνήθως οι έρευνες πάνω στην αναγνώριση ομιλητή εστιάζουν στις μεταβολές που υπόκειται ένα σήμα κατά τη μετάδοσή του μέσω διαύλων επικοινωνίας, όπως αυτές που συμβαίνουν κατά τη μετάδοση σήματος ομιλίας σε μία τηλεφωνική γραμμή. Στην περίπτωση της τραγουδιστής φωνής από την άλλη, οι αλλοιώσεις που συμβαίνουν έχουν συνήθως τη μορφή δομημένων παρεμβολών. Ένα παράδειγμα τέτοιων παρεμβολών είναι τα μουσικά όργανα που συμμετέχουν σε μία μουσική μίξη. Έτσι, οι προσεγγίσεις αναγνώρισης ομιλητής μπορούν να φανούν χρήσιμες στην περίπτωση της αναγνώρισης σόλο φωνητικών, αλλά δεν είναι απαραίτητο ότι θα έχουν παρόμοια ποσοστά επιτυχίας στην ευρύτερη εργασία της αναγνώρισης τραγουδιστή σε πολυφωνικές μίξεις.

Επιπλέον, η αναγνώριση τραγουδιστή συνδέεται και με πιο γενικές έρευνες ηχητικής ταξινόμησης και κατηγοριοποίησης. Μία εξ αυτών που σχετίζεται άμεσα με αυτή είναι η αυτόματη αναγνώριση μουσικών οργάνων. Η διαφορά μεταξύ της εργασίας αναγνώρισης τραγουδιστή και αυτής της αναγνώρισης μουσικών οργάνων είναι ότι σχεδόν όλες οι έρευνες πάνω στην εργασία αναγνώρισης μουσικών οργάνων εστιάζουν στην απομόνωση μουσικών τόνων. Άλλες εργασίες οι οποίες σχετίζονται με την αναγνώριση τραγουδιστή είναι ο διαχωρισμός ομιλίας-μουσικής (Martin, et al., 1998), (Scheirer & Slaney, 1997) και η κατηγοριοποίηση ήχου σε αυθαίρετες κατηγορίες (Wold, et al., 1996), (Li, 2000). Η έρευνα που παρουσιάζεται στη δημοσίευση (Martin & Kim, 1998) και αφορά αναγνώριση μουσικών οργάνων επιτυγχάνει 70% ακρίβεια σε μία βάση δεδομένων 14 διαφορετικών ορχηστρικών μουσικών οργάνων.

2.1.2 Πρακτικές εφαρμογές συστημάτων αναγνώριση τραγουδιστή

Η αναγνώριση καλλιτέχνη βρίσκει εφαρμογή σε πεδία που σχετίζονται με τα πολυμέσα, τις πολυμεσικές βάσεις δεδομένων, τις τηλεπικοινωνίες, καθώς και την ασφάλεια και υπάρχουν πολλά διαθέσιμα συστήματα που χρησιμοποιούν μεθόδους αναγνώρισης ομιλίας, καθώς και τραγουδιστής φωνής.

Αρκεί να αναλογιστούμε ότι η συνήθης μέθοδος που ακολουθούνταν πριν την επαρκή ανάπτυξη αυτοματοποιημένων συστημάτων ήταν η αναζήτηση μουσικής μέσω της χειροκίνητης καταγραφής του ονόματος του καλλιτέχνη ή του ονόματος του τραγουδιού για να καταλάβουμε πόσο έχουν βελτιώσει τη ζωή μας τέτοιου είδους συστήματα (Cano, et al.,

2005). Ένα από τα κύρια μειονεκτήματα της παραπάνω μεθόδου ήταν το γεγονός ότι ο χρήστης έπρεπε να γνωρίζει το όνομα του καλλιτέχνη. Παρόλα αυτά, υπάρχει και η περίπτωση ο χρήστης να έχει στη συλλογή του ένα κομμάτι που να μη γνωρίζει τον καλλιτέχνη που το ερμηνεύει και να επιθυμεί να εντοπίσει κομμάτια και άλμπουμ που παρουσιάζουν ηχητικές ομοιότητες ως προς αυτό, εξετάζοντάς τα είτε ως προς τη φωνή του τραγουδιστή, είτε συνολικά. Σε αυτή την περίπτωση ένα σύστημα αναγνώρισης του καλλιτέχνη που ερμηνεύει τα κομμάτια που τον ενδιαφέρουν αποτελεί τη μοναδική λύση στο πρόβλημά του. Χάρη σε ένα τέτοιο σύστημα προσφέρονται νέες λειτουργίες στα συστήματα διαχείρισης μουσικής, όπως η παροχή, στο χρήστη, πληροφοριών για τη φωνητική ταυτότητα ερμηνευτών (κυρίως ερασιτεχνών) μουσικών κομματιών, οι οποίες είναι εξαιρετικά δύσκολο να εντοπιστούν, καθώς και η ανάκτηση όλων των κομματιών, τα οποία ερμηνεύει ένας συγκεκριμένος καλλιτέχνης, σε μία παρεχόμενη μουσική βιβλιοθήκη.

Άλλες εφαρμογές της αναγνώρισης ερμηνευτή είναι ο εντοπισμός των αποσπασμάτων που ερμηνεύει ο υπό ερεύνηση τραγουδιστής, σε ηχογραφημένα κομμάτια σε συνθήκες στούντιο, αλλά και σε ζωντανές ηχογραφήσεις, καθώς και η διάκριση μεταξύ μίας πρωτότυπης ηχογράφησης και μίας διασκευής ερμηνευμένης από άλλους τραγουδιστές. Η αναγνώριση ερμηνευτή μπορεί επίσης, να δώσει τη δυνατότητα στις δισκογραφικές εταιρίες να σαρώνουν ταχέως ύποπτες ιστοσελίδες για «πειρατικό» περιεχόμενο. Αυτό θα χρησίμευε ιδιαιτέρως σε περιπτώσεις παράνομων ηχογραφήσεων συναυλιών, των οποίων ούτε η ίδια η εταιρία δεν κατέχει αντίγραφο του αρχείου ήχου προκειμένου να πραγματοποιηθεί σύγκριση. Μία άλλη πιθανή χρήση ενός τέτοιου συστήματος, θα μπορούσε να είναι η οργάνωση των μουσικών προτιμήσεων των συμμετεχόντων σε βραδιές καραόκε από τους οργανωτές των βραδιών αυτών, με τρόπο τέτοιο ώστε η υπηρεσία που προσφέρεται να είναι πιο προσωποποιημένη. Πολλές μέθοδοι αναγνώρισης ερμηνευτή έχουν μάλιστα εφαρμοστεί σε συστήματα που προτείνουν μουσική. Σε αυτά τα συστήματα προτείνονται τραγούδια από ερμηνευτές με παρόμοια φωνή με αυτήν του καλλιτέχνη που ερμηνεύει το τραγούδι που ενδιαφέρει το χρήστη. Στις επόμενες ενότητες παρουσιάζονται εκτενώς χαρακτηριστικά και αλγόριθμοι κατηγοριοποίησης που έχουν χρησιμοποιηθεί σε εργασίες μηχανικής μάθησης και χρησιμοποιήθηκαν στην υλοποίησή μας.

2.2 Χαρακτηριστικά

Στα πειράματα που έχουν εκτελεστεί και αφορούν στην εργασία της αυτόματης αναγνώρισης καλλιτέχνη καθώς και σε άλλα παρεμφερή ζητήματα, χρησιμοποιήθηκαν κάποια

συγκεκριμένα χαρακτηριστικά για την κατηγοριοποίηση των ηχητικών δειγμάτων. Η πλειοψηφία των συστημάτων που περιγράφουν και αναγνωρίζουν το περιεχόμενο ηχητικών σημάτων χρησιμοποιούν συνήθως έναν ή περισσότερους φασματικούς περιγραφείς που σχετίζονται με τη χροιά. Τέτοιοι είναι τα MFCC, τα GFCC, τα LPC χαρακτηριστικά, καθώς και τα παράγωγά τους. Στα συστήματα που περιγράφονται στις δημοσιεύσεις (Whitman, et al., 2001), (Berenzweig, et al., 2002), (Zhang, Ιούλιος 2003), (Lagrange, et al., 2012), (Mesaros, et al., 2007) και (Holzapfel & Στυλιανού, n.d.) για παράδειγμα, χρησιμοποιήθηκαν MFCC ως χαρακτηριστικά. Οι MFCC αποτελούν μία κλάση χαρακτηριστικών που συνήθως χρησιμοποιείται στις εργασίες αναγνώρισης και ταυτοποίησης ομιλητή (Davis & Mermelstein, 1980). Στο σύστημα που παρουσιάζεται στη δημοσίευση (Whitman, et al., 2001), τα MFCC υπολογίστηκαν από μικρά ηχητικά αποσπάσματα υποσυνόλων βάσεων δεδομένων που περιλαμβάνουν από 5 έως και 21 δημοφιλείς καλλιτέχνες.

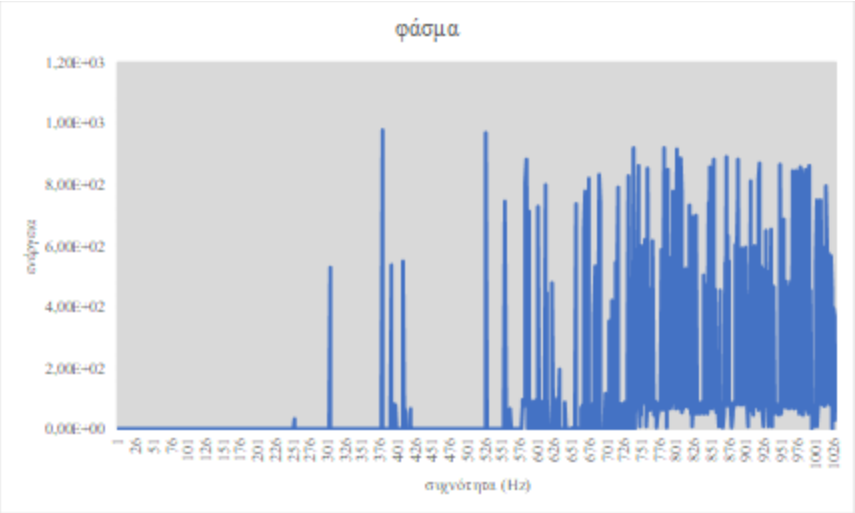
2.2.1 Cepstral συντελεστές συχνότητας Mel (Mel Frequency Cepstral Coefficients)

Το cepstrum συχνοτήτων mel έχει αποδειχθεί ότι αποτελεί έναν εξαιρετικά αποδοτικό τύπο cepstrum στην εργασία της αυτόματης αναγνώρισης ομιλίας και στη μοντελοποίηση του υποκειμενικού χαρακτηριστικού του ύψους και του συχνοτικού περιεχομένου των ηχητικών σημάτων. Αυτό ορίζεται ως το cepstrum που υπολογίζεται σε κλίμακα Mel και όχι σε φάσμα Fourier. Τα χαρακτηριστικά του mel cepstrum είναι δυνατό να αποτυπωθούν από τους MFCC. Τα MFCC επίσης διαθέτουν την ικανότητα αναπαράστασης του σχήματος του φάσματος με χρήση πολύ λίγων συντελεστών. Οι συντελεστές αυτοί είναι cepstral και υπολογίζονται σε μία συχνοτική κλίμακα mel. Ανάλυση cepstral ονομάζεται η μέθοδος του διαχωρισμού της περιβάλλουσας του φάσματος από τη λεπτομερή δομή του και ως cepstrum ορίζεται ο μετασχηματισμός Fourier ή ο διακριτός μετασχηματισμός συνημιτόνου (DCT) του λογάριθμου του φάσματος. Η χρήση της κλίμακας Mel επιτρέπει να δοθεί μεγαλύτερη έμφαση στις μεσαίες συχνότητες του σήματος. Οι συντελεστές cepstral υπολογίζονται μέσω του λογαριθμικού πλάτους διακριτού μετασχηματισμού συνημιτόνου του φάσματος ισχύος. Οι περιβάλλουσες αναπαριστώνται σε χαμηλότερη τάξη από τους cepstral συντελεστές, ενώ οι λεπτομερείς δομές βρίσκονται σε υψηλότερη τάξη. Για τους υπολογισμούς των MFCC εφαρμόζεται πρώτα ανάλυση mel τράπεζας φίλτρων και τέλος, λαμβάνονται οι MFCC από το λογαριθμικό πλάτος του DCT. Οι MFCC επομένως ορίζονται ως οι συντελεστές του Mel cepstrum. Στο σχήμα 2 παρουσιάζεται το φάσμα ενός σήματος, στο σχήμα 3 οι MFC συντελεστές και στο σχήμα 4 οι ζώνες MFCC. Ο πρώτος συντελεστής δεν αποθηκεύεται

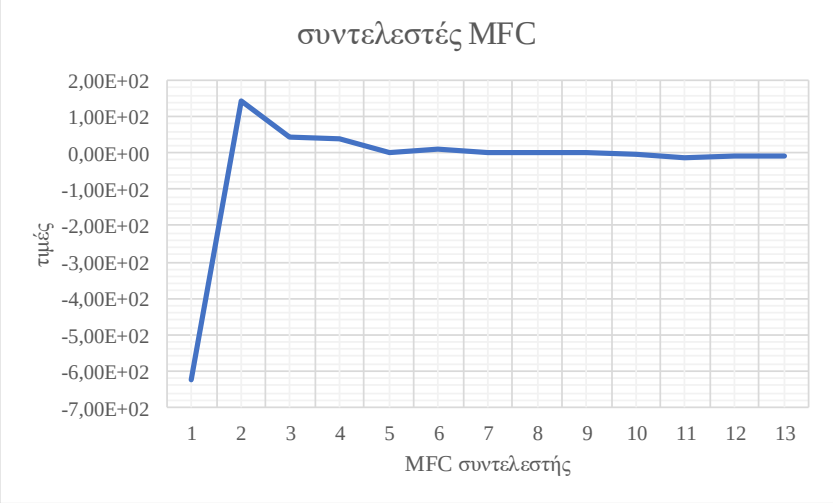
επειδή είναι ανάλογος της ενέργειας, ενώ οι επόμενοι 12 αποθηκεύονται για κάθε καρέ. Τα ΔMFCC και τα ΔΔMFCC αποτελούν την παράγωγο πρώτης και δεύτερης τάξης των MFCC στο χρόνο.

$$\Delta\text{MFCC} = \frac{\partial}{\partial t}\text{MFCC(t)} \tag{2.1}$$

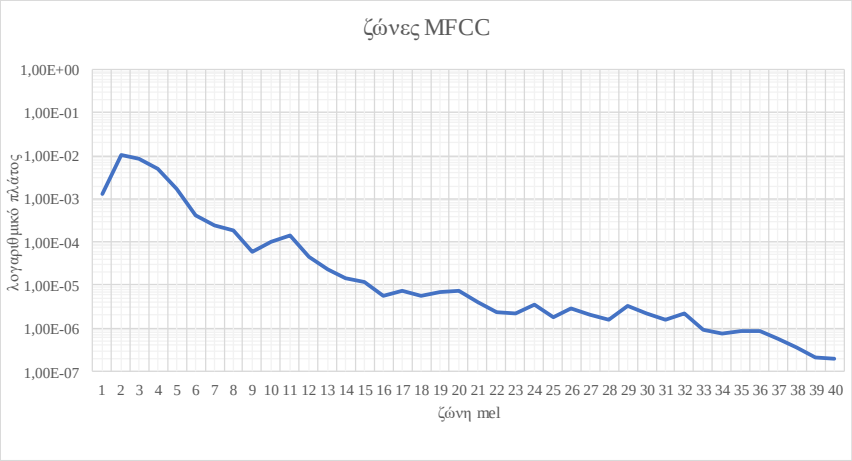
$$\Delta\Delta\text{MFCC} = \frac{\partial^2}{\partial^2 t}\text{MFCC(t)} \tag{2.2}$$



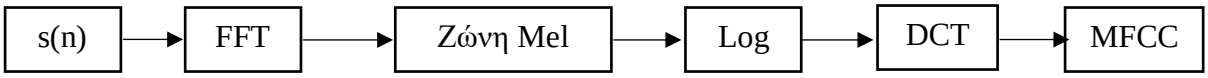
Σχήμα 2: Φάσμα αρχείου ήχου



Σχήμα 3: Συντελεστές MFC



Σχήμα 4: Ζώνες MFCC

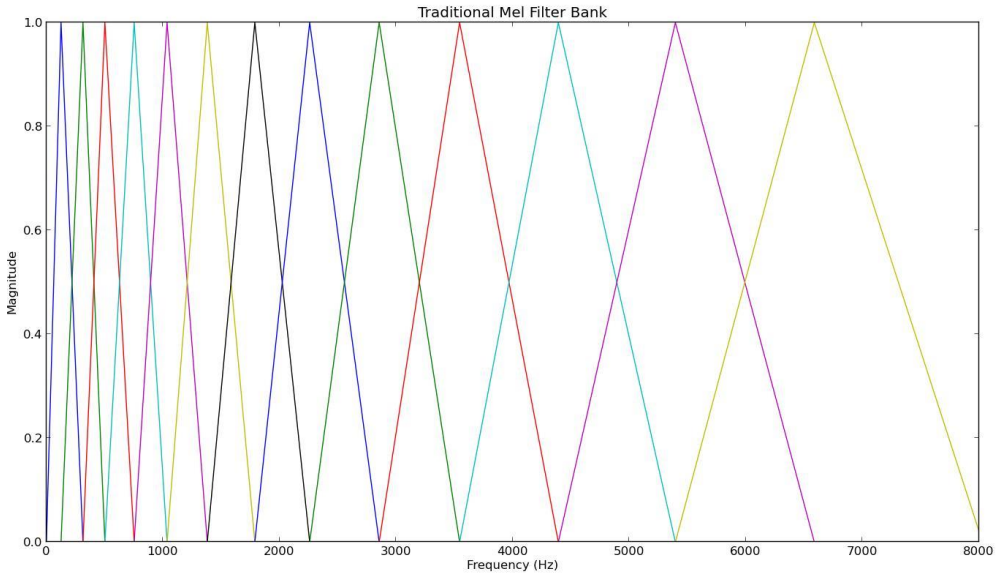


Σχήμα 5: Διάταξη συστήματος εξαγωγής MFC συντελεστών.

Η κλίμακα Mel είναι εξαιρετικά δημοφιλής στην κοινότητα των επιστημόνων που ανήκουν στον τομέα της αυτόματης αναγνώρισης ομιλίας, όπου και χρησιμοποιείται για τον υπολογισμό των MFCC. Η κλίμακα Mel είναι γραμμική στις χαμηλές συχνότητες κάτω των 1000 Hz και λογαριθμική στις ψηλές άνω των 1000 Hz. Οι ζώνες Mel που βασίζονται στην κλίμακα Mel αποτελούν ένα σύνολο από φίλτρα κρίσιμης ζώνης τα οποία δύνανται να μοντελοποιήσουν το σύστημα της ανθρώπινης ακοής και προσεγγίζει τα χαρακτηριστικά της ανθρώπινης ακουστικής αίσθησης. Η μετατροπή από Hz σε κλίμακα Mel πραγματοποιείται βάσει της σχέσης (2.3):

$$\begin{cases} M = f & \text{για } f < 1000, \\ M = f_c\left(1 + \log\frac{f}{f_c}\right) & \text{για } f > 1000. \end{cases} \quad (2.3)$$

Όπου M είναι η συχνότητα εκφρασμένη σε Mel, f, η συχνότητα εκφρασμένη σε Hz και η f_c ισούται με 1000Hz. Στο σχήμα 6 απεικονίζονται οι ζώνες Mel (Peeters, 2004).



Σχήμα 6: Ζώνες Mel *(Anon., n.d.)*

2.2.2 Συντελεστές γραμμικής πρόβλεψης (Linear Prediction Coefficients)

Η ανάλυση γραμμικής πρόβλεψης (LP) (Atal, 1974), (Shikano, 1986), αποτελεί μία μέθοδο για τον υπολογισμό της συνάρτησης μεταφοράς της φωνητικής εκτάσης, υποθέτοντας ότι το ηχητικό σήμα εμπεριέχει μόνο ανθρώπινη φωνή. Στην περίπτωση του μοντέλου LP, ένα σήμα $s(n)$ λογίζεται ως ένας γραμμικός συνδυασμός των προηγούμενών του δειγμάτων. Η προβλεπόμενη τιμή $s_W(n)$ υπολογίζεται από τη σχέση:

$$s_W(n) = \sum_{i=1}^{P} \alpha_i s_W(n-i) + g(n), \quad (2.4)$$

όπου το p αναπαριστά την τάξη της πρόβλεψης, τα α_i προσδιορίζονται ως γραμμικοί συντελεστές πρόβλεψης (LPC) και το $g(n)$ αναπαριστά το σφάλμα του μοντέλου. Οι LPC προσδιορίζονται μέσω της ελαχιστοποίησης του σφάλματος μέσης τετραγωνικής πρόβλεψης του $g(n)$.

2.2.3 Cepstral συντελεστές παραγμένοι από LP (LP-derived Cepstral Coefficients)

Οι LPCC (Atal, 1974) είναι cepstral συντελεστές ενός φάσματος LPC. Η cepstral ανάλυση του φάσματος LPC λειτουργεί και ως μέσο ορθογωνισμού, μία μέθοδος η οποία είναι γνωστή για την αποτελεσματικότητά της σε εφαρμογές αναγνώρισης προτύπων. Οι LPCC $c(n)$ παράγονται από τους LPC μέσω της ακόλουθης εξίσωσης:

$$c(n)\begin{cases} \log\sigma^2 \text{ για } n = 0, \\ \alpha_n + \sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)c(k)\alpha_{n-k} \text{για } 1 \le n \le p, \\ \sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)c(k)\alpha_{n-k} \text{ για } n > p, \end{cases} \quad (2.5)$$

όπου το σ^2 αναπαριστά την ενέργεια του σήματος, το α_n αναπαριστά τους LPC και το p αναπαριστά την τάξη του LPCC.

2.2.4 Mel Cepstral συντελεστές γραμμικής πρόβλεψης (Linear Prediction Mel Cepstral Coefficients)

Οι LPMCC είναι οι mel-cepstral συντελεστές του φάσματος LPC. Εκτός από το ότι μπορούν να εκτελέσουν ορθογωνισμό, οι LPMCC υπερέχουν των LPC επειδή είναι καταλληλότεροι για χρήση σε εφαρμογές όπου είναι απαραίτητη η προσομοίωση της ανθρώπινης ακουστικής αίσθησης. Αυτό είναι και το προτέρημα της χρήσης της συχνοτικής κλίμακας mel. Ένας απλός τρόπος παραγωγής των LPMCC είναι μέσω του υπολογισμού των MFCC του LPC φάσματος.

2.2.5 Cepstral συντελεστές οκταβικής συχνότητας (Octave Frequency Cepstral Coefficients)

Οι cepstral συντελεστές οκταβικής συχνότητας (OSCC/OFCC) προτάθηκαν από τους (Maddage, et al., 2004) και βασίζονται στην οκταβική κλίμακα του πιάνου, όπου η συχνότητα αποκοπής κάθε φίλτρου ζώνης ανατίθεται στη συχνότητα κάθε πλήκτρου μίας οκτάβας ενός πιάνου. Για τον υπολογισμό των OSCC το φάσμα πλάτους του σήματος φιλτράρεται από τριγωνικές τράπεζες φίλτρων, τα οποία φίλτρα τοποθετούνται σε οκτάβες (Monzo, 1998) γραμμικής συχνοτικής κλίμακας.

2.2.6 Φασματικό κέντρο βάρους (Spectral Centroid)

Το φασματικό κέντρο βάρους (spectral centroid) αποτελεί το βαρύκεντρο του φάσματος. Για να υπολογιστεί οφείλουμε να κάνουμε, αρχικώς, την υπόθεση ότι το φάσμα

είναι μία κατανομή, οι τιμές της οποίας είναι οι συχνότητες και οι πιθανότητες παρατήρησής τους είναι το κανονικοποιημένο πλάτος.

$$\mu = \int xp(x)\,\partial x \tag{2.6}$$

όπου x είναι τα παρατηρούμενα δεδομένα:

$$x = freq\_v(x) \tag{2.7}$$

και p(x) είναι η πιθανότητα παρατήρησης x:

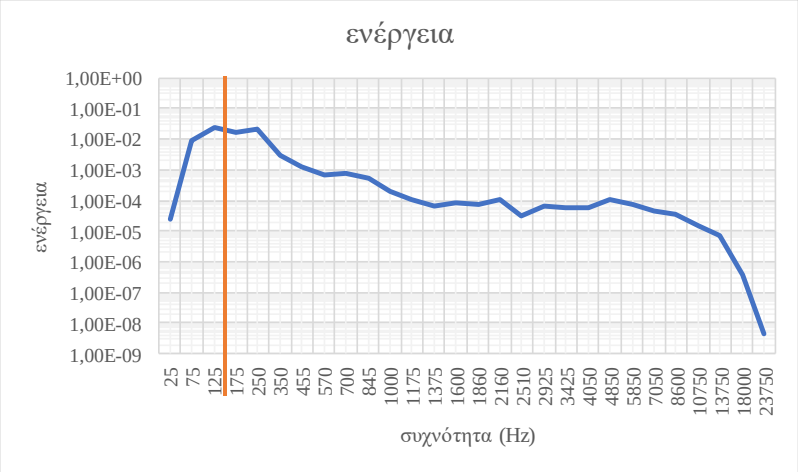
$$p(x) = \frac{ampl\_v(x)}{\sum_x ampl\_v(x)} \tag{2.8}$$

2.2.7 Φασματικό roll-off (Spectrall Roll-Off)

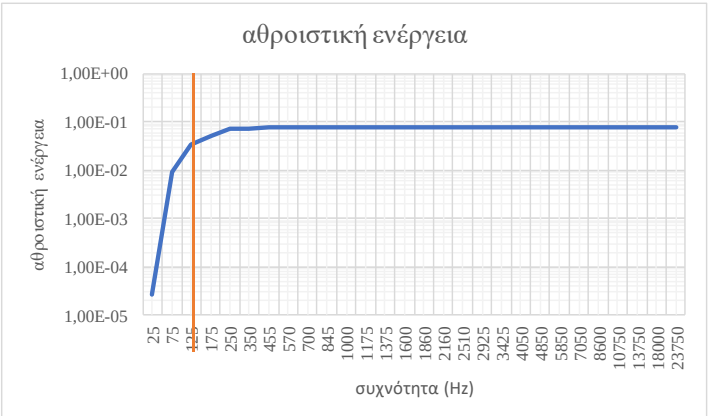
Το σημείο του φασματικού roll-off καθορίζεται από τη συχνότητα, κάτω από την οποία, η ενέργεια του σήματος ανέρχεται στο 95% της συνολικής ενέργειάς του. Συσχετίζεται, κατά κάποιον τρόπο, με τη συχνότητα τμήσης αρμονικών/θορύβου.

$$\sum_0^{f_c} a^2(f) = 0{,}95 \sum_0^{sr/2} a^2(f) \tag{2.9}$$

όπου f_c είναι η συχνότητα φασματικού roll-off, $sr/2$ είναι η συχνότητα Nyquist.



Σχήμα 7: Φάσμα ενέργειας ως προς τη συχνότητα



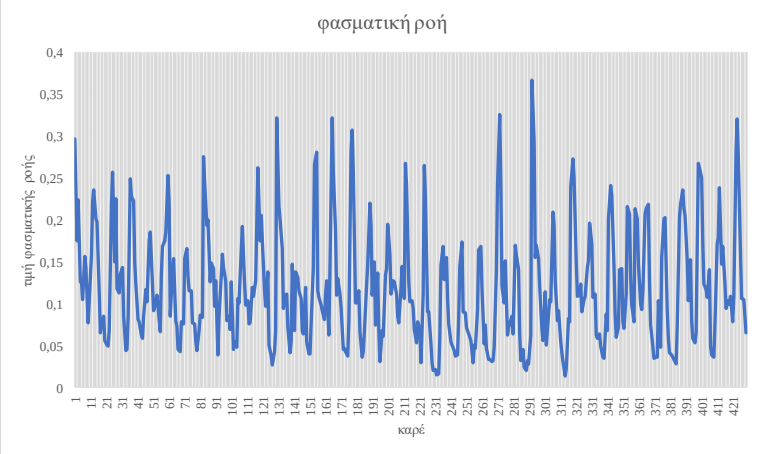
Σχήμα 8: Αθροιστική ενέργεια ως προς τη συχνότητα. Το φασματικό roll-off εντοπίζεται στα 125 Hz

2.2.8 Φασματική ροή (Spectral Flux)

Η φασματική ροή ορίζεται ως το δύο κανόνων, διάνυσμα διαφοράς καρέ από καρέ, στο φασματικό πλάτος. Η σχέση του δίνεται παρακάτω:

$$F_n = \left|\left||X_n(\omega)| - |X_{n+1}(\omega)|\right|\right| \quad (2.10)$$

όπου το |X_n (ω)| αναφέρεται στο φάσμα μέτρου του νιοστού καρέ του ηχητικού σήματος. Η έναρξη των φωνητικών συνήθως εντοπίζεται από την ύπαρξη υψηλών κορυφών στην τιμή της φασματικής ροής. Η φωνή τείνει να έχει ταχύτερο ρυθμό μεταβολών από την οργανική μουσική. Ένα παράδειγμα φασματικής ροής παρουσιάζεται στο σχήμα 9 (Zhang, 2003).



Σχήμα 9: Φασματική ροή ηχητικού αποσπάσματος

2.2.9 Δυναμική πολυπλοκότητα (Dynamic complexity)

Ως δυναμική πολυπλοκότητα ορίζεται η μέση απόλυτη απόκλιση από την προσέγγιση του επιπέδου της συνολικής ακουστότητας υπολογισμένη σε κλίμακα dB. Σχετίζεται με το δυναμικό εύρος και το ποσοστό διακύμανσης της ακουστότητας η οποία λαμβάνει χώρα σε μία ηχογράφηση. Η σιγή στην αρχή και στο τέλος των κομματιών αγνοείται καά τον υπολογισμό προκειμένου να αποφευχθεί η αλλοίωση των αποτελεσμάτων (Streich, 2007).

2.2.10 Φασματική αντίθεση (Spectral contrast)

Ως φασματική αντίθεση λογίζεται η φασματική κορυφή, η φασματική κοιλάδα και η διαφορά τους από κάθε υποζώνη. Συνήθως στη μουσική οι ισχυρές φασματικές κορυφές ανταποκρίνονται περίπου σε αρμονικά στοιχεία, ενώ τα μη αρμονικά χαρακτηριστικά ή οι θόρυβοι συνήθως εμφανίζονται ως φασματικές κοιλάδες. Η φασματική αντίθεση δηλαδή μπορεί χονδρικά να απαραστήσει τη σχετική κατανομή των αρμονικών και των μη αρμονικών στοιχείων του φάσματος (Jiang, et al., n.d.).

2.2.11 Ισχυρή φασματική κορυφή (Spectral Strong Peak)

Ως ισχυρή φασματική κορυφή ορίζεται ο λόγος του μεγέθους της μέγιστης κορυφής του φάσματος και του εύρους ζώνης της κορυφής κάτω από το threshold (κατώφλι). Το μισό είναι το πλάτος. Αυτός ο λόγος αποκαλύπτει αν στο φάσμα βρίσκεται κάποια πολύ τονισμένη μέγιστη κορυφή (για παράδειγμα, όσο πιο λεπτό και ψηλό είναι το μέγιστο του φάσματος, τόσο μεγαλύτερη είναι η τιμή του λόγου).

Να σημειωθεί ότι το εύρος ζώνης ορίζεται ως το πλάτος της κορυφής στο λογαριθμικό με βάση το 10, χώρο. Αυτό είναι διαφορετικό από αυτό που εφαρμόζεται στο (Gouyon & Herrera, 2001). Χρησιμοποιώντας το λογαριθμικό με βάση το 10, χώρο επιτρέπει στον αλγόριθμο να συγκρίνει ισχυρές κορυφές σε χαμηλές συχνότητες με αυτές στις υψηλές.

Εξαίρεση αποτελεί η περίπτωση που το φάσμα εισόδου εμπεριέχει λιγότερα από δύο στοιχεία.

2.2.12 Ακουστότητα (Loudness)

Η ακουστότητα ενός ηχητικού σήματος προσδιορίζεται από το νόμο της δύναμης του Steven. Η ακουστότητα υπολογίζεται ως η ενέργεια του σήματος εις τη δύναμη του 0,67 (Stevens, 1975).

2.2.13 Γαμματονικοί cepstral συντελεστές (GFCC)

Ενώ οι MFCC περνάνε μέσα από ζώνη mel φίλτρου, οι γαμματονικοί cepstral συντελεστές (GFCC) περνάνε μέσα από ζώνη γαμματονικού φίλτρου (Aertsen, et al., 1980). Οι GFCC υπολογίζονται λαμβάνοντας το λογάριθμο του κέντρου βάρους του διακριτού μετασχηματισμού συνημιτόνου (DCT) από το φάσμα ισχύος και ταυτοχρόνως υπολογίζονται και οι δέλτα αλλά και οι δέλτα-δέλτα συντελεστές.

Ένα γαμματονικό φίλτρο είναι ένα γραμμικό φίλτρο το οποίο περιγράφεται από μία κρουστική απόκριση η οποία είναι το προϊόν μίας γάμμα κατανομής και ενός συνημιτονοειδούς τόνου. Μοντελοποιεί τον κοχλία μέσω μιας τράπεζας αλληλεπικαλυπτόμενων ζωνοπερατών φίλτρων (Abdulla, 2002). Ένα σύνολο γαμματονικών φίλτρων με κεντρικές συχνότητες σε αναλογική λογαριθμική διάταξη χρησιμοποιήθηκε για την προσομείωση της βασικής μεμβράνης των κρουστικών αποκρίσεως και των χαρακτηριστικών πλάτους συχνότητας (Shixiong, et al., 2008). Αποτελεί ένα ευρέως διαδεδομένο μοντέλο ακουστικών φίλτρων στο ακουστικό σύστημα.

Η γαμματονική κρουστική απόκριση δίνεται από τη σχέση:

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \varphi) \quad (2.11)$$

όπου f_c είναι η κεντρική συχνότητα, φ η φάση του φορέα, α το πλάτος, n η τάξη του φίλτρου, b το εύρος ζώνης του φίλτρου και t ο χρόνος (Cai, et al., 2011).

2.3 Ταξινομητές

2.3.1 Μηχανές Υποστήριξης Διανύσματος (Support Vector Machines)

Οι μηχανές υποστήριξης διανύσματος (SVM) είναι μία εξαιρετικά χρήσιμη τεχνική στατιστικής μηχανικής μάθησης, η οποία έχει βρει εφαρμογές στον τομέα της αναγνώρισης προτύπων (Vapnik, 1998). Αυτές αποτελούν ένα σύστημα επιτηρούμενης μάθησης το οποίο επιχειρεί να εντοπίσει το μέγιστο όριο ενός υπερεπιπέδου που διαχωρίζει δύο κλάσεις δεδομένων. Αυτές γενικεύονται βέλτιστα σε μελλοντικά δεδομένα. Αυτού του είδους το υπερεπίπεδο ονομάζεται υπερεπίπεδο μέγιστου ορίου και μεγιστοποιεί την απόσταση μεταξύ των κοντινότερων σημείων κάθε κλάσης. Επίσης, οι SVM ελαχιστοποιούν ένα άνω όριο στο αναμενόμενό τους σφάλμα.

Κάθε υπερεπίπεδο που χωρίζει δύο κλάσεις δεδομένων, δοσμένων των σημείων δεδομένων $\{X_0, X_1, X_2, \dots, X_N\}$, καθώς και των ετικετών κλάσης $\{y_0, y_1, y_2, \dots, y_N\}$, όπου $x_i \in R^n$ και $y_i \in \{-1, 1\}$, έχει τη μορφή:

$$y_i(w^T X_i + b) > 0 \forall i, \quad (2.12)$$

ως $\{w_k\}$ συμβολίζεται το σύνολο όλων αυτών των υπερεπιπέδων. Το υπερεπίπεδο μέγιστου ορίου προσδιορίζεται από τη σχέση:

$$w = \sum_{i=0}^{N} \alpha_i y_i X_i, \quad (2.13)$$

και ως b συμβολίζεται το σύνολο των συνθηκών Karush Kuhn Tucker (Burgess, 1998). Το $\{\alpha_0, \alpha_i, \dots, \alpha_N\}$ μεγιστοποιείται στο:

$$L_D = \sum_{i=0}^{N} \alpha_i - \frac{1}{2}\sum_{i=0}^{N}\sum_{j=0}^{N} \alpha_i \alpha_j y_i y_j X_i^T X_j , \quad (2.14)$$

όταν ισχύει:

$$\sum_{i=0}^{N} \alpha_i y_i = 0, \alpha_i \geq 0 \forall i. \quad (2.15)$$

Για γραμμικά διαχωριζόμενα δεδομένα, μόνο ένα υποσύνολο των α_i θα είναι μη μηδενικό. Αυτά τα σημεία ονομάζονται διανύσματα υποστήριξης και όλες οι ταξινομήσεις που εκτελούνται από τα SVM εξαρτώνται μόνο από αυτά τα σημεία. Ένα σύνολο εκπαίδευσης το οποίο παραλείπει όλα τα εναπομείναντα παραδείγματα οδηγεί σε ένα πανομοιότυπο SVM. Για αυτό το λόγο τα SVM αποτελούν ένα εξαιρετικά ελκυστικό συμπλήρωμα στη σχετιζόμενη

ανάδραση. Ο χρόνος εκπαίδευσης και η προσπάθεια επισήμανσης με ετικέτες, των δειγμάτων, είναι δυνατό να μειωθούν σημαντικά, χωρίς να υπάρξει καμία απολύτως επίδραση στην ακρίβεια του αλγόριθμου κατηγοριοποίησης. Αυτό μπορεί να συμβεί εάν το σύστημα ανάδρασης είναι ικανό να αναγνωρίσει με ακρίβεια τα κρίσιμα δείγματα, τα οποία θα αποτελέσουν και τα διανύσματα υποστήριξης.

Τα σημεία δεδομένων X συμμετέχουν στους υπολογισμούς μόνο ως σημειακά προϊόντα. Ο μετασχηματισμός τους σε άλλο χώρο χαρακτηριστικών μπορεί να πραγματοποιηθεί μέσω μίας συνάρτησης $\Phi(X)$. Η αναπαράσταση των δεδομένων σε αυτό το χώρο χαρακτηριστικών δεν είναι αναγκαίο να υπολογιστεί αναλυτικά στην περίπτωση της ύπαρξης ενός κατάλληλου τελεστή πυρήνα Mercer για τον οποίο θα ισχύει:

$$K(X_i, X_j) = \Phi(X_i) \cdot \Phi(X_j). \quad (2.16)$$

Εάν στόχος είναι ο διαχωρισμός δεδομένων εκπαίδευσης σε δύο κλάσεις και τα δεδομένα δεν είναι δεν είναι δυνατό να διαχωριστούν με γραμμικό τρόπο στο χώρο των χαρακτηριστικών, κάτι το οποίο είναι εξαιρετικά σύνηθες, αλλά είναι δυνατό να διαχωριστούν μη γραμμικά, εφαρμόζεται ένας μη γραμμικός αλγόριθμος κατηγοριοποίησης SVM. Η διαδικασία περιλαμβάνει το μετασχηματισμό των διανυσμάτων εισόδου σε έναν υψηλό χώρο χαρακτηριστικών με τη χρήση του μη γραμμικού μετασχηματισμού Φ, με τον τρόπο που περιγράφηκε παραπάνω και το μετέπειτα γραμμικό διαχωρισμό τους στο χώρο των χαρακτηριστικών.

Για την κατασκευή ενός γραμμικού αλγόριθμου κατηγοριοποίησης SVM, το εσωτερικό προϊόν $\langle X_i, X_j \rangle$ αντικαθίσταται από μία συνάρτηση πυρήνα $K(X_i, X_j)$.

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right) \quad (2.17)$$

Με αυτό τον τρόπο υπάρχει η δυνατότητα να προβληθούν σε αυτό το χώρο υψηλότερης διάστασης μέσω του πυρήνα Mercer $K(\cdot)$ που προαναφέρθηκε. Στην πραγματικότητα, μόνο τα εσωτερικά προϊόντα των σημείων δεδομένων είναι απαραίτητα σε αυτό το χώρο υψηλότερης διάστασης. Με αυτό τον τρόπο η προβολή μπορεί να είναι έμμεση εάν ένα τέτοιο εσωτερικό προϊόν δύναται να υπολογιστεί άμεσα.

Ο SV αλγόριθμος είναι ικανός να κατασκευάσει μία πληθώρα από μηχανές μάθησης μέσω της χρήσης διάφορων συναρτήσεων πυρήνα. Οι πιο συχνά χρησιμοποιούμενες συναρτήσεις πυρήνα είναι ο πολυωνυμικός πυρήνας βαθμού d, η συνάρτηση ακτινικής βάσης

με Γκαουσιανό πυρήνα πλάτους $c > 0$ και τα νευρωνικά δίκτυα με συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη. Ο πολύωνυμικός πυρήνας βαθμού d εκφράζεται από τη σχέση:

$$K(X_i, X_j) = (\langle X_i, X_j \rangle + 1)^d, \quad (2.18)$$

η συνάρτηση ακτινικής βάσης με Γκαουσιανό πυρήνα πλάτους $c > 0$ περιγράφεται ως:

$$K(X_i, X_j) = \exp\left(-\|X_i - X_j\|^2 / c\right), \quad (2.19)$$

ενώ για τα νευρωνικά δίκτυα με συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη ισχύει:

$$K(X_i, X_j) = \tanh(k\langle X_i, X_j \rangle + \mu), \quad (2.20)$$

όπου οι παράμετροι k και μ ονομάζονται, αντίστοιχα, κέρδος και μετατόπιση.

Ο γενικός τύπος της συνάρτησης ακτινικής βάσης (RBF) είναι:

$$K(X_i, X_j) = e^{-\gamma D^2(X_i, X_j)}, \quad (2.21)$$

όπου στη θέση του $D^2(X_i, X_j)$ θα μπορούσε να υπάρχει οποιαδήποτε συνάρτηση απόστασης (Mandel, et al., 2006).

Ο χώρος των πιθανών συναρτήσεων ταξινόμησης αποτελείται από σταθμισμένους γραμμικούς συνδυασμούς, σταθμισμένων Γκαουσιανών, γύρω από στρατηγικά επιλεγμένες περιπτώσεις εκπαίδευσης, στο χώρο του πυρήνα (Christianini & Shawe-Taylor, 2000). Ο SVM αλγόριθμος εκπαίδευσης επιλέγει αυτές τις περιπτώσεις, τα «διανύσματα υποστήριξης» όπως λέγονται και τα σταθμίζει προκειμένου να βελτιστοποιήσει το περιθώριο μεταξύ των ορίων του ταξινομητή και των παραδειγμάτων εκπαίδευσης. Η χρήση ολόκληρων τραγουδιών ως παραδειγμάτων εκπαίδευσης αποτελούν μία εξαιρετικά σωστή προσέγγιση στην εργασία της κατηγοριοποίησης τραγουδιού, αφού τα παραδείγματα αυτά εφαρμόζονται απευθείας στην κατηγοριοποίηση (Mandel & Ellis, 2005).

2.3.2 Πολυεπίπεδο δίκτυο πρόσθιας τροφοδότησης (Multilayer Perceptron)

Το (MLP) αποτελεί ένα μοντέλο νευρωνικού δικτύο πρόσθιας τροφοδότησης το οποίο αναθέτει ένα σύνολο δεδομένων εισόδου σε ένα σύνολο κατάλληλων εξαγόμενων. Αυτό αποτελεί μία τροποποίηση του συμβατικού γραμμικού perceptron που χρησιμοποιεί τρία ή περισσότερα επίπεδα νευρώνων με μη γραμμικές συναρτήσεις ενεργοποίησης. Ωστόσο είναι πιο ισχυρό από τον percepton γιατί είναι ικανό να διαχωρίσει δεδομένα τα όποία δεν δύνανται να διαχωριστούν γραμικά ή με χρήση υπερεπιπέδου.

Ένα MLP που αποτελείται από νευρώνες χρησιμοποιεί μία μη γραμμική συνάρτηση ενεργοποίησης, η οποία αναπτύχθηκε προκειμένου να μοντελοποιήσει τη συχνότητα των πιθανών ενεργειών των βιολογικών νευρώνων του εγκεφάλου. Αυτή η συνάρτηση μοντελοποιείται με διάφορους τρόπους αλλά πρέπει να είναι πάντα κανονικοποιημένη και διαφοροποιημένη.

Οι δύο κύριες συναρτήσεις ενεργοποίησης που χρησιμοποιούνται σε τρέχουσες εφαρμογές είναι και οι δύο σιγμοειδείς και είναι οι εξής:

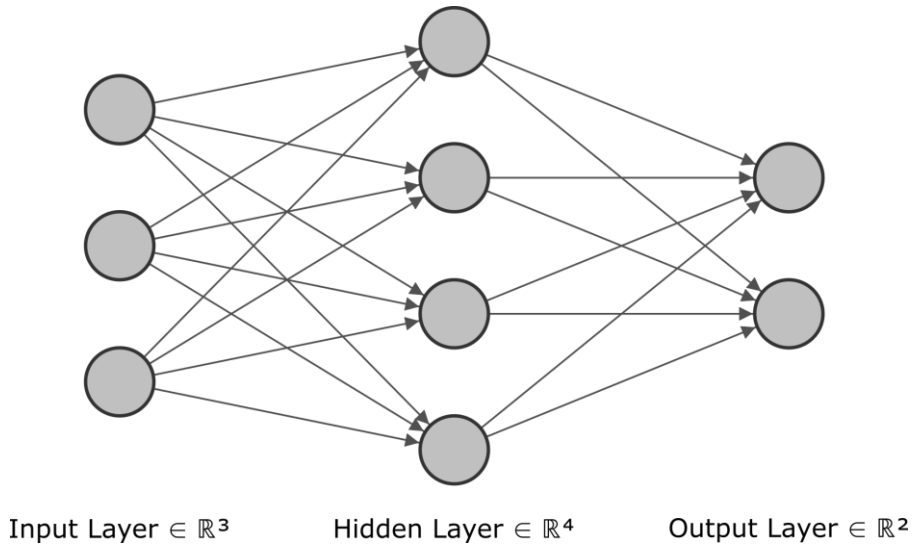
$$\varphi(v_i) = \tanh(v_i) \tag{2.22}$$

και

$$\varphi(v_i) = (1 + e^{v_i})^{-1} \tag{2.23}$$

Η (2.21) αποτελεί μία υπερβολική εφαπτομένη η οποία κυμαίνεται από το -1 μέχρι και το 1, ενώ η (2.22) είναι ίδιου σχήματος αλλά κυμαίνεται από το 0 έως και το 1. Όπου y_i είναι η έξοδος του ιοστού κόμβου (νευρώνα) και v_i είναι το σταθμισμένο άθροισμα των συνάψεων εισόδου.

Το MLP αποτελείται από ένα επίπεδο εισόδου και ένα εξόδου με ένα ή περισσότερα επίπεδα μη γραμμικά ενεργοποιήσιμων κόμβων. Κάθε κόμβος ενός επιπέδου συνδέεται με ένα συγκεκριμένο βάρος w_{ij} σε κάθε επόμενο κόμβο του επιπέδου που ακολουθεί. Στο σχήμα 10 παρουσιάζεται ένα βασικό MLP.



Σχήμα 10: Ένα νευρωνικό δίκτυο αποτελεί μία ομάδα διασυνδεδεμένων κόμβων.

Στο perceptron η εκμάθηση συμβαίνει μέσω την μεταβολής των βαρών των συνδέσεων ή των συναπτικών βαρών, μετά το πέρας της ολοκλήρωσης της επεξεργασίας κάθε τμήματος

δεδομένων. Η μεταβολή των βαρών των συνδέσεων βασίζεται στο ποσοστό του σφάλματος στην έξοδο σε σύγκριση με το αναμενόμενο αποτέλεσμα. Η εκμάθηση εκτελείται με χρήση της μεθόδου της οπίσθιας διάδοσης (backpropagation). Αυτή αποτελεί μία γενίκευση του αλγόριθμου των ελάχιστων μέσων τετραγώνων στο επίπεδο perceptron. Η αναπαράσταση του σφάλματος στον κόμβο εξόδου j στο νιοστό σημείο δεδομένων πραγματοποιείται μέσω της σχέσης:

$$\mathrm{e_j(n) = d_j(n) - y_j(n)} \tag{2.24}$$

όπου d είναι η τιμή στόχου και y είναι η τιμή που παράγεται από το perceptron. Έπειτα πραγματοποιούνται διορθώσεις στα βάρη των κόμβων. Οι διορθώσεις αυτές ελαχιστοποιούν την ενέργεια του σφάλματος σε ολόκληρη την έξοδο. Το σφάλμα αυτό δίνεται από τη σχέση:

$$\varepsilon(n) = \frac{1}{2}\sum_j e_j^2(n) \tag{2.25}$$

Μέσω της θεωρίας των διαφορικών, η μεταβολή σε κάθε βάρος προσδιορίζεται μέσω της σχέσης:

$$\Delta w_{ji}(n) = -\eta \frac{\partial \varepsilon(n)}{\partial v_j(n)} y_i(n) \tag{2.26}$$

όπου $\mathrm{y_i}$ είναι η έξοδος του προηγούμενου νευρώνα και η είναι ο ρυθμός εκμάθησης. Ο ρυθμός εκμάθησης επιλέγεται προσεκτικά προκειμένου να διαβεβαιωθεί ότι τα βάρη συγκεντρώνονται σε μία απόκριση η οποία δεν είναι ούτε πολύ συγκεκριμένη αλλά ούτε πολύ γενική. Σε εφαρμογές προγραμματισμού, εκτείνεται συνήθως από 0,2 μέχρι και 0,8. Η παράγωγος προς υπολογισμό εξαρτάται από το άθροισμα της σύναψης εισόδου v_j,το οποίο μεταβάλλεται. Είναι εύκολο να αποδειχθεί ότι για έναν κόμβο εξόδου αυτή η παράγωγος είναι δυνατό να απλοποιηθεί μέσω της σχέσης:

$$-\frac{\partial \mathrm{\varepsilon(n)}}{\partial \mathrm{v_j(n)}} = \mathrm{e_j(n)\varphi'\left(v_j(n)\right)} \tag{2.27}$$

όπου φ είναι η παράγωγος της συνάρτησης ενεργοποίησης που προαναφέρθηκε και παραμένει αμετάβλητη. Η ανάλυση παρουσιάζει μεγαλύτερες δυσκολίες στην περίπτωση της αλλαγής των βαρών εντός ενός κρυφού κόμβου και η σχετική παράγωγος δίνεται από τη σχέση:

$$-\frac{\partial \varepsilon(n)}{\partial v_j(n)} = \varphi'\left(v_j(n)\right)\sum_k -\frac{\partial \varepsilon(n)}{\partial v_k(n)} w_{kj}(n) \tag{2.28}$$

Η σχέση αυτή βασίζεται στη μεταβολή των βαρών του k πλήθους των κόμβων, το οποίο αναπαριστά το επίπεδο εξόδου. Έτσι, προκειμένου να μεταβληθούν τα βάρη του κρυμμένου επιπέδου πρέπει πρώτα να μεταβληθούν τα βάρη σύμφωνα με την παράγωγο της συνάρτησης

ενεργοποίησης. Εν κατακλείδι, αυτός ο αλγόριθμος αναπαριστά την οπίσθια διάδοση (backpropagation) της συνάρτησης ενεργοποίησης (Karaman, 2009).

2.3.3 Αλγόριθμος C4.5

Ο C4.5 (Quinlan, 1993) αποτελεί ένα σύστημα κατασκευής ταξινομητών. Αυτά τα συστήματα χρησιμοποιούνται ευρέως στην εξόρυξη δεδομένων (data mining). Λειτουργούν ως εξής, λαμβάνουν ως είσοδο ένα σύνολο περιπτώσεων, η καθεμία εκ των οποίων ανήκει σε ένα μικρότερο αριθμό κλάσεων. Οι κλάσεις περιγράφονται από τις αξίες τους (για ένα σταθερό σύνολο χαρακτηριστικών) και εξάγουν έναν ταξινομητή, ο οποίος μπορεί να προβλέψει με ακρίβεια την κλάση στην οποίο ανήκει η κάθε νέα περίπτωση.

Ο C4.5 είναι ο απόγονος του CLS (Hunt, et al., 1966) και του ID3 (Quinlan, 1979). Όπως ο CLS και ο ID3, έτσι και ο C4.5 δημιουργεί ταξινομητές, οι οποίοι λειτουργούν ως δέντρα απόφασης. Έχει επίσης τη δυνατότητα κατασκευής ταξινομητών σε μία πιο κατανοητή μορφή συνόλου κανόνων.

Ο C4.5 αρχικά αναπτύσσει ένα δέντρο από ένα σύνολο δοσμένων περιπτώσεων S. Για τη δημιουργία του δέντρου, χρησιμοποιείται ο αλγόριθμος «διαίρει και βασίλευε» ως εξής:

- Αν όλες οι περιπτώσεις στο S ανήκουν στην ίδια κλάση ή το S είναι μικρό, τα φύλλα του δέντρου επισημειώνονται με την κλάση που συναντάται συχνότερα στο S.
- Διαφορετικά, επιλέγουμε μία δοκιμαστική η οποία βασίζεται σε ένα μόνο χαρακτηριστικό με δύο ή περισσότερα αποτελέσματα. Κάνουμε αυτήν τη δοκιμαστική κλάση τη ρίζα του δέντρου, με μία διακλάδωση για κάθε εξαγόμενο της δοκιμαστικής. Στη συνέχεια, διαχωρίζουμε το S σε ανταποκρινόμενα υποσύνολα S1, S2, … στο εξαγόμενο της κάθε περίπτωσης. Έπειτα επαναλαμβάνουμε την ίδια διαδικασία για κάθε υποσύνολο.

Πολλές δοκιμαστικές κλάσεις είναι δυνατό να επιλεγούν για αυτό το τελευταίο βήμα. Ο C4.5 χρησιμοποιεί δύο κριτήρια για την εύρεσή τους. Το ένα από αυτά είναι το κέρδος πληροφορίας. Αυτό ελαχιστοποιεί τη συνολική εντροπία των υποομάδων $\{S_i\}$. Το αρνητικό αυτού του κριτηρίου είναι ότι τείνει να μεροληπτεί υπέρ των δοκιμαστικών κλάσεων με πολλά εξαγόμενα. Επίσης, ελαχιστοποιεί τον προεπιλεγμένο λόγο κέρδους, ο οποίος διαχωρίζει το κέρδος των πληροφοριών, από τις πληροφορίες των αποτελεσμάτων της δοκιμής.

Τα χαρακτηριστικά μπορούν να είναι είτε αριθμητικά είτε ονομαστικά και αυτά είναι που καθορίζουν τη μορφή των εξαγομένων της δοκιμής. Για ένα αριθμητικό χαρακτηριστικό A ισχύει $\{A \leq h, A > h\}$ όπου το κατώφλι h βρίσκεται ταξινομώντας τις S στις τιμές του A και

επιλέγοντας το διαχωρισμό μεταξύ των διαδοχικών τιμών που μεγιστοποιούν το παραπάνω κριτήριο. Κάθε χαρακτηριστικό Α με διακριτές τιμές διαθέτει ένα προεπιλεγμένο εξαγόμενο για κάθε υποσύνολο.

Στη συνέχεια, το αρχικό δέντρο «κλαδεύεται» προκειμένου να αποτραπεί η υπερπροσαρμογή. Ο αλγόριθμος «κλαδέματος» βασίζεται σε μία απαισιόδοξη προσέγγιση του ποσοστού των σφαλμάτων για ένα σύνολο Ν περιπτώσεων, οι Ε από τις οποίες δεν ανήκουν στην πιο συχνή κλάση. Αντί του λόγου Ε/Ν, ο C4.5 προσδιορίζει το άνω όριο της διωνυμικής ικανότητας όταν Ε γεγονότα έχουν παρατηρηθεί σε Ν δοκιμές, με μία βεβαιότητα η οποία καθορίζεται από το χρήστη (η προκαθορισμένη τιμή αυτής είναι η 0.25).

Το «κλάδεμα» εκτελείται από τα φύλλα προς τη ρίζα. Για κάθε υποδέντρο, ο C4.5 προσθέτει τα υπολογισμένα σφάλματα των βρόγχων και συγκρίνει αυτό το αποτέλεσμα με τα υπολογισμένα σφάλματα των υποδέντρων. Στην περίπτωση που το υποδέντρο αντικατασταθεί από ένα φύλλο (αν δηλαδή το φύλλο δεν έχει μεγαλύτερο αριθμό σφαλμάτων από το υποδέντρο), το δέντρο «κλαδεύεται». Παρομοίως, ο C4.5 ελέγχει το υπολογισμένο σφάλμα αν το υποδέντρο αντικατασταθεί από έναν εκ των βρόγχων του. Όποτε κρίνεται ωφέλιμο το δέντρο μπορεί να τροποποιηθεί αναλόγως. Η διαδικασία «κλαδέματος» ολοκληρώνεται σε ένα πέρασμα του δέντρου.

Ο C4.5 αντικαταστάθηκε το 1997 από ένα εμπορικό σύστημα, το See5/C5.0 (ή C5.0) για συντομία, ο οποίος προσθέτει νέες δυνατότητες στο ήδη υπάρχον σύστημα και βελτιώνει κατά πολύ την απόδοση. Περιλαμβάνει ενδεικτικά:

- Μία παραλλαγή του boosting (Freund & Schapire, 1997), το οποίο κατασκευάζει ένα σύνολο ταξινομητών που στη συνέχεια ορίζουν την τελική κατηγοριοποίηση. Το boosting συχνά βελτιώνει δραματικά την προβλεπόμενη ακρίβεια.
- Νέους τύπους δεδομένων, για παράδειγμα ημερομηνίες, μη εφαρμόσιμες αξίες, μεταβλητές τιμές εσφαλμένης κατηγοριοποίησης και μηχανισμούς προγενέστερου φιλτραρίσματος χαρακτηριστικών.
- Σύνολα κανόνων χωρίς σειρά προτεραιότητας. Όταν δηλαδή μία περίπτωση κατηγοριοποιείται, όλοι οι εφαρμόσιμοι κανόνες εντοπίζονται και συμμετέχουν στην κατηγοριοποίηση. Με αυτόν τον τρόπο βελτιώνεται τόσο η ερμηνευσιμότητα, όσο και η ακρίβεια πρόβλεψής τους.

- Βελτιωμένη δυνατότητα κλιμάκωσης, τόσο των δέντρων απόφασης, όσο και συγκεκριμένων συνόλων κανόνων. Η κλιμάκωση βελτιώνεται με τη χρήση πολλαπλών thread του υπολογιστή, αφού το C5.0 είναι ικανό να χρησιμοποιήσει τους πολλαπλούς πυρήνες και επεξεργαστές των σύγχρονων υπολογιστών (Wu, et al., 2007).

2.3.4 Δείκτες επαλήθευσης αλγορίθμων (TP, FP, TN, FN, ανάκληση, ακρίβεια, μέτρηση-f)

Παρακάτω ακολουθούν οι περιγραφές των πινάκων και των σχημάτων που επισυνάπτονται και παρουσιάζουν τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν.

Πίνακας 1: Ο πίνακας σύγχυσης ενός προβλήματος δύο κλάσεων.

	αρνητική πρόβλεψη	θετική πρόβλεψη
πραγματικό αρνητικό (TN)	a	b
πραγματικό θετικό (TP)	c	d

Ένας πίνακας σύγχυσης (confusion matrix) μεγέθους n × n σε συνεργασία με έναν ταξινομητή, μας προσφέρει την ικανότητα πρόβλεψης της ακριβούς κατηγοριοποίησης (όπου n είναι ο αριθμός των διαφορετικών κλάσεων). Ο πίνακας 1 παρουσιάζει έναν πίνακα σύγχυσης για n = 2. Οι καταχωρήσεις του συμβολίζουν τα εξής:

- με “a” συμβολίζεται το πλήθος των σωστών αρνητικών προβλέψεων (tn),
- με “b” το πλήθος των εσφαλμένων θετικών (fp),
- με “c” το πλήθος των εσφαλμένων αρνητικών (fn)
- και με “d” το πλήθος των σωστών θετικών (tp).

Η ακρίβεια της πρόβλεψης και το σφάλμα ταξινόμησης μπορούν να εξαχθούν από αυτόν τον πίνακα μέσω των παρακάτω σχέσεων:

$$\text{Ακρίβεια} = \frac{a+d}{a+b+c+d} \tag{2.29}$$

$$\Sigma\text{φάλμα} = \frac{b+c}{a+b+c+d} \tag{2.30}$$

Ορίζουμε το αποτέλεσμα της διαφωνίας που σχετίζεται με τον πίνακα σύγχυσης στην εξίσωση (3). Σύμφωνα με αυτήν την εξίσωση η διαφωνία ισούται με 1 όταν μία από τις ποσότητες b ή c είναι 0. Στην περίπτωση αυτή ο ταξινομητής κατηγοριοποιεί λανθασμένα τα παραδείγματα μόνο της μίας κλάσης. Η διαφωνία ισούται με 0 όταν το b και το c είναι τα ίδια.

$$D = \begin{cases} 0 & \text{αν } b = c = 0, \\ \frac{|b-c|}{\max\{b,c\}} & \text{αλλιώς.} \end{cases} \quad (2.31)$$

Μία μεθοδολογία για την επιλογή χαρακτηριστικών είναι η επιλογή μόνο εκείνων που έχουν μία καλή δυνατότητα διαχωρισμού από μόνα του και συμπληρώνουν το ένα το άλλο. Ας πάρουμε για παράδειγμα δύο χαρακτηριστικά A1 και A2, τα οποία έχουν παρόμοια ακρίβεια κατηγοριοποίησης. Αυτά θα τα χρησιμοποιήσουμε ως ένα υποσύνολο χαρακτηριστικών εάν και εφόσον έχουν έντονη διαφωνία όσον αφορά τα παραδείγματα που κατηγοριοποιούν λάθος. Μία έντονη διαφωνία παρατηρείται για τις τιμές του D οι οποίες είναι πιο κοντά στο 1 και στα δύο χαρακτηριστικά, αλλά όταν τα τελευταία έχουν και διαφορετικούς παρονομαστές στην εξίσωση.

Πραγματικά θετικά (true positive) στιγμιότυπα μίας κλάσης $i \in |1, \dots, k|$ ονομάζεται το πλήθος εκείνων, τα οποία ο αλγόριθμος ταξινόμησε σωστά. Το πλήθος τους δίνεται από την παρακάτω σχέση:

$$tp_i = \sum y'_i | Y = y_i \quad (2.32)$$

Εσφαλμένα θετικά (false positive) καλείται το πλήθος των παραδειγμάτων εξέτασης που ταξινομήθηκαν σε μία κατηγορία μίας κλάσης i, χωρίς να ανήκουν σε αυτήν. Το πλήθος τους δίνεται από τη σχέση:

$$fp_i = \sum y'_i | Y \neq y_i \quad (2.33)$$

Ανάκληση (Recall) ή ευαισθησία (Sensitivity — όπως ονομάζεται στην ψυχολογία) ονομάζεται το ποσοστό των πραγματικών θετικών περιπτώσεων που έχουν προβλεφθεί σωστά ως θετικές. Αυτή καταμετρά της κάλυψη των αληθινών θετικών περιπτώσεων μέσω του κανόνα αυτών που έχουν προβλεφθεί ως θετικές (+P). Η ανάκληση προσδιορίζεται από την εξίσωση:

$$R = \text{ανάκληση} = \text{ευαισθησία} = tpr = \frac{tp}{tp+fn} = \frac{tp}{rp} = \frac{TP}{RP} = \frac{A}{(A+C)} \quad (2.34)$$

Η **ακρίβεια (Precision) ή αυτοπεποίθηση** (Confidence — όπως ονομάζεται στην εξόρυξη δεδομένων) δηλώνει το ποσοστό των περιπτώσεων που έχουν προβλεφθεί ως θετικές και είναι σωστά πραγματικά θετικές. Η ακρίβεια προσδιορίζεται από την εξίσωση (2.36):

$$P = \text{ακρίβεια} = \text{αυτοπεποίθηση} = tpa = \frac{tp}{tp+fp} = \frac{tp}{pp} = \frac{TP}{PP} = \frac{A}{(A+B)} \quad (2.35)$$

Με τον όρο μέτρηση — f ονομάζουμε τον αρμονικό μέσο της ακριβείας και της ανάκλησης. Αυτή είναι ένα συγκεντρωτικό μέτρο που αφορά την ικανότητα ταξινόμησης της κλάσης $i \in |1, \dots, k|$. Η μέτρηση — f προσδιορίζεται από την εξίσωση:

$$F = \frac{1}{\left(a\frac{1}{P} + (1-a)\frac{1}{R}\right)} \tag{2.36}$$

όπου α: ο συντελεστής που δηλώνει το βάρος προς το P ή το R.

Κεφάλαιο 3ο: Βάσεις δεδομένων

Υπάρχουν πολλές βάσεις δεδομένων για πολλά ζητήματα ανάκτησης πληροφορίας και οι περισσότεροι ερευνητές τείνουν να δημιουργούν τις δικές τους για τις ανάγκες των πειραμάτων τους.

3.1 Οι βάσεις δεδομένων που έχουν χρησιμοποιηθεί από άλλους ερευνητές

(Zhang, 2003): Η βάση δεδομένων που χρησιμοποιήθηκε από τον Tong Zhang, το 2003, στο δημοσίευσή του με τίτλο "Automatic Singer Identification", αποτελούνταν από 45 τραγούδια ερμηνευμένα από 8 τραγουδιστές, τέσσερις άντρες και τέσσερις γυναίκες, τέσσερις Βρετανοί και τέσσερις Κινέζοι. Στον κάθε τραγουδιστή αντιστοιχούσαν τρία με εννιά δείγματα. Τα κομμάτια αυτά είναι ποικίλων τεχνοτροπιών, άλλα έχουν αρκετά γρήγορο ρυθμό, άλλα σχετικά αργό. Στα περισσότερα εξ αυτών η φωνή του τραγουδιστή αποτελεί κυρίαρχο στοιχείο στο μεγαλύτερό τους μέρος, ή τουλάχιστον σε πολλά σημεία. Τα υπόλοιπα, ωστόσο, διέπονται από μία ισχυρή μουσική συνοδεία στο μεγαλύτερο μέρος τους.

(Bartsch & Wakefield, 2004): Η βάση δεδομένων των Bartsch και Wakefield στη δημοσίευση "Singing Voice Identification Using Spectral Envelope Estimation", ήταν μία συλλογή από ηχογραφήσεις φωνών, δώδεκα εκπαιδευμένων γυναικών, στην κλασική μουσική. Οι ηχογραφήσεις αυτές πραγματοποιήθηκαν στο σχολείο μουσικής του Μίσιγκαν. Οι πέντε εξ αυτών των τραγουδιστριών ήταν μέτζο-σοπράνο (m01, …, m05) και οι υπόλοιπες επτά ήταν σοπράνο (s01, …, s07). Αυτές εκτέλεσαν μία σειρά από ασκήσεις με πέντε νότες, με τις εξής βαθμίδες: do-re-mi-re-do. Ερμηνεύθηκαν από τον κάθε τραγουδιστή τα εξής ιταλικά φωνήεντα /α/. /ε/, /i/, /o/ και /u/, σε ημιτονικά διαστήματα, για πάνω από δύο οκτάβες. Οι τραγουδιστές καθοδηγήθηκαν να τραγουδήσουν με δυναμικές που μπορούσαν να τις εκτελέσουν άνετα, παράγοντας ταυτόχρονα έναν πλήρη τόνο με κανονικό βιμπράτο. Όλα τα δείγματα ήχου ήταν μονοφωνικά και με συχνότητα δειγματοληψίας 44100 Hz.

(Maddage, et al., 2004): Η βάση δεδομένων των Maddage, Xu και Wang αποτελείται από 110 κομμάτια, ερμηνευμένα από 8 τραγουδιστές, οι πέντε εκ των οποίων είναι άντρες και οι τρεις γυναίκες. Τα κομμάτια συλλέχθηκαν από εμπορικούς συμπαγείς δίσκους με συχνότητα δειγματοληψίας 44100 Hz, 16 bit ανά δείγμα και σε στερεοφωνική μορφή. Η βάση δεδομένων περιέχει 14 κομμάτια ερμηνευμένα στα Αγγλικά, 28 κομμάτια στα κινέζικα, καθώς και 18 στη γλώσσα Σινχάλα.

DB-S-1 (Tsai & Wang, 2006)**:** Η DS-S-1 αποτελούνταν από 200 σόλο κομμάτια, τα οποία ερμηνεύονταν από 10 άντρες και 10 γυναίκες, με δέκα τραγούδια ανά τραγουδιστή. Όλα τα κομμάτια αλιεύθηκαν από μουσικά CD Μανδαρινικής ποπ μουσικής και τα τμήματα με φωνητικά και χωρίς, καθώς και με την ταυτότητα του τραγουδιστή επισημειώθηκαν χειροκίνητα. Η διάρκεια των κομματιών κυμαινόταν από 135 έως 391 s και ο ρυθμός δειγματοληψίας όλων των κομματιών κατέβηκε από τα 44100 Hz στα 22050 Hz.

DB-S-2 (Tsai & Wang, 2006)**:** Η DB-S-2 αποτελούνταν από 42 σόλο κομμάτια, τα οποία ερμηνεύονταν από 13 γυναίκες και οκτώ άντρες. Κανένας από τους ερμηνευτές και τις ερμηνεύτριες των κομματιών της DB-S-2 δεν περιλαμβανόταν στην DB-S-1. Όλα τα κομμάτια αλιεύθηκαν από μουσικά CD Μανδαρινικής ποπ μουσικής και τα τμήματα με φωνητικά και χωρίς, καθώς και με την ταυτότητα του τραγουδιστή επισημειώθηκαν χειροκίνητα. Η διάρκεια των κομματιών κυμαινόταν από 135 έως 391 s και ο ρυθμός δειγματοληψίας όλων των κομματιών κατέβηκε από τα 44100 Hz στα 22050 Hz.

DB-S-1-E (Tsai & Wang, 2006)**:** Η DB-S-1, χωρίστηκε σε δύο υποομάδες. Η μία εξ αυτών ονομάστηκε DB-S-1-T και περιείχε πέντε κομμάτια ανά τραγουδιστή. Όλα τα κομμάτια αλιεύθηκαν από μουσικά CD Μανδαρινικής ποπ μουσικής και τα τμήματα με φωνητικά και χωρίς, καθώς και με την ταυτότητα του τραγουδιστή επισημειώθηκαν χειροκίνητα. Η διάρκεια των κομματιών κυμαινόταν από 135 έως 391 s και ο ρυθμός δειγματοληψίας όλων των κομματιών κατέβηκε από τα 44100 Hz στα 22050 Hz.

DB-S-1-T (Tsai & Wang, 2006)**:** Η DB-S-1-E περιέχει τα υπόλοιπα πέντε κομμάτια από την DB-S-1 που δεν εμπεριέχονται στην DB-S-1-E. Όλα τα κομμάτια αλιεύθηκαν από μουσικά CD Μανδαρινικής ποπ μουσικής και τα τμήματα με φωνητικά και χωρίς, καθώς και με την ταυτότητα του τραγουδιστή επισημειώθηκαν χειροκίνητα. Η διάρκεια των κομματιών κυμαινόταν από 135 έως 391 s και ο ρυθμός δειγματοληψίας όλων των κομματιών κατέβηκε από τα 44100 Hz στα 22050 Hz.

DB-D (Tsai & Wang, 2006)**:** Η DB-D εμπεριέχει 22 ντουέτα. Όλα τα κομμάτια αλιεύθηκαν από μουσικά CD Μανδαρινικής ποπ μουσικής και τα τμήματα με φωνητικά και χωρίς, καθώς και με την ταυτότητα του τραγουδιστή επισημειώθηκαν χειροκίνητα. Η διάρκεια των κομματιών κυμαινόταν από 135 έως 391 s και ο ρυθμός δειγματοληψίας όλων των κομματιών κατέβηκε από τα 44100 Hz στα 22050 Hz.

DB-I (Tsai & Wang, 2006)**:** Η DB-I αποτελείται από 174 ορχηστρικά κομμάτια. Όλα τα κομμάτια αλιεύθηκαν από μουσικά CD Μανδαρινικής ποπ μουσικής και τα τμήματα με

φωνητικά και χωρίς, καθώς και με την ταυτότητα του τραγουδιστή επισημειώθηκαν χειροκίνητα. Η διάρκεια των κομματιών κυμαινόταν από 135 έως 391 s και ο ρυθμός δειγματοληψίας όλων των κομματιών κατέβηκε από τα 44100 Hz στα 22050 Hz.

DB-ALBUM1 (Li & Wang, 2005)**,** (Nwe & Li, 2007)**:** Η συγκεκριμένη βάση δεδομένων αποτελείται από 84 δημοφιλή σόλο κομμάτια Βρετανών και Κινέζων καλλιτεχνών. Καθένας εκ των 12 καλλιτεχνών της βάσης δεδομένων ερμηνεύει από 7 κομμάτια. Τα τραγούδια καλύπτουν μία χρονική περίοδο από το 1990 μέχρι και το 2004. Τα τραγούδια στο σύνολο εκπαίδευσης, τα οποία είχαν χωριστεί ως σημεία με φωνητικά και σημεία χωρίς, ταξινομήθηκαν χειροκίνητα ως εξής: εισαγωγή, κουπλέ, ρεφρέν, γέφυρα και κλείσιμο. Στα τραγούδια του συνόλου δοκιμής από την άλλη, επισημάνθηκαν χειροκίνητα τα τμήματα με φωνητικά και αυτά χωρίς. Έτσι, η βάση δεδομένων αποτελείται από τμήματα με φωνητικά συνολικής διάρκειας 200 λεπτών, στο σύνολο δοκιμής και 112 λεπτών στο σύνολο δοκιμής. Αυτά, είναι σε στερεοφωνική μορφή, με ρυθμό δειγματοληψίας 44100 Hz και 16 bit ανά δείγμα.

DB-ALBUM2 (Nwe et al., 2007): Αυτή η βάση δεδομένων αποτελείται από τρεις τραγουδιστές και επτά τραγούδια ανά τραγουδιστή, δηλαδή 21 συνολικά. Η συνολική διάρκεια των τμημάτων με φωνητικά είναι 30 λεπτά στο σύνολο εκπαίδευσης και 21 στο σύνολο δοκιμής.

(Li & Wang, Μάιος 2007): Αυτή η βάση δεδομένων περιλαμβάνει 10 κομμάτια από καραόκε CD, συχνότητας δειγματοληψίας 16000 Hz και ανάλυσης 16-bit. Τα δέκα εξ αυτών είναι ροκ και τα άλλα πέντε κάουντρι.

(Li & Wang, Μάιος 2007): 25 εξαγμένα τμήματα από την παραπάνω βάση δεδομένων, με μέση διάρκεια 3,9 s και συνολική 97,5 s, ερμηνευμένα και από άντρες αλλά και από γυναίκες. Σε κάποια εκ των δειγμάτων, η φωνή είναι παρούσα καθ' όλη τους τη διάρκεια, σε άλλα όμως βρίσκεται ή στην αρχή ή στη μέση ή στο τέλος τους.

(Σοφιανός, et al., 2010): Αυτή η βάση δεδομένων αποτελείται από 10 ηχητικά αποσπάσματα. Η διάρκεια των αποσπασμάτων κυμαίνεται από 4 έως και 23 s.

TrainDB – SingVD (Nwe & Li, 2008)**:** 35 κομμάτια τα οποία επισημαίνονται χειροκίνητα ως προς την ύπαρξη ή την ανυπαρξία φωνητικών. Τα κομμάτια που εμπεριέχονται στη συγκεκριμένη βάση δεδομένων δεν αλληλεπικαλύπτονται με τα κομμάτια των βάσεων δεδομένων DevelopmentDB – Sing VD, TestDB – SingVD, TrainDB – SingerID και TestDB – SingerID.

DevelopmentDB – SingVD (Nwe & Li, 2008)**:** 25 κομμάτια τα οποία επισημαίνονται χειροκίνητα ως προς την ύπαρξη ή την ανυπαρξία φωνητικών. Τα κομμάτια που εμπεριέχονται στη συγκεκριμένη βάση δεδομένων δεν αλληλεπικαλύπτονται με τα κομμάτια των βάσεων δεδομένων TrainDB – Sing VD, TestDB – SingVD, TrainDB – SingerID και TestDB – SingerID.

TestDB – SingVD (Nwe & Li, 2008)**:** 45 κομμάτια τα οποία επισημαίνονται χειροκίνητα ως προς την ύπαρξη ή την ανυπαρξία φωνητικών. Τα κομμάτια που εμπεριέχονται στη συγκεκριμένη βάση δεδομένων δεν αλληλεπικαλύπτονται με τα κομμάτια των βάσεων δεδομένων TrainDB – SingVD, DevelopmentDB – SingVD, TrainDB – SingerID και TestDB – SingerID.

TrainDB – SingerID (Nwe & Li, 2008)**:** 48 κομμάτια, από 12 σόλο τραγουδιστές από τη βάση δεδομένων DB-ALBUM1, τα οποία επισημαίνονται χειροκίνητα ως προς την ύπαρξη ή την ανυπαρξία φωνητικών. Τα κομμάτια που εμπεριέχονται στη συγκεκριμένη βάση δεδομένων δεν αλληλεπικαλύπτονται με τα κομμάτια των βάσεων δεδομένων TrainDB – SingVD, DevelopmentDB – SingVD, TestDB – SingVD, DevelopmentDB – SingerID, και TestDB - SingerID.

DevelopmentDB – SingerID (Nwe & Li, 2008)**:** 25 κομμάτια από 12 σόλο τραγουδιστές από τη βάση δεδομένων DB-ALBUM1, τα οποία επισημαίνονται χειροκίνητα ως προς την ύπαρξη ή την ανυπαρξία φωνητικών. Τα κομμάτια που εμπεριέχονται στη συγκεκριμένη βάση δεδομένων δεν αλληλεπικαλύπτονται με τα κομμάτια των βάσεων δεδομένων TrainDB – SingVD, DevelopmentDB – SingVD, TestDB – SingVD, TrainDB – SingerID και TestDB – SingerID.

TestDB – SingerID (Nwe & Li, 2008)**:** 36 κομμάτια από 12 σόλο τραγουδιστές από τη βάση δεδομένων DB-ALBUM1, τα οποία επισημαίνονται χειροκίνητα ως προς την ύπαρξη ή την ανυπαρξία φωνητικών. Τα κομμάτια που εμπεριέχονται στη συγκεκριμένη βάση δεδομένων δεν αλληλεπικαλύπτονται με τα κομμάτια των βάσεων δεδομένων TrainDB – Sing VD, DevelopmentDB – SingVD, TrainDB – SingerID και DevelopmentDB – SingerID.

(Sridhar & Geetha, 2008): Η βάση δεδομένων που χρησιμοποιήθηκε για την εκτέλεση των πειραμάτων του συγκεκριμένου συγγράμματος περιλάμβανε 600 κομμάτια καρνατικής μουσικής από 60 καλλιτέχνες, 10 από τον καθένα.

(Chanrungutai & Ratanamahatana, 2008): Στο συγκεκριμένο σύγγραμμα χρησιμοποιήθηκε μία βάση δεδομένων με 10 κομμάτια με δειγματοληψία 16.000 Hz και με βάθος bit 16, από

διάφορους τραγουδιστές και καθένα χωρίστηκε σε τρία μέρη. Δύο κομμάτια τραγουδίστηκαν πάνω σε MIDI backing tracks και τα υπόλοιπα σε αληθινά.

"artist20" (Shirali-Shahreza & Shirali-Shahreza, 2009)**:** Αυτή βάση δεδομένων βασίζεται κυρίως στη "uspop2002". Υπάρχουν 20 καλλιτέχνες σε αυτήν, με έξι άλμπουμ ανά καλλιτέχνη και 1413 κομμάτια συνολικά. Αυτή η βάση δεδομένων δημιουργήθηκα για την επαλήθευση εργασιών αναγνώρισης καλλιτέχνη και έχει ένα καθορισμένο κανονικοποιημένο σχήμα για διασταυρωμένη επικύρωση 6-fold.

"uspop2002" (Berenzweig, et al., 2004)**:** Αποτελείται από 706 άλμπουμ, 400 καλλιτεχνών, με συνολικά 8764 κομμάτια. Στην αρχή ήταν 8772 κομμάτια αλλά ένας δίσκος εμφανίστηκε δύο φορές με ελαφρώς διαφορετικά ονόματα.

The "asset400" Pop Music Artist set (Ellis, et al., 2003)**:** Στην αρχή εμπεριείχε 414 καλλιτέχνες, που μετέπειτα έγιναν 412 λόγω διπλότυπων και στη συνέχεια κατέληξαν να περιοριστούν στους 400.

CAL500 Expansion (CAL500exp) (Wang, et al., 2014)**:** 3.223 τμήματα των τριών με 16 δευτερολέπτων από 500 κομμάτια.

CAL500 (Turnbull, et al., 2008)**:** 500 κομμάτια, που χρησιμοποιείται για συμβατική ετικετοποίηση σε επίπεδο κομματιών.

CAL10k[1]**:** 10.870 κομμάτια.

MSD[2]**:** 1.000.000 κομμάτια.

MajorMiner[3]**:** 2.600 τμήματα των 10 δευτερολέπτων.

Magnatagatune[4]**:** 25.860 τμήματα των 30 δευτερολέπτων.

(Turk, 1991): 925 τμήματα των 10 δευτερολέπτων.

(Chang, 2009): Για τη βάση δεδομένων του συγκεκριμένου συγγράμματος επιλέχθηκαν τυχαία 20 μουσικά κομμάτια με φωνή από την CAL500.

RWC Music Database[5]**:** Η RWC Music Database περιέχει 215 κομμάτια συνολικά και αποτελείται από την Popular Music Database, από τη Royalty-Free Music Database, από την Classical Music Database και από την Jazz Music Database.

[1] http://calab1.ucsd.edu/~datasets/cal10k/
[2] http://millionsongdataset.com/
[3] http://majorminer.org/search/human
[4] http://mirg.city.ac.uk/codeapps/the-magnatagatune-dataset
[5] https://staff.aist.go.jp/m.goto/RWC-MDB/

The Popular Music Database[6]: Η συγκεκριμένη βάση δεδομένων αποτελεί μέρος της μεγαλύτερης RWC Music Database και περιλαμβάνει 100 κομμάτια πρωτότυπα ηχογραφημένων συμπαγών μουσικών δίσκων, πρότυπα MIDI αρχεία και αρχεία κειμένου με στίχους. Τα 20 κομμάτια εκ των 100 της βάσης δεδομένων έχουν αγγλικούς στίχους με την τεχνοτροπία εκείνη των ποπ κομματιών που βρίσκονταν στα αμερικάνικα chart το 1980. Τα υπόλοιπα 80 έχουν ιαπωνικούς στίχους ποπ μουσικής των ιαπωνικών chart του 1990. Τα 50 τραγούδια από τα 100 της βάσης δεδομένων τα ερμηνεύουν 15 άντρες, τα 44, 13 γυναίκες και τα υπόλοιπα 6, 6 συγκροτήματα φωνητικών. Όλα τα κομμάτια είναι αυθεντικά ποπ κομμάτια.
Royalty-Free Music Database[7]: Η Royalty-Free Music Database αποτελεί μέρος της RWC Music Database και περιλαμβάνει 15 κομμάτια πρωτότυπα ηχογραφημένων συμπαγών μουσικών δίσκων, πρότυπα MIDI αρχεία και αρχεία κειμένου με στίχους. Τα 10 είναι πολύ δημοφιλή πρότυπα ποπ τραγούδια με αγγλικούς στίχους και τα υπόλοιπα 5 είναι δημοφιλή παιδικά τραγούδια με ιαπωνικούς στίχους. Όλα τα κομμάτια είναι πρωτότυπης σύνθεσης και ηχογράφησης και ηχογραφήθηκαν από 16 ανθρώπους συμπεριλαμβανομένων δύο συνθετών και τριών τραγουδιστών.
Classical Music Database[8]: Η Classical Music Database αποτελεί μέρος της RWC Music Database Και περιέχει 50 κομμάτια πρωτότυπα ηχογραφημένων συμπαγών μουσικών δίσκων, πρότυπα MIDI αρχεία και αρχεία κειμένου με στίχους και περιλαμβάνει 4 συμφωνίες, 2 κονσέρτα, 4 ορχηστρικά, 10 δωματίου, 24 σόλο, και 6 φωνητικών. Όλα είναι αυθεντικά ηχογραφημένα, αλλά όχι όλες οι μεταβάσεις. Μία συγκεκριμένη επιλέχθηκε και ηχογραφήθηκε για πολλές κατηγορίες όπως στις συμφωνίες και στα κονσέρτα. Επιλέχθηκαν προκειμένου να αναπαραστήσουν μία μεγάλη ποικιλία ενορχηστρώσεων, τεχνοτροπιών, περιόδων, συνθετών και διαθέσεων. Δεν είναι απλά μία ανθολογία από πολύ γνωστά μουσικά κομμάτια, αλλά εντάχθηκαν κομμάτια που χρησιμοποιήθηκαν σε προηγούμενες έρευνες ή παρουσιάζουν ενδιαφέρον από ερευνητική άποψη. Αυτά ηχογραφήθηκαν από 115 ανθρώπους, συμπεριλαμβανομένων μίας φιλαρμονικής ορχήστρας με 72 οργανοπαίκτες και 1 μαέστρο, 16 πιανιστών και τεσσάρων βιολιστών.

[6] https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-p.html
[7] https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-r.html
[8] https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-c.html

Jazz Music Database[9]**:** Η Jazz Music Database αποτελεί μέρος της RWC Music Database και εμπεριέχει 50 κομμάτια πρωτότυπα ηχογραφημένων συμπαγών μουσικών δίσκων, πρότυπα MIDI αρχεία και αρχεία κειμένου με στίχους. Η συγκεκριμένη βάση δεδομένων αποτελείται από 35 ενορχηστρωτικές παραλλαγές (5 κομμάτια που το καθένα έχει 7 διαφορετικές ενορχηστρώσεις), από 9 διαφορετικές παραλλαγές τεχνοτροπιών και από 6 fusion (crossover). Όλα τα κομμάτια δημιουργήθηκαν αποκλειστικά για τη βάση δεδομένων, εκτός από τη σύνθεση και από το γράψιμο των στίχων για τέσσερα κομμάτια εναλλασόμενων τεχνοτροπιών. Πρώτα τα κομμάτια με τις παραλλαγμένες ενορχηστρώσεις ηχογραφήθηκαν προκειμένου να δημιουργήσουν διαφορετικές συνθέσεις από το ίδιο κομμάτι. 5 πρότυπα κομμάτια jazz δημιουργήθηκαν και εκτελέστηκαν με μοντέρνη τεχνοτροπία jazz χρησιμοποιώντας επτά ενορχηστρώσεις, μία με σόλο πιάνο, μία με σόλο κιθάρα, μία ντούο με βιμπράφωνο και πιάνο, με φλάουτο και πιάνο, ή με πιάνο και μπάσο, ένα τρίο με πιάνα και ένα τρίο με πιάνα και τρομπέτα ή τενόρο σαξόφωνο, μία οκταφωνία με τρίο πιάνων, κιθάρα, άλτο σαξόφωνο, βαρύτονο σαξόφωνο και δύο τενόρα σαξόφωνα και μία με τρίο πιάνων, βιμπράφωνο ή φλάουτο. Έπειτα, οι διάφορες παραλλαγές ηχογραφήθηκαν έτσι ώστε να αναπαριστούν διάφορα είδη jazz. Τα εννιά κομμάτια που περιλαμβάνουν τέσσερα πολύ γνωστά κομμάτια, καλύπτουν τα είδη vocal jazz (δύο κομμάτια), big band jazz (δύο κομμάτια), modal jazz (δύο κομμάτια), funky jazz (δύο κομμάτια) και free jazz (ένα κομμάτι). Τέλος, τα fusion κομμάτια ηχογραφήθηκαν και συνδυάζουν στοιχεία jazz με άλλες τεχνοτροπίες όπως pop, rock ή latin. Όλα τα κομμάτια ηχογραφήθηκαν από 53 ανθρώπους, συμπεριλαμβανομένων τεσσάρων συνθετών και ενός στιχουργού.

(Fujihara & Goto, 2010): Για την εκτέλεση των πειραμάτων που παρουσιάζονται στο συγκεκριμένο σύγγραμμα χρησιμοποιήθηκαν 40 κομμάτια για εκπαίδευση και για δοκιμή, από δέκα διαφορετικούς τραγουδιστές, οι πέντε εκ των οποίων ήταν άντρες και οι άλλοι πέντε γυναίκες. Τα κομμάτια αυτά περισυλλέχθηκαν από τη βάση δεδομένων RWC Music Database: Popular Music (RWC-MDB-P-2001). Επίσης, χρησιμοποιήθηκαν 25 κομμάτια από 16 διαφορετικούς τραγουδιστές (8 άντρες και 8 γυναίκες), ως δεδομένα εκπαίδευσης της επιλογής καρέ, επίσης από τη βάση δεδομένων RWC-MDB-P-2001. Οι τραγουδιστές διέφεραν από αυτούς που χρησιμοποιήθηκαν για επαλήθευση.

[9] https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-j.html

Επίσης, διεξάχθηκαν πειράματα με χρήση ηχογραφήσεων από CD τα οποία ήταν διαθέσιμα στην Ιαπωνία. Χρησιμοποιήθηκαν 246 κομμάτια από 20 καλλιτέχνες, 8 εκ των οποίων ήταν άνδρες και 12 γυναίκες. Οι καλλιτέχνες επιλέχθηκαν από τη λίστα με τα CD που πούλησαν περισσότερα αντίτυπα στην Ιαπωνία.

DB-KAR-ACC-VOC: Η συγκεκριμένη βάση δεδομένων εμπεριέχει το ένα από τα δύο κανάλια από 308 κομμάτια μανδαρινικής μουσικής εξαγμένα από καραόκε VCD. Το κάθε κανάλι εμπεριείχε μία μίξη φωνητικών και συνοδείας μουσικής. Όλα τα δεδομένα μετατράπηκαν σε μορφή PCM wave με ρυθμό δειγματοληψίας 22.05 kHz και επίπεδο κβαντοποίησης 16-bit από 44.1 kHz. Η τυπική διάρκεια κάθε κομματιού ήταν περίπου τρία λεπτά. Δημιουργήθηκε για τις ανάγκες των πειραμάτων που περιγράφονται στο σύγγραμμα (Tsai & Lin, Ιούλιος 2010).

DB-KAR-ACC: Αυτή η βάση δεδομένων εμπεριέχει το δεύτερο κανάλι από τα κομμάτια που χρησιμοποιήθηκαν στην DB-KAR-ACC-VOC. Το κανάλι αυτό εμπεριέχει μόνο συνοδεία μουσικής. Όλα τα δεδομένα μετατράπηκαν σε μορφή PCM wave με ρυθμό δειγματοληψίας 22.05 kHz και επίπεδο κβαντοποίησης 16-bit από 44.1 kHz. Η τυπική διάρκεια κάθε κομματιού ήταν περίπου τρία λεπτά. Αυτή η βάση δεδομένων δημιουργήθηκε επίσης για τις ανάγκες του συγγράματος (Tsai & Lin, Ιούλιος 2010).

DB-SOL: Αυτή η βάση δεδομένων περιέχει ηχογραφήσεις σόλο φωνητικών από δέκα ερασιτέχνες άντρες τραγουδιστές και δέκα ερασιτέχνες γυναίκες τραγουδίστριες. Οι ηλικίες τους κυμαίνονταν από τα 20 έως τα 35 έτη και οι τραγουδιστικές τους ικανότητες προσέγγιζαν αυτές των επαγγελματιών. Καθένας ζητήθηκε να ερμηνεύσει οποιαδήποτε 10 τραγούδια της επιλογής του μεταξύ των 308 μανδαρινικών τραγουδιών από τις βάσεις δεδομένων DB-KAR-ACC-VOC και DB-KAR-ACC. Έτσι, η συγκεκριμένη βάση δεδομένων εμπεριέχει συνολικά 200 σόλο ηχογραφήσεις. Όλα τα δεδομένα ηχογραφήθηκαν σε μορφή PCM wave με ρυθμό δειγματοληψίας 22.05 kHz και επίπεδο κβαντοποίησης 16-bit. Η τυπική διάρκεια κάθε ηχογράφησης ήταν περίπου τρία λεπτά. Δημιουργήθηκε επίσης για τις ανάγκες του συγγράμματος (Tsai & Lin, Ιούλιος 2010).

DB-MIX-ACC-VOC: Η μίξη των σόλο φωνητικών της βάσης δεδομένων DB-SOL με τις συνοδείες μουσικής της DB-KAR-ACC. Οι περιοχές που δεν εμπεριείχαν φωνητικά στις σόλο ηχογραφήσεις και στις ηχογραφήσεις με μίξη φωνητικών και μουσικής αφαιρέθηκαν χειροκίνητα. Αυτή η βάση δεδομένων δημιουργήθηκε για τις ανάγκες του συγγράματος (Tsai & Lin, Ιούλιος 2010).

DB-MAN: Η DB-MAN είναι μία συλλογή από τυχαία επιλεγμένα τραγούδια από δημοφιλή μουσικά CD μανδαρινικής μουσικής. Περιλαμβάνει 300 κομμάτια εκτελεσμένα από 20 γυναίκες τραγουδίστριες και 20 άντρες τραγουδιστές. Κάθε τραγουδιστής και τραγουδίστρια ερμήνευσε 10 κομμάτια. Όλα τα κομμάτια μετατράπηκαν από ρυθμό δειγματοληψίας 44.1 kHz σε 22.05 kHz και τα σημεία με φωνητικά και χωρίς επισημάνθηκαν χειροκίνητα. Χρησιμοποιήθηκε για τις ανάγκες του συγγράμματος (Tsai & Lin, Ιούλιος 2010).

DB-ENG: Η DB-ENG είναι μία συλλογή από τυχαία επιλεγμένα κομμάτια από δημοφιλή αγγλικά μουσικά CD. Εμπεριέχει 200 κομμάτια ερμηνευμένα από δέκα άντρες τραγουδιστές και δέκα γυναίκες τραγουδίστριες. Κάθε τραγουδιστής και τραγουδίστρια ερμήνευσε 10 κομμάτια. Όλα τα κομμάτια μετατράπηκαν από ρυθμό δειγματοληψίας 44.1 kHz σε 22.05 kHz και τα σημεία με φωνητικά και χωρίς επισημάνθηκαν χειροκίνητα. Χρησιμοποιήθηκε για τις ανάγκες του συγγράμματος (Tsai & Lin, Ιούλιος 2010).

DB-JAP-ENG: Η DB-JAP-ENG δημιουργήθηκε από κομμάτια που εξάχθηκαν από τη μουσική βάση δεδομένων RWC: Popular (RWC-MDB-P-2001). Περιλαμβάνει 28 ιαπωνικά κομμάτια και 12 αγγλικά ερμηνευμένα από πέντε γυναίκες ερμηνεύτριες και πέντε άντρες ερμηνευτές. Κάθε ερμηνευτής και ερμηνεύτρια τραγούδησε τέσσερα κομμάτια. Όλα τα κομμάτια μετατράπηκαν από ρυθμό δειγματοληψίας 44.1 kHz σε 22.05 kHz και τα σημεία με φωνητικά και χωρίς επισημάνθηκαν χειροκίνητα. Χρησιμοποιήθηκε για τις ανάγκες του συγγράμματος (Tsai & Lin, Ιούλιος 2010).

(Kopparapu & Laxminarayana, 2010): Η βάση δεδομένων που χρησιμοποιήθηκε για τους σκοπούς του συγκεκριμένου συγγράμματος περιλαμβάνει πολλά ηχητικά σήματα τα οποία είχαν τμήματα με ομιλία (ή φωνή γενικότερα) και μουσική. Ο ευδιάκριτος διαχωρισμός φωνής και μουσικής αποτελεί ένα βασικό στοιχείο της ινδικής κλασικής μουσικής και για αυτό το λόγο χρησιμοποιήθηκε ένα μεγάλο σύνολο ινδικής κλασικής μουσικής σε μορφή WAV. Η συνολική διάρκεια της βάσης δεδομένων είναι 275 s. Η δειγματοληψία των κομματιών της βιβλιοθήκης ήταν τα 22.05 kHz. Τα σημεία με φωνή επισημάνθηκαν χειροκίνητα με το γράμμα V και τα σημεία με μουσικής επισημάνθηκαν με το γράμμα M, με τη χρήση μίας ημιαυτόματης διαδικασίας. Τέλος, ελέγχθηκαν χειροκίνητα προκειμένου να γίνει σωστά η κατηγοριοποίηση. Η βάση δεδομένων αποτελούνταν από 250 τμήματα, τα 150 εκ των οποίων ήταν τμήματα με φωνή και τα 100 με μουσική και κάθε τμήμα είχε μέση διάρκεια 2 s.

(Ezzaidi, et al., 2010): Οι Ezzaidi, Bahoura και Rouat χρησιμοποίησαν 68 κομμάτια από τη βάση δεδομένων μουσικών ειδών RWC. Τα κομμάτια είχαν διάρκεια περίπου τέσσερα λεπτά

και άνηκαν σε 25 μουσικά είδη. Οι ερμηνευτές των κομματιών ήταν και άντρες και γυναίκες και τα κομμάτια που χρησιμοποιήθηκαν βρίσκονται κατηγοριοποιημένα στη βάση δεδομένων RWC και είναι όλα τα κομμάτια που ξεκινούν από G01_, G02_, G03_, G04_, G05_ και G07_. Η διάρκεια ολόκληρης της βάσης δεδομένων είναι περίπου πέντε ώρες, οι τρεις ώρες εκ των οποίων εμπεριείχαν μόνο μουσική.

(Maazouri & Bahi, n.d.): Η βάση δεδομένων που χρησιμοποιήθηκε από τους Maazouri και Bahi περιλαμβάνει αλγερινά τραγούδια από 50 τραγουδιστές, αλιευμένα από την ιστοσελίδα Zikdalgerie.com[10].

(Cai, et al., 2011): Η βάση δεδομένων που χρησιμοποιήθηκε στο συγκεκριμένο σύγγραμμα αποτελείται από δέκα τραγουδιστές και τραγουδίστριες από την Κίνα, οι πέντε εκ των οποίων ήταν άντρες και οι πέντε γυναίκες. Όλα τα κομμάτια είχαν ρυθμό δειγματοληψίας 22050 Hz και μετατράπηκαν σε μορφή μονοφωνικού WAV.

DB-Singing: Αυτή η βάση δεδομένων περιλαμβάνει 600 αποσπάσματα μανδαρινικής pop μουσικής ερμηνευμένα από 20 ερασιτέχνες άντρες τραγουδιστές, με ηλικίες μεταξύ 20 και 39 ετών. Κάθε τραγουδιστής ερμήνευσε από 30 αποσπάσματα και όλα τα αποσπάσματα ηχογραφήθηκαν με τη χρήση μία μηχανής καραόκε σε σιωπηλό δωμάτιο, σε δειγματοληψία 22.05 kHz, με βάθος bit 16 bit και σε μορφή μονοφωνικού PCM wave. Η συνοδευτική μουσική του καραόκε εξάχθηκε σε ακουστικά, έτσι ώστε να μην καταγραφεί στις ηχογραφήσεις. Έτσι δημιουργήθηκε μία βάση δεδομένων με ηχητικά αποσπάσματα που εμπεριείχαν μόνο φωνητικά. Η διάρκεια κάθε αποσπάσματος κυμαινόταν από 17 μέχρι και 26 δευτερόλεπτα. Η συγκεκριμένη βάση δεδομένων δημιουργήθηκε για τις ανάγκες του συγγράμματος (Tsai & Lee, 2012).

DB-Speech: Για τη δημιουργία της βάσης δεδομένων DB-Speech, οι ερμηνευτές των αποσπασμάτων της DB-Singing, απήγγειλαν τους στίχους των αποσπασμάτων που ερμήνευσαν για τη δημιουργία της DB-Singing, σε κανονική ταχύτητα. Οι διαδικασία που ακολουθήθηκε για την ηχογράφηση των αποσπασμάτων ήταν παρόμοια με εκείνη που ακολουθήθηκε για τη δημιουργία της βάσης δεδομένων DB-Singing. Τα ηχητικά αποσπάσματα τα οποία περιλαμβάνονταν στη συγκεκριμένη βάση δεδομένων ήταν 600, δηλαδή 30 ανά καλλιτέχνη. Η συγκεκριμένη βάση δεδομένων δημιουργήθηκε για τις ανάγκες του συγγράμματος (Tsai & Lee, 2012).

[10] https://www.zikdalgerie.com/

(Regnier & Peeters, 2011): Η βάση δεδομένων που χρησιμοποιήθηκε για τις ανάγκες του συγκεκριμένου συγγράμματος αποτελούνταν από 54 κομμάτια. Τα κομμάτια αυτά ερμηνεύθηκαν από 18 καλλιτέχνες, καθένας από τους οποίους ερμήνευσε το μέρος των κύριων φωνητικών από 3 κομμάτια. Από κάθε κομμάτι επιλέχθηκαν 50 νότες, κατά προσέγγιση. Τελικώς, η βάση δεδομένων αποτελούνταν από 2592 νότες. Τα δύο κομμάτια εκ των τριών που ερμήνευσε ο κάθε καλλιτέχνης χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου, ενώ το εναπομείναν χρησιμοποιήθηκε ως τραγούδι προς ερεύνηση.

(Patil, et al., n.d.): Η βάση δεδομένων που χρησιμοποιήθηκε για τις ανάγκες του συγκεκριμένου συγγράμματος περιλάμβανε 500 κομμάτια από 20 διαφορετικούς δημοφιλείς καλλιτέχνες, 14 άντρες και 6 γυναίκες. Ο κάθε καλλιτέχνης ερμήνευσε 25 διαφορετικά Bollywood Hindi κομμάτια, τα οποία περισυλλέχθηκαν κυρίως από ταινίες και άλμπουμ και ήταν διαθέσιμα στο κοινό σε μορφή CD ή DVD. Ο κάθε καλλιτέχνης ήταν μέλος της βιομηχανίας του Bollywood για πάνω από 20 με 25 χρόνια. Για να διατηρηθεί η ομοιομορφία της βάσης δεδομένων, η συχνότητα δειγματοληψίας όλων των κομματιών κατέληξε στα 22.050 Hz και το βάθος bit τους στα 16 bit, σε PCM μορφή. Επίσης, όλα μετατράπηκαν από στερεοφωνικά σε μονοφωνικά.

Από τη βάση δεδομένων αφαιρέθηκαν χειροκίνητα, εξ ακοής, τα ορχηστρικά τμήματα, τα τμήματα όπου τα φωνητικά επαναλαμβάνονταν, δηλαδή τμήματα που έχουν ήδη επιλεχθεί και τα τμήματα που ήταν ερμηνευμένα από άλλον τραγουδιστή (εάν το κομμάτι ερμηνεύονταν από παραπάνω από έναν τραγουδιστή). Έτσι, το κάθε επεξεργασμένο κομμάτι εμπεριέχει σόλο φωνητικά, ένα τμήμα με θόρυβο τύπου humming (αν είναι υπαρκτός στο εκάστοτε κομμάτι) και κυρίως διαφορετικά κουπλέ από κάθε τραγουδιστή. Το καθένα από τα προκύπτοντα κομμάτια, δηλαδή αυτά που περισυλλέχθηκαν μετά από την απομάκρυνση κάθε μη απαραίτητου τμήματος, είχε διαφορετική χρονική διάρκεια. Ωστόσο, η ελάχιστη χρονική διάρκεια κάθε κομματιού ήταν 60 s και οι διάρκειές τους κυμαίνονταν από 60 s μέχρι και 166 s, με μέση διάρκεια τα 102 s. Μερικά κομμάτια μπορεί να εμπεριέχουν λίγα ορχηστρικά μέρη, αλλά η μέγιστη διάρκειά τους είναι μόλις 500 ms. Ο σκοπός της δημιουργίας μίας τέτοιας βάσης δεδομένων είναι ότι μπορεί να χρησιμοποιηθεί ως δεδομένα απόλυτης αλήθειας όταν η διαδικασία αυτόματου διαχωρισμού τμημάτων με φωνητικά και χωρίς, εκτελείται σε ένα κομμάτι.

MIR-1k[11]: Η συγκεκριμένη βάση δεδομένων χρησιμοποιήθηκε για τις ανάγκες του συγγράμματος (Jeong & Lee, 2014) και αποτελείται από 1.000 μουσικά αποσπάσματα ερμηνευμένα από ερασιτέχνες τραγουδιστές. Χρησιμοποιήθηκε για την ποσοτική επαλήθευση του προτεινόμενου αλγορίθμου διαχωρισμού φωνητικών-μουσικής από τους Jeong και Lee, καθώς και για την επαλήθευση και άλλων αλγορίθμων που χρησιμοποιήθηκαν για τη σύγκριση με το δικό τους. Τα κανάλια των φωνητικών ηχογραφήθηκαν ξεχωριστά από τη μουσική συνοδεία και μιξαρίστηκαν με τρόπο ώστε να έχουν -5 dB, 0 dB και 5 dB λόγο φωνητικών-συνοδείας VAR.

Ο ρυθμός δειγματοληψίας ήταν 16 kHz και το παράθυρο ανάλυσης 1.024 δείγματα με λόγο αλληλεπικάλυψης ¾. Οι παράμετροι α και φ ρυθμίστηκαν στο 0.25 και 0.25 E, αντίστοιχα. Το πλήθος των επαναλήψεων ήταν 200. Για να γίνει η σύγκριση πιο δίκαιη, τα αποτελέσματα φιλτραρίστηκαν με τη χρήση ενός υψιπερατού φίλτρου με 110 Hz συχνότητα αποκοπής, η οποία είναι η ίδια που χρησιμοποιήθηκε στον αλγόριθμο Tachibana. Ωστόσο, τα καλύτερα δυνατά αποτελέσματα πάρθηκαν με τη συχνότητα αποκοπής στα 120 Hz.

Βάση δεδομένων Beach Boys: Αυτή η βάση δεδομένων εμπεριέχει πραγματικά μουσικά κομμάτια με μεγαλύτερη χρονική διάρκεια από αυτή των μουσικών κομματιών της βάσης δεδομένων MIR-1k. Για τη δημιουργία της βάσης δεδομένων επιλέχθηκε το άλμπουμ Good Vibrations: Years of the Beach Boys, επειδή εμπεριέχει πολλά μουσικά κομμάτια, τα φωνητικά των οποίων ηχογραφήθηκαν στο ένα κανάλι και η μουσική συνοδεία στο άλλο και ήταν πολύ δύσκολο να βρεθούν αυθεντικές πολυκάναλες ηχογραφήσεις. Παρόλο που ήταν περιοριστικό που όλες οι ηχογραφήσεις προέρχονταν από τον ίδιο καλλιτέχνη και από το ίδιο είδος μουσικής, η συγκεκριμένη βάση δεδομέων θεωρήθηκε εξαιρετικά χρήσιμη σε πολλά πειράματα, για την επαλήθευση αλγορίθμων φωνητικού διαχωρισμού.

Οι παράμετροι ήταν ίδιες με αυτές που χρησιμοποιήθηκαν στη βάση δεδομέων MIR-1k, με μοναδική διαφοροποίηση τη συχνότητα αποκοπής που ορίστηκε στα 100 Hz για δίκαιη σύγκριση. Η βάση δεδομένων Beach Boys δημιουργήθηκε για τις ανάγκες του συγγράμματος (Jeong & Lee, 2014).

Minnowmatch (Whitman, et al., 2001)**:** Η βάση δεδομένων Minnowmatch αποτελείται από περισσότερα από 250 κομμάτια από άλμπουμ περισσότερων από 20 καλλιτεχνών και

[11] https://sites.google.com/site/unvoicedsoundseparation/mir-1k

συγκροτημάων. Ορισμένα από τα κομμάτια που εμπεριέχονται στη συγκεκριμένη βάση δεδομένων δεν εμπεριέχουν φωνητικά.

(Kim & Whitman, 2002): Για τους σκοπούς του συγκεκριμένου συγγράμματος χρησιμοποιήθηκαν διάφορα υποσέτ από το σετ δοκιμής NECI Minnowmatch με κάποιες ελάχιστες προσθήκες. Τα κομμάτια τα οποία δεν περιλάμβαναν φωνητικά δε χρησιμοποιήθηκαν στα πειράματα. Σε ορισμένες περιπτώσεις, διαφορετικοί τραγουδιστές ερμηνεύουν διαφορετικά τραγούδια σε κάθε συγκρότημα. Για αυτό το λόγο, κάθε κομμάτι ταξινομήθηκε ως προς τον τραγουδιστή που ερμήνευε το εκάστοτε τραγούδι και το τελικό σετ δοκιμής συμπεριλάμβανε 17 διαφορετικούς σόλο τραγουδιστές και λίγα περισσότερα από 200 κομμάτια. Ο ρυθμός δειγματοληψίας όλων των κομματιών μειώθηκε από τον πρότυπο ρυθμό δειγματοληψίας συμπαγών δίσκων 44,1 kHz στα 11,025 kHz. Αυτό έγινε προκειμένου να μειωθεί ο όγκος των δεδομένων προς αποθήκευση και η επεξεργαστική ισχύ που απαιτούνταν. Ακόμα και μετά από αυτή τη δραστική μείωση του ρυθμού δειγματοληψίας, ο κύριος όγκος της ενέργειας των φωνητικών βρίσκονταν κάτω από την τιμή του ρυθμού Nyquist, που είναι η μισή της συχνότητας δειγματοληψίας.

Για να ελεγχθεί η ακρίβεια του ανιχνευτή τμημάτων με φωνητικά, ένα υποσέτ με 20 τραγούδια, δηλαδή σχεδόν το 10% της βάσης δεδομένων, διαμερίστηκαν χειροκίνητα εξ ακοής σε τμήματα με φωνητικά και τμήματα χωρίς προκειμένου να δημιουργηθεί ένα σετ με δεδομένα αληθείας. Για την αναγνώριση τραγουδιστή χρησιμοποιήθηκε περίπου η μισή βάση δεδομένων, συγκεκριμένα τα κομμάτια των άλμπουμ με περιττό αριθμό. Για την επαλήθευση της απόδοσης του ταξινομητή χρησιμοποιήθηκαν όσα κομμάτια δε χρησιμοποιήθηκαν για την αναγνώριση.

(Mesaros, et al., 2007): Η βάση δεδομένων που χρησιμοποιήθηκε στο συγκεκριμένο σύγγραμμα αποτελούνταν από 13 τραγουδιστές και των δύο φύλων με ποικίλλες τραγουδιστικές ικανότητες. Ηχογραφήθηκαν 4-6 μελωδίες με χρονικές διάρκειες από 20 μέχρι και 30 δευτερόλεπτα και συχνότητα δειγματοληψίας 44100 Hz και ανάλυση 16 bit, ανά τραγουδιστή. Σε κάθε τραγουδιστή δόθηκε η ίδια συνοδεία μουσικής και έτσι εξασφαλίστηκε ότι οι διαδικασίες που ακολουθήθηκαν δε διαμορφώνονταν ανάλογα με τον εκάστοτε τραγουδιστή. Το σετ εκπαίδευσης περιλάμβανε όλα τα δεδομένα του εκάστοτε τραγουδιστή εκτός από το κομμάτι που βρισκόταν υπό εξέταση.

DB1: Η DB1 περιλαμβάνει 9 τραγουδιστές με 8 δείγματα ανά τραγουδιστή, δηλαδή 72 αρχεία. Η ηχογραφήσεις πραγματοποιήθηκαν σε λίγο θορυβώδες περιβάλλον, με χρήση ενός φορητού

συστήματος ηχογράφησης, από τραγουδιστές κλασικής μουσικής της Βόρειας Ινδίας. Οι λόγοι που χρησιμοποιήθηκαν θορυβώδη ηχητικά αποσπάσματα ήταν δύο. Ο πρώτος λόγος ήταν ότι βάσει του συγγράμματος Schouten, 1968, η χροιά έχει πέντε κύρια χαρακτηριστικά, ένα εκ των οποίη ήταν το ότι η χροιά είναι ένα εύρος μεταξύ τόνου και θορύβου.

Έτσι θεωρητικά, εάν ο εξωτερικός θόρυβος προστεθεί σε αυτή τη βάση δεδομένων, τότε όλος ο θόρυβος συμπεριλαμβανομένου και κάποιου μέρους της χροιάς θα φιλτραριστεί. Έτσι, το σύστημα δε θα αναγνωρίσει τον τραγουδιστή. Ο δεύτερος λόγος ήταν ότι ο στόχος ήταν η επίτευξη της γενίκευσης της απόδοσης του συστήματος και η δημιουργία ενός στιβαρού συστήματος, ανεπηρέαστου από φυσιολογικό θόρυβο που προέρχεται από το σύστημα εισόδου των ηχογραφήσεων. Ολόκληρη η βάση δεδομένων φιλτραρίστηκε με τη χρήση μίας τεχνικής ανάστροφου φίλτρου χτένας και έπειτα ορίστηκε ως ρυθμός δειγματοληψίας τα 11.025 Hz με τη μέθοδο διαμόρφωσης παλμικού κώδικα PCM και σε μορφή ασυμπίεστου αρχείου με ανάλυση 16 bit. Κάθε κομμάτι είχε διάρκεια 5 s και ήταν μονοφωνικό. Δημιουργήθηκε για τις ανάγκες του συγγράμματος (Deshmukh & Bhirud, 2014).

Κεφάλαιο 4ο: Υλοποίηση συστήματος αυτόματης αναγνώρισης καλλιτέχνη

Για την υλοποίηση του συστήματος εκτελέστηκαν τα εξής βήματα. Πρώτα, δημιουργήθηκε η βάση δεδομένων που χρησιμοποιήθηκε για την εξαγωγή των χαρακτηριστικών και την κατηγοριοποίηση. Η διαδικασία της δημιουργίας της περιλάμβανε την επιλογή των καλλιτεχνών, οι οποίοι αποτελούν τις κλάσεις του συστήματος, την επιλογή των κομματιών και την περισυλλογή τους, τη χειροκίνητη επισήμανση των τμημάτων που εμπεριέχουν μίξη φωνής και μουσικής και αποκλειστικά φωνή ή μουσική. Τα υπόλοιπα απορρίφθηκαν. Έπειτα διαχωρίστηκαν από τα αρχικά κομμάτια και κανονικοποιήθηκε η βάση μέσω της περαιτέρω διχοτόμησης και μετατροπής των εξαγμένων ηχητικών αποσπασμάτων σε αποσπάσματα συγκεκριμένης χρονικής διάρκειας. Στη συνέχεια, απορρίφθηκαν όλα όσα κατέληξαν να είναι μικρότερης διάρκειας από την επιθυμητή. Επόμενη δουλειά μας ήταν να δημιουργήσουμε βάσεις δεδομένων ανάλογα με το ηχητικό περιεχόμενο των ηχητικών τμημάτων που απέμειναν. Έτσι, δημιουργήσαμε μία βάση δεδομένων η οποία εμπεριείχε δεδομένα με φωνή ως αποκλειστικό ηχητικό περιεχόμενο, μία με μουσική και μία με μίξη μουσικής και φωνής και μία η οποία εμπεριείχε το περιεχόμενο και των τριών προηγούμενων βάσεων. Τέλος, προσπαθήσαμε οι κλάσεις τους να εμπεριέχουν ίσο αριθμό αποσπασμάτων.

Για την εξαγωγή των χαρακτηριστικών και την κατηγοριοποίηση επιλέξαμε τις βάσεις που εμπεριείχαν μόνο μουσική και μουσική με φωνή. Για να επιλέξουμε το αποδοτικότερο σύνολο χαρακτηριστικών εκτελέσαμε μία σειρά πειραμάτων. Έτσι, αφού καταλήξαμε σε ένα συγκεκριμένο σύνολο χαρακτηριστικών πραγματοποιήσαμε μία σειρά κατηγοριοποιήσεων με χρήση των δύο αυτών βάσεων με χρήση των αλγορίθμων SVM, MLP και C4.5. Έπειτα, επαναλάβαμε τα ίδια πειράματα αφού πραγματοποιήσαμε επιλογή χαρακτηριστικών. Τέλος, επιχειρήσαμε την επαλήθευση του βέλτιστου μοντέλου μέσω της κατηγοριοποίησης ενός ολόκληρου κομματιού ανά καλλιτέχνη, κανένα εκ των οποίων δεν υπήρχε στη βάση δεδομένων.

4.1 Συλλογή μουσικών δεδομένων

Στην αρχή δοκιμάστηκε η χρήση κάποιων εκ των ήδη υπαρχόντων βάσεων, ωστόσο ή εμπεριείχαν κομμάτια ασιατών καλλιτεχνών άγνωστων για εμάς ή καταφέραμε να τις εντοπίσουμε με ήδη αναλυμένα τα χαρακτηριστικά τους και όχι σε μορφή αρχείων ήχου και πολλές από τις ήδη υπάρχουσες δεν καταφέραμε να τις εντοπίσουμε κάπου διαθέσιμες προς

μεταφόρτωση. Ωστόσο, ένας ακόμα λόγος που δημιουργήθηκε μία καινούργια βάση δεδομένων για τις ανάγκες των πειραμάτων της συγκεκριμένης πτυχιακης εργασίας είναι και για να διερευνηθεί η διαφοροποίηση του ηχητικού περιεχομένου των καλλιτεχνών στο πέρασμα του χρόνου. Έτσι επιλέχθηκαν κομμάτια καλλιτεχνών που κυκλοφόρησαν σε διαφορετικές χρονολογίες και καλλιτέχνες που ήταν κ0αι είναι ενεργοί μουσικά, πολλά έτη. Επίσης, έγινε προσπάθεια να επιλεχθούν συγκροτήματα των οποίων ο τραγουδιστής δεν άλλαζε συχνά ανά τα έτη.

Τα ηχητικά αποσπάσματα των βάσεων δεδομένων μας δημιουργήθηκαν χειροκίνητα, μέσω της ακρόασης των αρχικών κομματιών και της επισήμανσης τμημάτων τους ως “Music_Vocals”, “Music”, “Vocals” και “Null”. Έτσι δημιουργήθηκε η βάση δεδομένων All και από αυτή δημιουργήθηκαν τα υποσύνολα: Music, Vocals και Music_Vocals, που εμπεριείχαν μόνο ηχητικά αποσπάσματα με μουσική χωρίς συνοδεία φωνής, με φωνή χωρίς συνοδεία μουσικής και με μουσική με συνοδεία φωνής, αντίστοιχα. Τα τμήματα τα οποία επισημάνθηκαν ως “Null”, ήταν τα τμήματα τα οποία δεν μπορούσαν να ενταχθούν σε καμία από τις υπόλοιπες κατηγορίες και εν τέλει απορρίφθηκαν εντελώς και δεν εντάχθηκαν σε καμία βάση δεδομένων. Ο σκοπός αυτού του διαχωρισμού είναι να εξεταστεί η απόδοση του συστήματος σε διαφορετικό ηχητικό περιεχόμενο. Έτσι, θα μπορέσουμε να διαπιστώσουμε ποια στοιχεία ενός κομματιού εξετάζει κυρίως το σύστημά μας. Η αποκοπή των επισημανσμένων τμημάτων και η επισήμανσή τους πραγματοποιήθηκε με χρήση του δωρεάν λογισμικού Praat[12]. Για την αποκοπή των ηχητικών αποσπασμάτων χρησιμοποιήσαμε κώδικα στο Praat. Έπειτα, για την κανονικοποίηση των βάσεων δεδομένων χρησιμοποιήθηκε ένας αλγόριθμος σε Python[13], ο οποίος απέρριψε όλα τα ηχητικά αποσπάσματα τα οποία ήταν μικρότερης διάρκειας από τη χρονική διάρκεια που ορίστηκε ανά βάση δεδομένων και διαχώρισε όλα όσα την ξεπερνούσαν. Ο λόγος που επιλέχθηκε διαφορετική χρονική διάρκεια ηχητικών αποσπασμάτων ανά βάση δεδομένων, ήταν επειδή η χρονική κατανομή των τμημάτων, αλλά και το πλήθος αυτών, με σκέτη συνοδεία μουσικών οργάνων, σκέτα φωνητικά, αλλά και μίξη αυτών, υπήρξε άνιση στα τραγούδια και έγινε προσπάθεια να ενταχθούν όσο το δυνατόν περισσότερα ηχητικά αποσπάσματα ανά βάση δεδομένων από διαφορετικά άλμπουμ –έτσι περιορίζεται και το album effect (Kim, et al., 2006)– και διαφορετικές χρονολογίες κυκλοφορίας, προκειμένου να δημιουργηθεί το καλύτερο δυνατό

[12] http://www.fon.hum.uva.nl/praat/
[13] https://www.python.org/

ηχητικό αποτύπωμα για τον εκάστοτε καλλιτέχνη, αλλά και για να εξεταστούν οι ηχητικές μεταβολές στο μουσικό περιεχόμενο των καλλιτεχνών στο πέρασμα του χρόνου. Στη βάση δεδομένων Vocals, δύο συγκροτήματα δεν ήταν δυνατό να συμπεριληφθούν, καθώς δεν εντοπίστηκε σε κανένα κομμάτι τους από όσα διατίθονταν, τμήμα το οποίο να εμπεριείχε φωνητικά χωρίς συνοδεία μουσικών οργάνων. Τα αποσπάσματα ανά κλάση βάσης δεδομένων ήταν ισάριθμα, έτσι ώστε να διατηρηθεί, όσο γίνεται, μία ομοιογένεια εντός των βάσεων δεδομένων αλλά και να δημιουργηθούν κατά το δυνατόν ισάξιες κλάσεις, παρόλα αυτά τα κομμάτια από τα οποία προήλθαν αυτά τα ηχητικά αποσπάσματα δεν είναι ίσα ανά κλάση, αλλά ούτε τα άλμπουμ από όπου προήλθαν τα κομμάτια και κατ' επέκταση τα ηχητικά αποσπάσματα δεν είσαι ίσα ανά κλάση. Από την All λείπουν δύο ηχητικά αποσπάσματα από δύο κλάσεις για να είναι ίσα τα ηχητικά αποσπάσματα ανά κλάση, λόγω της αδυναμίας που αναφέραμε και προηγουμένως να εντοπίσουμε ηχητικά αποσπάσματα με σόλο φωνή σε δύο καλλιτέχνες.

Music_Vocals: Η βάση δεδομένων Music_Vocals, που δημιουργήθηκε για τις ανάγκες των πειραμάτων μας, περιλαμβάνει 450 ηχητικά αποσπάσματα διάρκειας 30 δευτερολέπτων, από 315 μουσικά κομμάτια εξαγμένα από 203 άλμπουμ, τα οποία εμπεριέχουν φωνητικά αναμεμειγμένα με μουσική. Στη βάση δεδομένων περιλαμβάνονται 14 δημοφιλή συγκροτήματα βρετανικής, αμερικάνικης, γερμανικής και αυστραλιανής προέλευσης που ανήκουν στα ευρύτερα είδη της ροκ και της μέταλ, καθώς και ένας ποπ ροκ αμερικάνος τραγουδιστής, στον καθένα εκ των οποίων αντιστοιχούν 30 ηχητικά αποσπάσματα, αποκομμένα από 21 τραγούδια. Οι καλλιτέχνες που περιλαμβάνονται στη συγκεκριμένη βάση δεδομένων είναι οι: The Rolling Stones, Aerosmith, Bon Jovi, Scorpions, ZZ Top, The Beach Boys, Eagles, Mötley Crüe, The Hollies, Def Leppard, Status Quo, AC/DC, Donny Osmond, Black Sabbath και Iron Maiden. Τα κομμάτια των Aerosmith και του Donny Osmond προέρχονται από 14 άλμπουμ, των The Rolling Stones, των The Hollies και του Donny Osmond από 21, των Bon Jovi από 14, των Scorpions από 7, των ZZ Top από 12, των The Beach Boys από 19, των Eagles από 6, των Mötley Crüe και των Def Leppard από 9, των AC/DC από 15, των Black Sabbath από 16 και των Iron Maiden από 8. Όλα τα ηχητικά δείγματα έχουν μετατραπεί σε μορφή μονοφωνικού wave, με ρυθμό δειγματοληψίας 44,1 kHz και βάθος bit 16 bit ανά δείγμα. Οι καλλιτέχνες της βάσης δεδομένων ανήκουν στα ακόλουθα γενικευμένα είδη: rock, heavy metal και pop.

Music: Η βάση δεδομένων Music, που δημιουργήθηκε επίσης για τις ανάγκες της συγκεκριμένης εργασίες, περιλαμβάνει 525 ηχητικά αποσπάσματα διάρκειας 7 δευτερολέπτων, εξαγμένα από 360 μουσικά κομμάτια από 204 άλμπουμ, τα οποία εμπεριέχουν αποκλειστικά ηχητική πληροφορία μουσικών οργάνων. Σε αυτή τη βάση δεδομένων περιλαμβάνονται οι ίδιοι 15 καλλιτέχνες που περιλαμβάνονται και στη Music_Vocals, στον καθένα εκ των οποίων αντιστοιχούν 35 ηχητικά αποσπάσματα, αποκομμένα από 24 τραγούδια. Τα κομμάτια των The Rolling Stones, των Aerosmith και των AC/DC εξάχθηκαν από 15 άλμπουμ, των Bon Jovi από 12, των Scorpions από 7, των ZZ Top από 14, των The Beach Boys από 21, των Eagles από 6, των Mötley Crüe και των Def Leppard από 9, των The Hollies και των Status Quo από 22, του Donny Osmond από 10, των Black Sabbath από 19 και των Iron Maiden από 8. Η αρχική μορφή όλων των κομματιών των The Rolling Stones, των ZZ Top και των The Beach Boys ήταν αυτή του στερεοφωνικού FLAC, με ρυθμό δειγματοληψίας 44100 Hz και βάθος bit 16, των Aerosmith ήταν στερεοφωνικά WAV, με ρυθμό δειγματοληψίας 44100 Hz και βάθος bit 16, των Bon Jovi, των Mötley Crüe, των The Hollies, των Def Leppard, των Status Quo και του Donny Osmond, στερεοφωνικά MP3 με ρυθμό δειγματοληψίας 44100 Hz και βάθος bit 16, των Scorpions ήταν τα 15 ηχητικά αποσπάσματα που προέρχονται από 9 κομμάτια, στερεοφωνικά FLAC με ρυθμό δειγματοληψίας 192 kHz και βάθος bit 24, όπως και τα κομμάτια των Eagles, των AC/DC και των Black Sabbath και τα υπόλοιπα 20 ηχητικά αποσπάσματα που προέρχονται από 13 κομμάτια, ήταν στερεοφωνικά FLAC με ρυθμό δειγματοληψίας 96 kHz και βάθος bit 24 και των Iron Maiden τα 10 ηχητικά αποσπάσματα που εξάχθηκαν από 6 μουσικά κομμάτια ήταν στερεοφωνικά MPEG-4, με ρυθμό δειγματοληψίας 48 kHz και βάθος bit 24, τα 16 ηχητικά αποσπάσματα 12 κομματιών ήταν στερεοφωνικά FLAC με ρυθμό δειγματοληψίας 44100 Hz και βάθος bit 24 και τα υπόλοιπα 9 ηχητικά αποσπάσματα 6 κομματιών ήταν FLAC με ρυθμό δειγματοληψίας 44100 Hz και βάθος bit 16. Για την ηχητική μορφή των αποσπασμάτων χρησιμοποιήθηκαν οι ίδιες ρυθμίσεις με εκείνες της Music_Vocals.

Vocals: Η βάση δεδομένων Vocals, είναι η τρίτη βάση δεδομένων που δημιουργήθηκε για τις ανάγκες αυτής της εργασίας και περιλαμβάνει 26 ηχητικά αποσπάσματα διάρκειας ενός δευτερολέπτου, από 13 μουσικά κομμάτια και 13 άλμπουμ, δηλαδή από 1 άλμπουμ ανά καλλιτέχνη, τα οποία εμπεριέχουν μόνο πληροφορία φωνής. Αυτή η βάση περιλαμβάνει τους 13 από τους 15 καλλιτέχνες των συλλογών Music, Music_Vocals και All, σε καθέναν από τους οποίους αντιστοιχούν δύο ηχητικά αποσπάσματα αποκομμένα από ένα κομμάτι τους. Οι

αρχικές μορφές του κομματιών των Rolling Stones και των Beach Boys ήταν στερεοφωνικά FLAC με ρυθμό δειγματοληψίας 44100 Hz και βάθος bit 16, των Aerosmith, των AC/DC και των Black Sabbath, στερεοφωνικά WAV με ρυθμό δειγματοληψίας 44100 Hz και βάθος bit 16, των Bon Jovi, των ZZ Top, των Mötley Crüe, των The Hollies, των Status Quo και του Donny Osmond, στερεοφωνικά MP3 με ρυθμό δειγματοληψίας 44100 Hz και βάθος bit 16, των Def Leppard ήταν στερεοφωνικά AIFF-C, με ρυθμό δειγματοληψίας 44100 Hz και βάθος bit 16 και των Iron Maiden ήταν στερεοφωνικά MPEG-4 με ρυθμό δειγματοληψίας 48000 Hz και βάθος bit 24. Οι καλλιτέχνες που δεν περιλαμβάνονται είναι οι Scorpions και οι Eagles. Και σε αυτή τη βάση δεδομένων χρησιμοποιήθηκαν οι ίδιες ρυθμίσεις στα αρχεία ήχου με αυτές που χρησιμοποιήθηκαν και στη Music_Vocals και στη Music.

All: Η All εμπεριέχει όλα τα ηχητικά αποσπάσματα των βάσεων δεδομένων Music_Vocals, Music και Vocals, δηλαδή 1001 ηχητικά αποσπάσματα, από 458 μουσικά κομμάτια τα οποία εξάχθηκαν από 221 άλμπουμ. Τα 22 εκ των 222 άλμπουμ είναι των The Rolling Stones, τα 15 των Aerosmith, τα 12 των Bon Jovi, τα 7 των Scorpions, τα 14 των ZZ Top, τα 26 των The Beach Boys, τα 6 των Eagles, τα 9 των Mötley Crüe, τα 22 των The Hollies, τα 9 των Def Leppard, τα 24 των Status Quo, τα 15 των AC/DC, τα 14 του Donny Osmond, τα 19 των Black Sabbath και τα 8 των Iron Maiden.

Για τις ανάγκες της πειραματικής διαδικασίας χρησιμοποιήθηκαν μόνο η Music_Vocals και η Music γιατί είναι οι μόνες οι οποίες εμπεριέχουν ίσο αριθμό ηχητικό αποσπασμάτων ανά κλάση, ίσης διάρκειας ηχητικά αποσπάσματα, αλλά και ηχητικά αποσπάσματα σε όλες τις κλάσεις. Η Vocals εμπεριέχει ούτως ή άλλως πολύ λίγα ηχητικά αποσπάσματα για να έχει νόημα η συμπερίληψή της στην πειραματική διαδικασία. Από τις Music_Vocals και Music εξαιρέθηκαν 4 ηχητικά αποσπάσματα ανά κλάση από 2 κομμάτια ανά καλλιτέχνη τα οποία και αποτέλεσαν δύο διαφορετικά σύνολα δοκιμών –ένα για τη Music_Vocals και ένα για τη Music, τα TestMV και TestM, αντίστοιχα– για τη διεξαγωγή της πειραματικής διαδικασίας. Έτσι, δημιουργήθηκαν και τα σύνολα εκπαίδευσης MUSVOC και MUS.

4.2 Χαρακτηριστικά εκπαίδευσης & αλγορίθμων ταξινόμησης

Για την εξαγωγή χαρακτηριστικών χρησιμοποιήθηκαν οι βάσεις MUSVOC και MUSIC. Έγινε χρήση της βιβλιοθήκης ανοιχτού κώδικα MIR Essentia[14]. Η βιβλιοθήκη αυτή

[14] https://essentia.upf.edu/

διαθέτει πληθώρα από αλγόριθμους εξαγωγής χαρακτηριστικών βασισμένους στη διεθνή βιβλιογραφία. Σε πολλές μελέτες χρησιμοποιείται ως βασικό εργαλείο εξαγωγής χαρακτηριστικών και ως βιβλιοθήκη χρησιμοποιείται σε πολυάριθμες εφαρμογές εμπορικών οργανισμών.

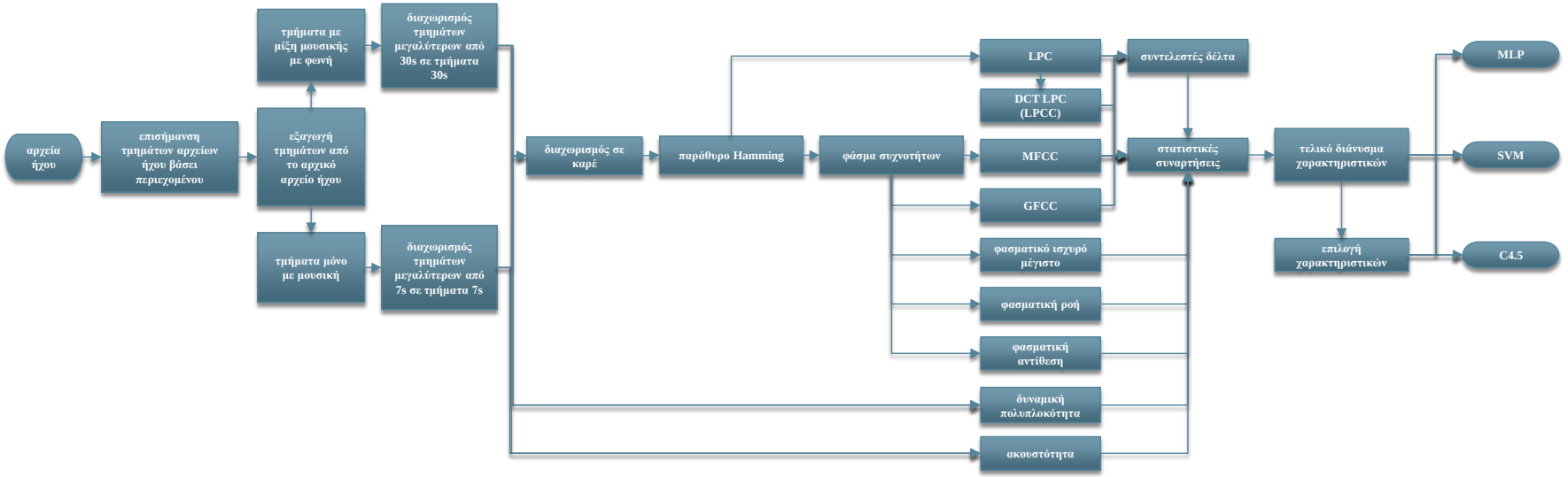
Πραγματοποιήθηκαν πολλές δοκιμές έως ότου καταλήξουμε στο σύνολο χαρακτηριστικών που καταλήξαμε τελικά, οι οποίες παρουσιάζονται στον πίνακα 2.

Πίνακας 2: Αποτελέσματα κατηγοριοποίησης SVM, στις βάσεις δεδομένων MUS_VOC και All, με διάφορα σύνολα χαρακτηριστικών και διαφορετική διάρκεια μέσου όρου με χρήση συνόλου δοκιμής.

Σύνολο χαρακτηριστικών	**Διάρκεια αποσπάσματος μέσου όρου (s)**	**Ακρίβεια (%)**	**Βάση δεδομένων**
SpFlx, StrPk, SCC, GFCC, MFCC, LPC, LPCC, Dlt, DynCmplx, Ldns	30	**78,33**	All
SpFlx, StrPk, SCC, GFCC, MFCC, LPC, LPCC, Dlt, DynCmplx, Ldns	30	75	MUSVOC
HPCP, GFCC, MFCC, LPC, LPCC, Dlt	30	73,33	MUSVOC
SpFlx, MMF, StrPk, SCC, HPCP, GFCC, MFCC, LPC, LPCC, Dlt	30	71,67	MUSVOC
MMF, StrPk, SSC, HPCP, GFCC, MFCC, LPC, LPCC, Dlt	30	70,00	MUSVOC
MFCC, LPCC, LPC, Dlt	30	63,33	MUSVOC
HPCP, GFCC, MFCC, LPC, LPCC, Dlt	10	58,33	MUSVOC
GFCC	30	56,67	MUSVOC
MFCC, DMFCC	30	51,67	MUSVOC
HPCP, GFCC, MFCC, LPC, LPCC, Dlt	2,5	45,00	MUSVOC

Έτσι καταλήξαμε στο σύνολο χαρακτηριστικών που περιλαμβάνει τα εξής χαρακτηριστικά: φασματική ροή (SpFlx), ισχυρό μέγιστο (strong peak, StrPk), φασματική αντίθεση (spectral contrast, SCC), φασματικοί συντελεστές γαμματονικής συχνότητας (GFCC), φασματικοί συντελεστές mel συχνότητας (MFCC), συντελεστές γραμμικής πρόβλεψης (LPC), φασματικοί συντελστές γραμμικής πρόβλεψης (LPCC), τους συντελεστές δέλτα (Dlt) των MFCC (DMFCC), LPC (DLPC, LPCC (DLPCC), GFCC (DGFCC), δυναμική πολυπλοκότητα (dynamic complexity, DynCmplx) και ακουστότητα (loudness, Ldns).

Η προτεινόμενη διαδικασία εξαγωγής χαρακτηριστικών περιγράφεται στο σχήμα 8.



Σχήμα 11: Διάγραμμα ροής πειραματικής διαδικασίας

Η μεθοδολογία που χρησιμοποιήθηκε για την εξαγωγή των διανυσμάτων των χαρακτηριστικών των μουσικών κομματιών αποτελείται από διάφορες επιμέρους εργασίες, καθεμία εκ των οποίων εξάγει χαρακτηριστικά τα οποία αργότερα θα τροφοδοτήσουν τους αλγόριθμους κατηγοριοποίησης για τη διενέργηση των πειραμάτων.

Κάθε μουσικό κομμάτι εισάγεται στο λογισμικό και εξάγονται χαρακτηριστικών δυναμικής πολυπλοκότητας (dynamic complexity) και ακουστότητας (loudness) για το σύνολο της διάρκειάς του. Έπειτα πραγματοποιείται διαχωρισμός του σε τμήματα διάρκειας 2048 δειγμάτων για περαιτέρω επεξεργασία με αλληλεπικάλυψη (overlapping) διάρκειας 1024 δειγμάτων. Αυτό έχει ως αποτέλεσμα την αύξηση της ακρίβειας των αλγορίθμων εξαγωγής χαρακτηριστικών. Σε κάθε τμήμα εφαρμόζεται μία συνάρτηση παραθύρου hamming για την εξομάλυνση των άκρων του τμήματος, καθώς και για τη διατήρηση της συχνοτικής πληροφορίας του. Στη συνέχεια το κάθε επεξεργασμένο τμήμα τροφοδοτεί τον αλγόριθμο εξαγωγής LPC και εξαγωγής φάσματος.

Η εξαγωγή του φάσματος αποτελεί ένα απαραίτητο βήμα για την εξαγωγή των ακόλουθων χαρακτηριστικών και υπολογίζεται με 1024 διακριτές τιμές ανά τμήμα μέσω της χρήσης του αλγορίθμου FFT. Από το εκάστοτε φάσμα του τμήματος υπολογίζονται τα χαρακτηριστικά GFCC (Γαμματονικοί φασματικοί συντελεστές), MFCC, spectral strong peak (Φασματική ισχυρή κορυφή), spectral flux (φασματική ροή) και spectral contrast (φασματική αντίθεση). Χρησιμοποιήθηκαν οι προεπιλεγμένες ρυθμίσεις των χαρακτηριστικών, οι οποίες είναι, για τους GFCC η υψηλότερη συχνότητα υπολογισμού είναι τα 22050 Hz, το πλήθος των gammatone filter banks (γαμματονικά ζωνοπερατά φίλτρα) είναι 40 και ο αριθμός των συνιστωσών 13, για τους MFCC χρησιμοποιήθηκε ο δεύτερος τύπος DCT (διακριτής συνημιτονοειδούς συνάρτησης) με υψηλότερη συχνότητα υπολογισμού τα 11000 Hz, με πλήθος mel scale banks (ζωνών κλίμακας mel) 40 και πλήθος συνιστωσών 13. Η υψηλότερη συχνότητα υπολογισμού της φασματικής αντίθεσης είναι τα 11000 Hz με πλήθος ζωνών 6 και η διαφορά της φασματικής ροής υπολογίστηκε με τον αλγόριθμο L2-Norm.

LPC (συντελεστές γραμμικής πρόβλεψης): υπολογίζονται από τα τμήματα του μουσικού κομματιού προτού πραγματοποιηθεί εξαγωγή του φάσματός του. Οι LPC με τη σειρά τους τροφοδοτούνται στον αλγόριθμο DCT-II για την εξαγωγή των LPCC με τις προεπιλεγμένες από το πρόγραμμα, τιμές των μεταβλητών. Είναι σύνηθες να υπολογίζονται και οι διαφορές του κάθε τμήματος από το προηγούμενό του για τα χαρακτηριστικά LPC, LPCC, GFCC και

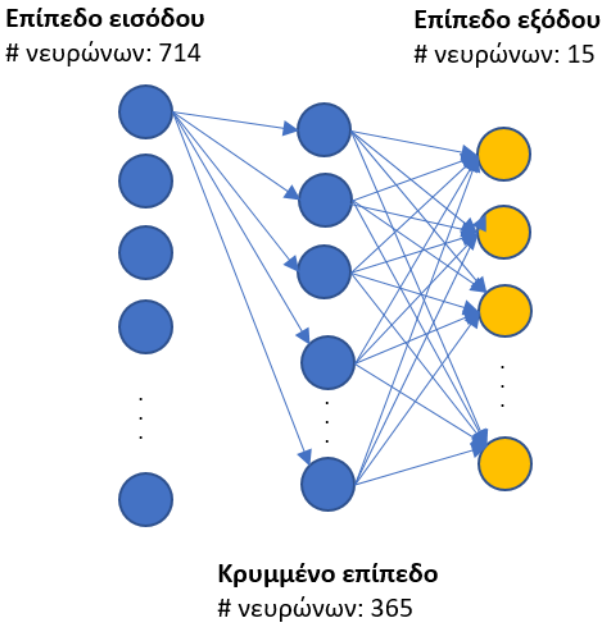
MFCC, οπότε υπολογίστηκε και η 1η παράγωγος ή αλλιώς delta για τα συγκεκριμένα χαρακτηριστικά. Στο τελικό διάνυσμα συμπεριλαμβάνονται και οι αρχικές τιμές των χαρακτηριστικών.

Στα διανύσματα χαρακτηριστικών που εξήχθησαν στα προηγούμενα στάδια εφαρμόστηκαν οι παρακάτω στατιστικές συναρτήσεις: μέσος όρος, τυπική απόκλιση, ελάχιστο, μέγιστο, διάμεσος και διακύμανση. Αυτό εξυπηρετεί την παραγωγή ενός διανύσματος χαρακτηριστικών ανά αρχείο για τη μετέπειτα εισαγωγή του στους αλγόριθμους κατηγοριοποίησης. Το τελικό διάνυσμα χαρακτηριστικών που παράγεται έπειτα από την επεξεργασία κάθε αρχείου ήχου αποτελείται συνολικά από 714 τιμές.

Οι αλγόριθμοι κατηγοριοποίησης που χρησιμοποιήθηκαν για την πραγματοποίηση της κατηγοριοποίησης ήταν οι SVM, MLP και C4.5. Η κατηγοριοποίηση έγινε με τη χρήση του προγράμματος Weka[15]. Ο λόγος που χρησιμοποιήθηκε ο αλγόριθμος SVM είναι γιατί πραγματοποιήθηκε πείραμα στο Matlab[16] όπου χρησιμοποιήθηκαν όλοι οι διαθέσιμοι αλγόριθμοί του και ο SVM έβγαλε τη μεγαλύτερη ακρίβεια. Ο MLP είναι ένας εξαιρετικά δημοφιλής αλγόριθμος στον τομέα της εξόρυξης δεδομένων και είναι ικανός να εκτελέσει πολύ πολύπλοκες κατηγοριοποιήσεις και ο C4.5 χρησιμοποιήθηκε για λόγους σύγκρισης, ως ο απλούστερος εκ των τριών. Χρησιμοποιήθηκαν οι προεπιλεγμένες ρυθμίσεις των αλγορίθμων του προγράμματος. Κάναμε δύο κύκλους πειραμάτων, έναν με επιλογή χαρακτηριστικών και έναν χωρίς για να διαπιστώσουμε εάν αυτή η επιλογή χαρακτηριστικών θα ήταν ικανή να βελτιώσει την ακρίβεια των αλγορίθμων κατηγοριοποίησης. Η επιλογή χαρακτηριστικών πραγματοποιήθηκε μέσω της χρήσης του αλγορίθμου Info Gain Evaluation και με τη μέθοδο Ranker. Για την επαλήθευση της κατηγοριοποίησης δημιουργήθηκαν δύο σύνολα δοκιμών με κομμάτια αντλημένα από τις ίδιες τις βάσεις τα οποία και εξαιρέθηκαν από τα σύνολα εκπαίδευσης, ένα για τη βάση δεδομένων MUSVOC και ένα για τη MUS τα οποία εμπεριείχαν τέσσερα ηχητικά αποσπάσματα (δύο ανά τραγούδι) ανά καλλιτέχνη. Επιλέξαμε αυτή τη μέθοδο έναντι του cross-validation (διασταυρούμενη επικύρωση) προκειμένου να διασφαλίσουμε ότι κατά τη διαδικασία της επαλήθευσης δε θα εμπεριέχονται στα σύνολα δοκιμής και εκπαίδευσης κοινά κομμάτια, έτσι ώστε να γνωρίζουμε εάν το σύστημά μας θα είναι ικανό να αναγνωρίσει κομμάτια των καλλιτεχνών που εμπεριέχονται στη βάση δεδομένων μας τα οποία δεν εμπεριέχονται στη βάση και να αποφευχθεί το υπερταίριασμα

[15] https://www.cs.waikato.ac.nz/ml/weka/
[16] https://www.mathworks.com/products/matlab.html

(overfitting) (Leinweber, 2007). Στο τέλος επιλέξαμε από ένα ολόκληρο κομμάτι ανά καλλιτέχνη, το οποίο δεν υπήρχε ούτε στο σύνολο εκπαίδευσης ούτε στο σύνολο δοκιμών και δοκιμάσαμε να εκτελέσουμε ξανά τη διαδικασία της κατηγοριοποίησης με τον αλγόριθμο και τη βάση δεδομένων που μας έβγαλε τα καλύτερα αποτελέσματα προκειμένου να διαπιστώσουμε κατά πόσο είναι ικανή να γενικεύει η υλοποίησή μας.



Σχήμα 12: Σχηματική απεικόνιση του MLP που χρησιμοποιήθηκε κατά την πειραματική διαδικασία

Πίνακας 3: Χαρακτηριστικά βάσεων δεδομένων και αλγόριθμοι που χρησιμοποιήθηκαν στην υλοποίησή μας

Σύνολα πειραμάτων	**Δεδομένα**	**Αλγόριθμοι**	**Κλάσεις**
MUS	525 ηχητικά αποσπάσματα από 204 άλμπουμ	SVM MLP C4.5	15
MUSVOC	450 ηχητικά αποσπάσματα από 203 άλμπουμ		

Κεφάλαιο 5ο: Αξιολόγηση υλοποίησης-Πειράματα

5.1 Πειράματα

Πίνακας 4: Ποσοστά σωστής κατηγοριοποίησης αλγόριθμόν στη βάση δεδομένων MUSVOC χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με (OPTIMAL_SET)

ACCURACY (%)	MLP	SVM	C4.5
FULL_SET	75	72	35
OPTIMAL_SET	75	72	35

Πίνακας 5: Ακρίβειες αλγόριθμων κατηγοριοποίησης στη βάση δεδομένων MUSVOC χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με (OPTIMAL_SET)

PRECISION	MLP	SVM	C4.5
FULL_SET	0,801	0,747	0,407
OPTIMAL_SET	0,801	0,747	0,407

Πίνακας 6: Ανακλήσεις αλγόριθμων κατηγοριοποίησης στη βάση δεδομένων MUSVOC χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με (OPTIMAL_SET)

RECALL	MLP	SVM	C4.5
FULL_SET	0,75	0,717	0,35
OPTIMAL_SET	0,75	0,717	0,35

Πίνακας 7: Ποσοστά σωστής κατηγοριοποίησης αλγόριθμων στη βάση δεδομένων MUS χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με (OPTIMAL_SET)

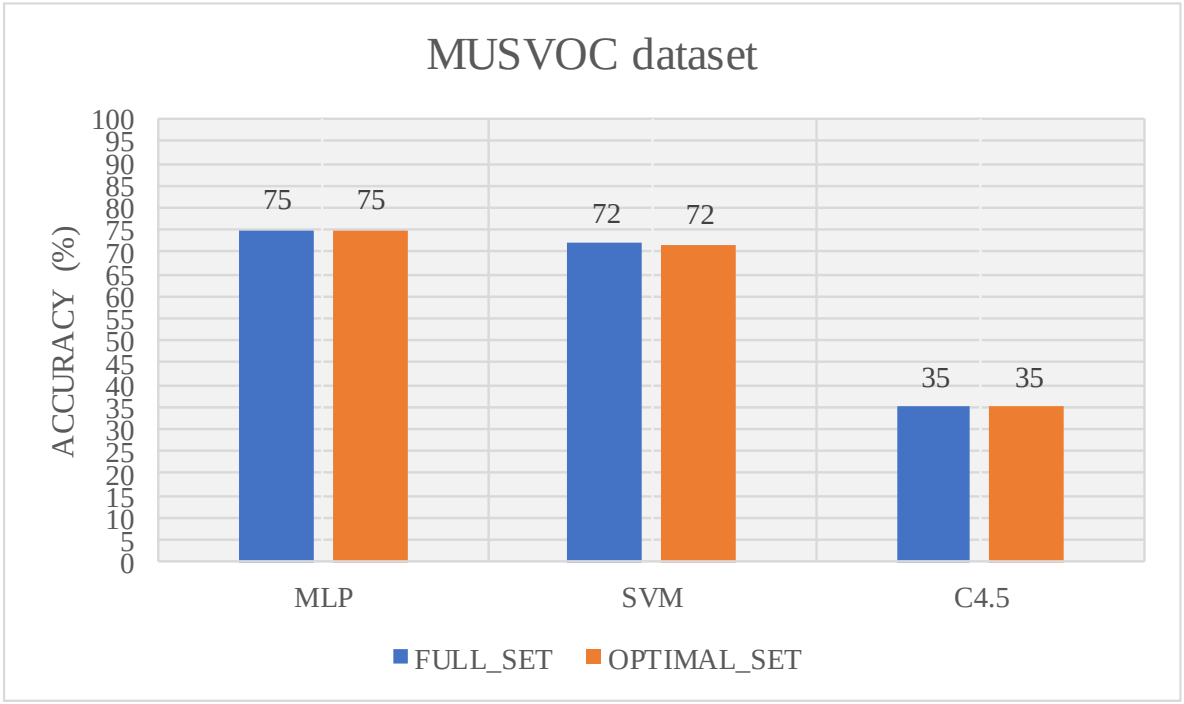
ACCURACY	MLP	SVM	C4.5
FULL_SET	32	40	27
OPTIMAL_SET	32	40	27

Πίνακας 8: Ακρίβειες αλγόριθμων κατηγοριοποίησης στη βάση δεδομένων MUS χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με (OPTIMAL_SET)

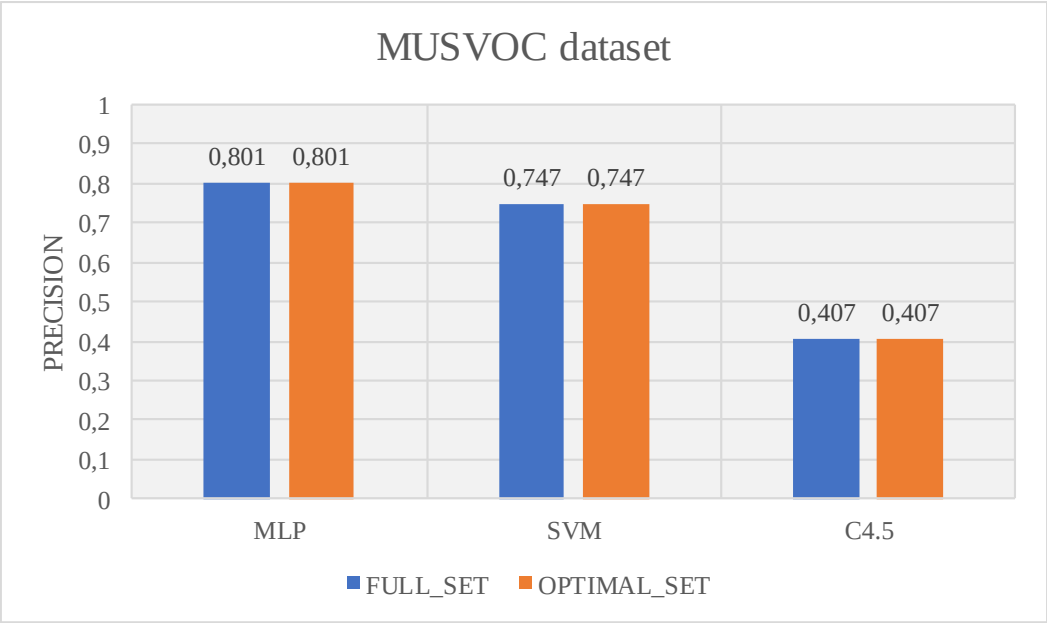
PRECISION	MLP	SVM	C4.5
FULL_SET	0,306	0	0,287
OPTIMAL_SET	0,306	0	0,287

Πίνακας 9: Ανακλήσεις αλγόριθμων κατηγοριοποίησης στη βάση δεδομένων MUS χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με (OPTIMAL_SET)

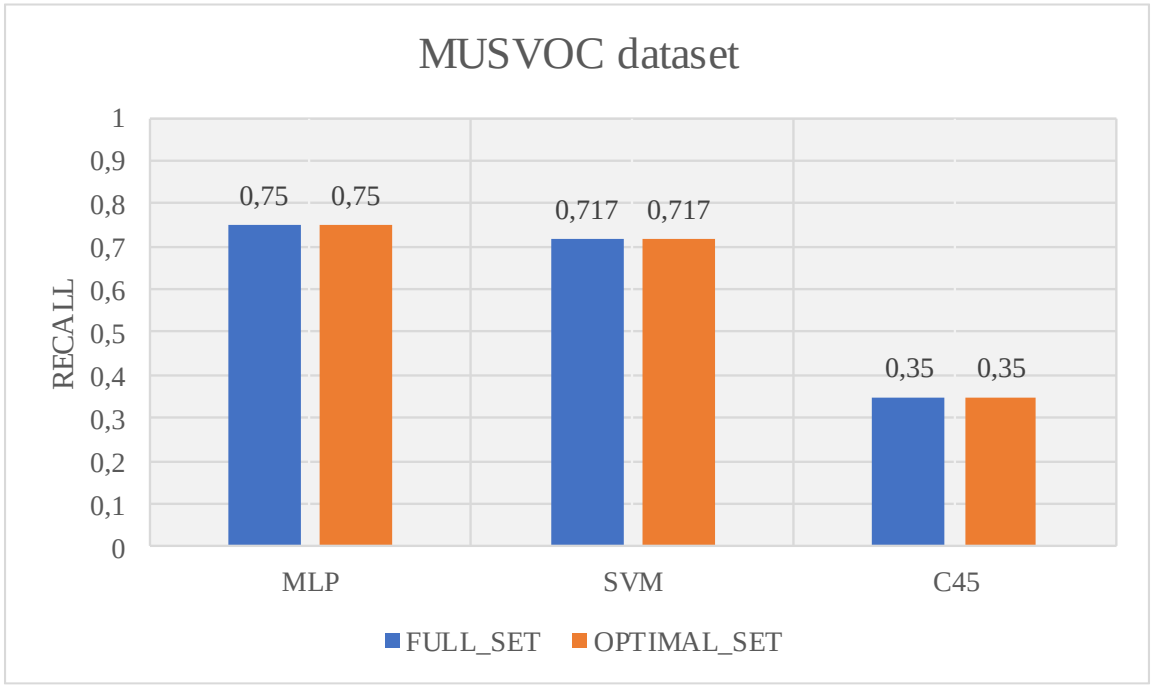
RECALL	MLP	SVM	C45
FULL_SET	0,317	0,4	0,267
OPTIMAL_SET	0,317	0,4	0,267



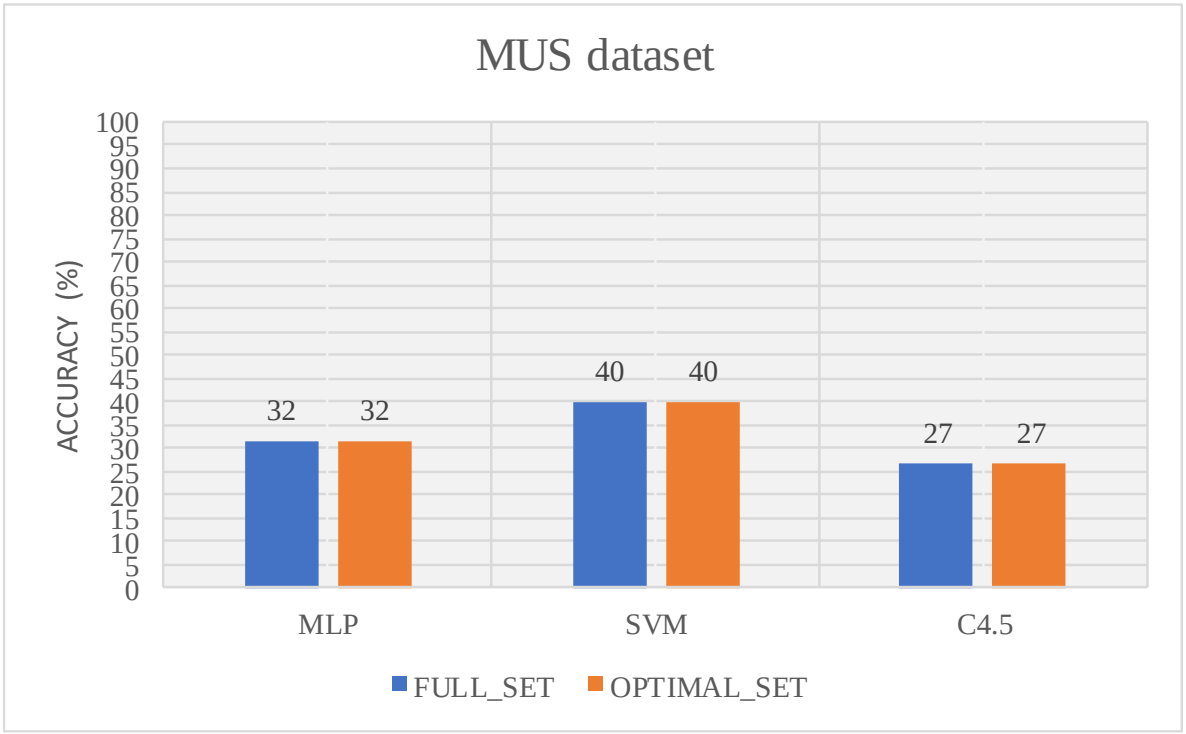
Σχήμα 13: Ποσοστά σωστής κατηγοριοποίησης της βάσης δεδομένων MUSVOC για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)



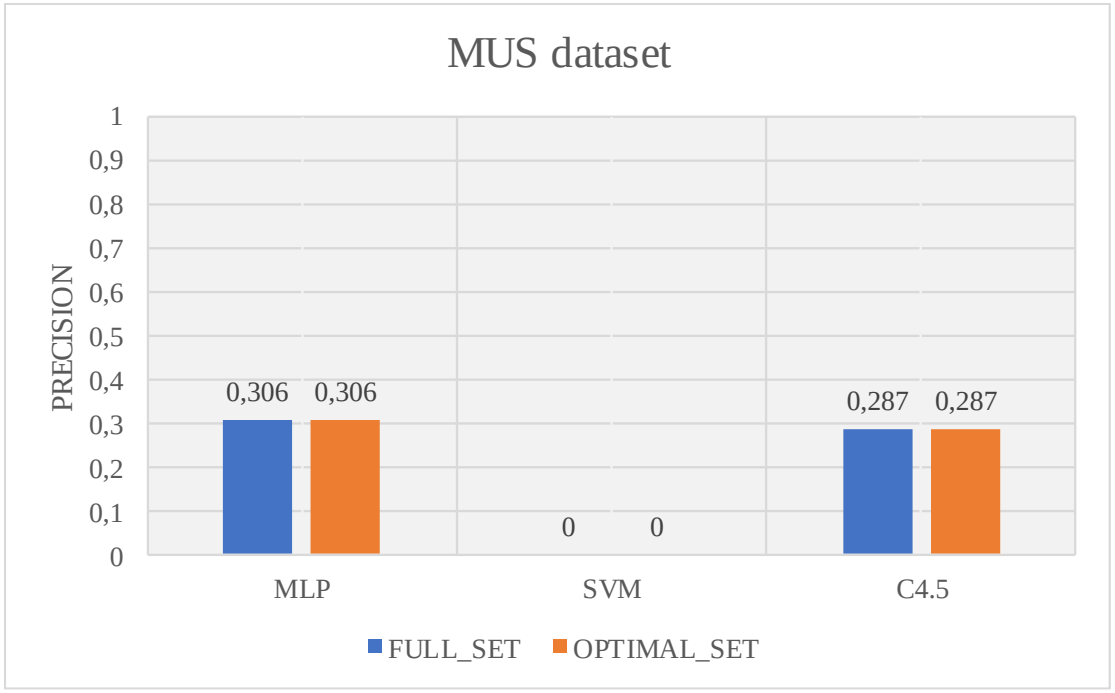
Σχήμα 14: Ακρίβειες της βάσης δεδομένων MUSVOC για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)



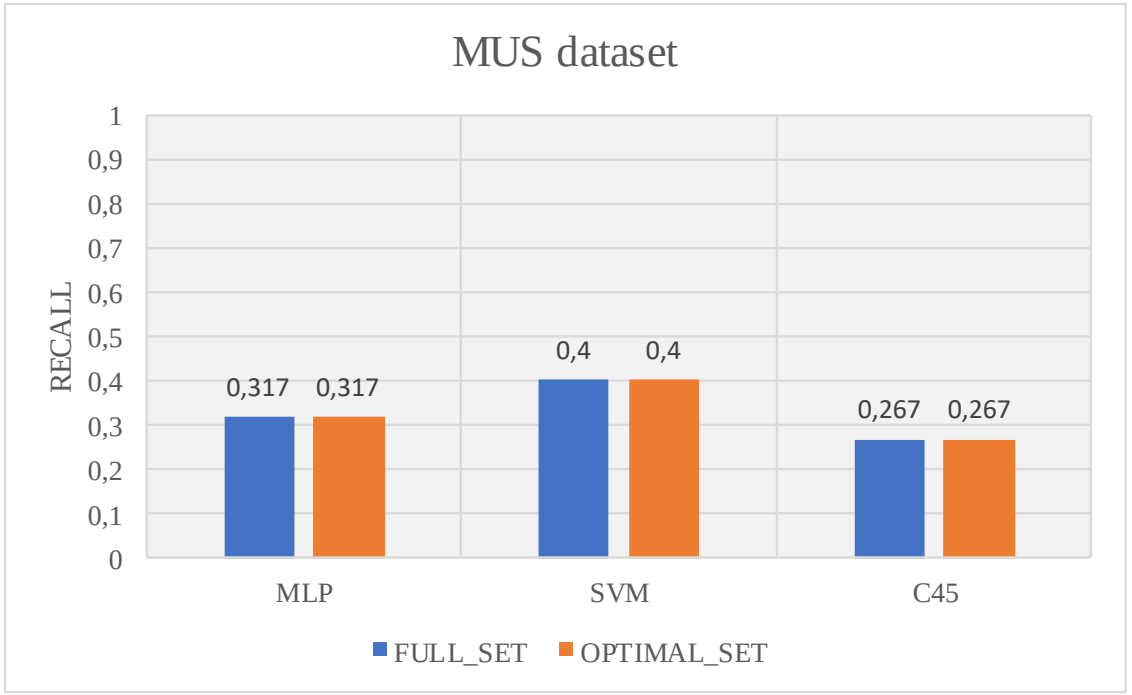
Σχήμα 15: Ανακλήσεις της βάσης δεδομένων MUSVOC για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)



Σχήμα 16: Ποσοστά σωστής κατηγοριοποίησης της βάσης δεδομένων MUS για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)



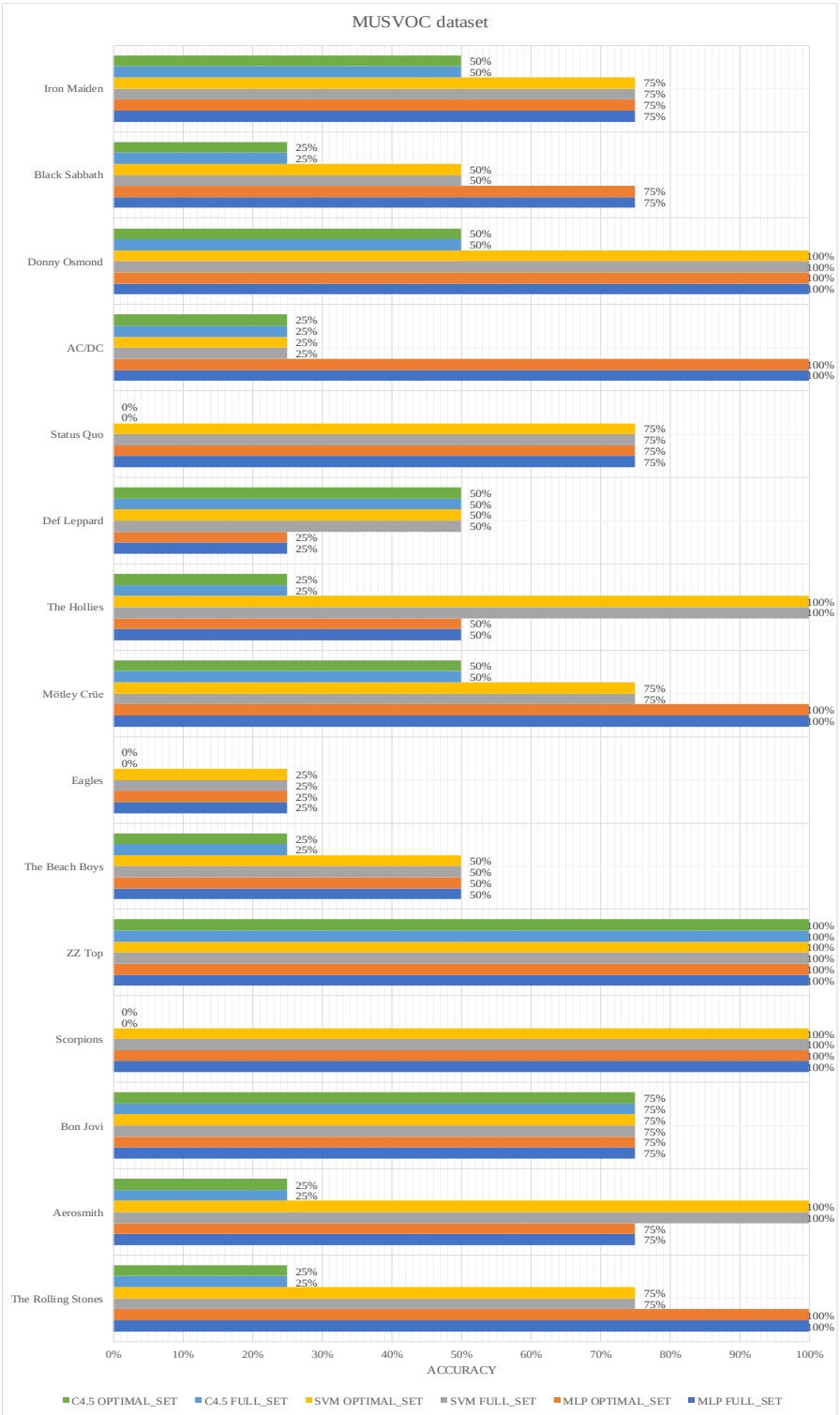
Σχήμα 17: Ακρίβειες της βάσης δεδομένων MUS για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)



Σχήμα 18: Ανακλήσεις της βάσης δεδομένων MUS για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)

ACCURACY	MLP FULL_SET	MLP OPTIMAL_SET	SVM FULL_SET	SVM OPTIMAL_SET	C4.5 FULL_SET	C4.5 OPTIMAL_SET
The Rolling Stones	100%	100%	75%	75%	25%	25%
Aerosmith	75%	75%	100%	100%	25%	25%
Bon Jovi	75%	75%	75%	75%	75%	75%
Scorpions	100%	100%	100%	100%	0%	0%
ZZ Top	100%	100%	100%	100%	100%	100%
The Beach Boys	50%	50%	50%	50%	25%	25%
Eagles	25%	25%	25%	25%	0%	0%
Mötley Crüe	100%	100%	75%	75%	50%	50%
The Hollies	50%	50%	100%	100%	25%	25%
Def Leppard	25%	25%	50%	50%	50%	50%
Status Quo	75%	75%	75%	75%	0%	0%
AC/DC	100%	100%	25%	25%	25%	25%
Donny Osmond	100%	100%	100%	100%	50%	50%
Black Sabbath	75%	75%	50%	50%	25%	25%
Iron Maiden	75%	75%	75%	75%	50%	50%

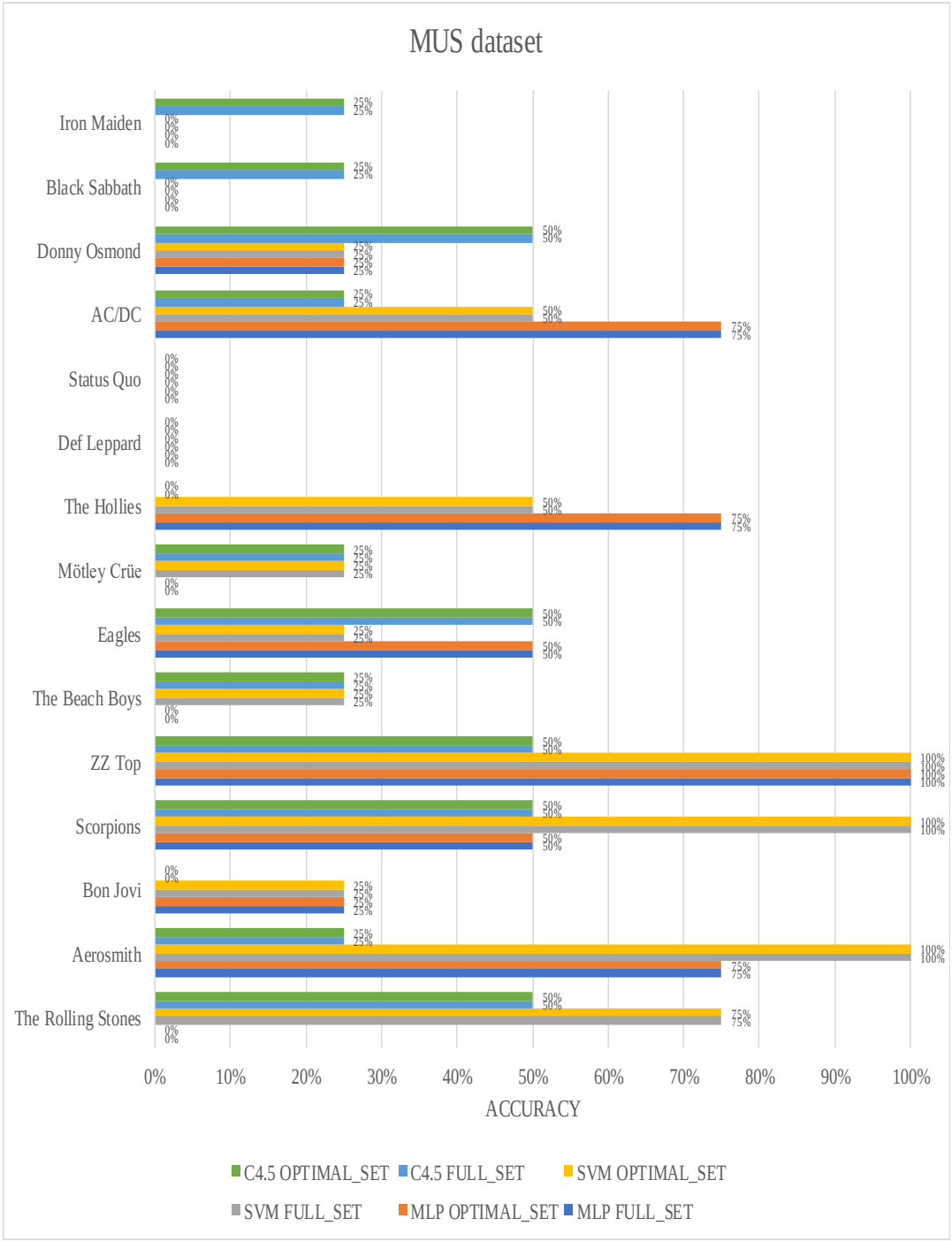
Σχήμα 19: Ποσοστά σωστής κατηγοριοποίησης ανά καλλιτέχνη, της βάσης δεδομένων MUSVOC για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)



Σχήμα 20: Ποσοστά σωστής κατηγοριοποίησης ανά καλλιτέχνη, της βάσης δεδομένων MUSVOC για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)

ACCURACY	MLP FULL_SET	MLP OPTIMAL_SET	SVM FULL_SET	SVM OPTIMAL_SET	C4.5 FULL_SET	C4.5 OPTIMAL_SET
The Rolling Stones	0%	0%	75%	75%	50%	50%
Aerosmith	75%	75%	100%	100%	25%	25%
Bon Jovi	25%	25%	25%	25%	0%	0%
Scorpions	50%	50%	100%	100%	50%	50%
ZZ Top	100%	100%	100%	100%	50%	50%
The Beach Boys	0%	0%	25%	25%	25%	25%
Eagles	50%	50%	25%	25%	50%	50%
Mötley Crüe	0%	0%	25%	25%	25%	25%
The Hollies	75%	75%	50%	50%	0%	0%
Def Leppard	0%	0%	0%	0%	0%	0%
Status Quo	0%	0%	0%	0%	0%	0%
AC/DC	75%	75%	50%	50%	25%	25%
Donny Osmond	25%	25%	25%	25%	50%	50%
Black Sabbath	0%	0%	0%	0%	25%	25%
Iron Maiden	0%	0%	0%	0%	25%	25%

Σχήμα 21: Ποσοστά σωστής κατηγοριοποίησης ανά καλλιτέχνη, της βάσης δεδομένων MUS για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)



Σχήμα 22: Ποσοστά σωστής κατηγοριοποίησης ανά καλλιτέχνη, της βάσης δεδομένων MUS για τους αλγόριθμους MLP, SVM και C4.5 χωρίς επιλογή χαρακτηριστικών (FULL_SET) και με επιλογή χαρακτηριστικών (OPTIMAL_SET)

Πίνακας 10: Πίνακας σύγχυσης, κατηγοριοποίησης ολόκληρων κομματιών εκτός συνόλου δοκιμής και εκπαίδευσης με χρήση του αποτελεσματικότερου αλγορίθμου και βάσης δεδομένων

Πίνακας σύγχυσης															
a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**n**	**o**	**←ταξινομήθηκαν ως**
9	0	0	0	1	0	0	0	0	0	0	0	0	0	0	**a=The Rolling Stones**
1	**5**	0	0	0	0	0	0	0	0	2	0	0	0	0	**b=Aerosmith**
0	0	**7**	0	0	2	0	0	0	0	0	0	0	0	0	**c=Bon Jovi**
0	0	0	**5**	0	0	0	0	0	0	0	2	0	0	1	**d=Scorpions**
0	0	0	0	**5**	0	0	0	1	0	0	0	0	0	0	**e=ZZ Top**
0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	0	**f=The Beach Boys**
2	0	0	0	1	0	**0**	0	5	0	0	1	0	0	0	**g=Eagles**
0	0	0	0	1	0	0	**5**	0	0	0	0	0	0	0	**h=Mötley Crüe**
0	0	0	0	0	1	0	0	**3**	0	0	0	0	0	0	**i=The Hollies**
0	0	0	0	0	0	0	0	0	**0**	0	7	0	1	0	**j=Def Leppard**
4	1	0	0	1	0	0	0	0	0	**0**	0	0	0	0	**k=Status Quo**
4	0	0	0	1	0	0	0	0	0	2	**0**	0	0	0	**l=AC/DC**
0	0	0	0	0	1	0	0	0	0	0	0	**4**	0	0	**m=Donny Osmond**
0	0	0	0	0	0	0	0	0	0	0	3	0	**15**	0	**n=Black Sabbath**
0	0	0	0	1	0	0	0	0	0	0	1	0	8	**0**	**o=Iron Maiden**

5.2 Αξιολόγηση υλοποίησης

Αξιολογώντας τα αποτελέσματα των πειραμάτων μας καταλήγουμε στο συμπέρασμα ότι η μέγιστη ακρίβεια επιτυγχάνεται με τη χρήση του αλγορίθμου MLP, η οποία είναι έως 75%, ενώ με τον SVM πήραμε ακρίβεια έως περίπου 72% ενώ με τον C4.5 έως 35%. Τα αποτελέσματα αυτά είναι αναμενόμενα αν αναλογιστούμε την πολυπλοκότητα των αλγορίθμων. Συγκρίνοντας την απόδοση των αλγορίθμων στη βάση MUSVOC και MUS, 75% και περίπου 32% για τον MLP, περίπου 72% και 40% για τον SVM και 35% και περίπου 27% για τον C4.5, αντίστοιχα, καταλήγουμε στο συμπέρασμα ότι το σύστημά μας λαμβάνει σοβαρά υπόψη του τη φωνή του εκάστοτε τραγουδιστή για την αναγνώριση. Ενδιαφέρον παρουσιάζει το γεγονός ότι στη βάση δεδομένων MUS, ως πιο αποδοτικός αλγόριθμος παρουσιάζεται ο SVM. Εκτελέστηκε και ένα πείραμα στο οποίο επιχειρήθηκε να κατηγοριοποιηθεί από ένα ολόκληρο κομμάτι ανά καλλιτέχνη. Κανένα από αυτά τα κομμάτια δεν εμπεριέχονταν ούτε στο σύνολο εκπαίδευσης, ούτε στο σύνολο δοκιμών. Κατηγοριοποίηθηκαν σωστά οι 10 από τους 15 καλλιτέχνες. Οι καλλιτέχνες που κατηγοριοποιήθηκαν λάθος είναι οι: Eagles, Def Leppard, Status Quo, AC/DC και Iron Maiden. Συγκεκριμένα, οι Eagles κατηγοριοποιήθηκαν ως The Hollies, οι Def Leppard ως AC/DC, οι Status Quo ως The Rolling Stones, οι AC/DC ως The Rolling Stones και οι Iron Maiden ως Black Sabbath. Από τις ακρίβειες που εξάγαμε έπειτα από επιλογή χαρακτηριστικών καταλήγουμε στο συμπέρασμα ότι αυτή η πρακτική δε βελτιώνει καθόλου την απόδοση κανενός αλγορίθμου, τουλάχιστον στη συγκεκριμένη υλοποίηση, ωστόσο βελτιώνει ελάχιστα το χρόνο εκπαίδευσης του μοντέλου. Στη δική μας περίπτωση αυτή η βελτίωση κυμάνθηκε από 1 ως 2 s για τη δοκιμή του μοντέλου στα δεδομένα δοκιμής με τη μεγαλύτερη βελτίωση στην ταχύτητα να παρατηρείται στον MLP που είναι και ο πιο αργός εκ των τριών. Ο αλγόριθμος που εκτελεί τη δοκιμή του μοντέλου γρηγορότερα, εκ των τριών αυτών είναι ο C4.5.

Παρατηρώντας τις τιμές του δείκτη FP, που είναι η χαμηλότερη στην κλάση The Rolling Stones, στα πειράματα με τη μεγαλύτερη ακρίβεια, δηλαδή στη βάση δεδομένων MUSVOC με MLP αλγόριθμο, καταλήγουμε στο συμπέρασμα ότι κατηγοριοποιούνται αρκετά κομμάτια από άλλες κλάσεις σε αυτή την κλάση. Αυτό μπορεί να οφείλεται στην ιδιαίτερα μεγάλη ποικιλομορφία των μουσικών τους στιλ και στις διαφορές του ηχοχρώματος της παραγωγής μεταξύ των κομματιών της βάσης δεδομένων που δημιουργήσαμε. Αντιθέτως, στις κλάσεις Eagles και Def Leppard όπου

η ανάκληση (recall) είναι η μικρότερη στα συγκεκριμένα πειράματα, συμπεραίνουμε ότι ο αλγόριθμος δημιούργησε πολύ ειδικό μοντέλο πρόβλεψης.

Από τα πειράματα που περιγράφονται στον πίνακα 2 μπορούμε να βγάλουμε το συμπέρασμα ότι ο μέσος όρος της χαρακτηριστικών της συνολικής διάρκειας των ηχητικών αποσπασμάτων βελτιώνει την κατηγοριοποίηση σημαντικά σε σχέση με την πρακτική του υπολογισμού μέσου όρου χαρακτηριστικών σε πιο τακτά χρονικά διαστήματα.

Κεφάλαιο 6ο: Μελλοντική έρευνα/βελτιώσεις

Στην παρούσα πειραματική διαδικασία χρησιμοποιήθηκαν μουσικά έργα από 15 καλλιτέχνες και μέχρι 21 άλμπουμ ανά καλλιτέχνη. Σαφώς η βάση δεδομένων χρήζει επέκτασης ώστε να δημιουργηθούν και για άλλους καλλιτέχνες, αλλά και να ενισχυθούν τα ήδη υπάρχοντα μοντέλα με περισσότερα κομμάτια. Επίσης, θα ήταν θεμιτό να απορριφθούν ηχητικά αποσπάσματα από κοινά μουσικά κομμάτια προκειμένου να υπάρχει η δυνατότητα της χρήσης της πρακτικής της διασταυρούμενης επικύρωσης (cross-validation) (Kohavi, 1995) χωρίς το φόβο του υπερταιριάσματος (overfitting). Άλλη μία πρόταση για τη βελτίωση της βάσης θα ήταν οι κλάσεις να εμπεριείχαν κομμάτια προερχόμενα από τον ίδιο αριθμό άλμπουμ. Επιπλέον, το σύνολο δοκιμής δε θα έπρεπε να εμπεριείχε τόσο λίγα δεδομένα, αλλά για να μπορούσαμε να το επεκτείναμε με περισσότερα θα έπρεπε να εμπλουτίζαμε και το σύνολο δοκιμής.

Όσον αφορά τους αλγορίθμους κατηγοριοποίησης και κατ' επέκταση στον MLP ο οποίος απέδωσε καλύτερα έχουμε να προτείνουμε την εκτεταμένη έρευνα στη δομή και τον τύπο του εκάστοτε νευρωνικού δικτύου. Ιδιαίτερα δημοφιλείς τύποι νευρωνικών δικτύων είναι τα recurrent neural networks και convoluted neural networks. Θα θέλαμε να προτείνουμε τη χρήση του framework Tensorflow[17] ως αλγοριθμική προσέγγιση και ως υλικό τη χρήση της GPU για επιτάχυνση της πειραματικής διαδικασίας. Χαρακτηριστική αποτελεί η χρήση του Tensorflow για την κατηγοριοποίηση της βάση YouTube-8M (Zhong, et al., 2017), η οποία αποτελεί εν γένει μία πολύπλοκη εργασία κατηγοριοποίησης, αλλά η ταχύτητα και η ευελιξία του framework επιτρέπει την περάτωση της συγκεκριμένης εργασίας με ικανοποιητικά αποτελέσματα. Αυτές οι διαδικασίες αναμένεται να εφαρμοστούν κατά κόρον στο μέλλον για εργασίες όπως η παρούσα, καθώς διαφοροποιούνται από την εργασία δημιουργίας ηχητικού αποτυπώματος που συναντάται στη βιομηχανία και εντάσσονται σε μία πιο αφηρημένη εργασία που προσεγγίζει την αντίληψη της μουσικής από τον αθρώπινο εγκέφαλο (Warren, 2008).

[17] https://www.tensorflow.org/

Παράρτημα

Πίνακας 11: Χαρακτηριστικά κάποιων βάσεων δεδομένων που έχουν χρησιμοποιηθεί ή δημιουργηθεί από διάφορους ερευνητές

βάση ηχητικών δεδομένων	**κλάσεις**	**πλήθος κλάσεων**	**μορφή αρχείων**
(Zhang, 2003)	Andy Williams Elvis Presley Barbra Streisand Ella Fitzgerald, Liu, Huan, Liu, Wen-Zheng, Deng, Li-Jun, Meng, Ting-Wei	8	-
(Bartsch & Wakefield, 2004)	m01, m02, m03, m04, m05, s01, s02, s03, s04, s05, s06, s07	12	44100 Hz, μονοφωνικά
(Maddage, et al., 2004)	Bryan Adams, Michael Bolton, Eric Clapton, Shania Twain, Huang Pingyuan, Li Qi, Liu Ruoying, Clarance Wijewardana	8	44100 Hz, 16-bit, στερεοφωνικά (ποιότητα CD)
DB-S-2 (Tsai & Wang, 2006)	-	21	22050 Hz
DB-S-1 (Tsai & Wang, 2006)	-	20	22050 Hz
DB-S-1-T (Tsai & Wang, 2006)	-		22050 Hz
DB-S-1-E (Tsai & Wang, 2006)	-		22050 Hz
DB-D (Tsai & Wang, 2006)	-	-	22050 Hz
DB-I (Tsai & Wang, 2006)	-	-	22050 Hz
DB-ALBUM1 (Nwe & Li, 2007)	Michael Bolton, Richard Marx, Madonna	3	-
DB-ALBUM1 (Li & Wang, 2005), (Nwe & Li, 2007)	Michael Bolton, Richard Marx, Adu, Ou De Yang, Jay Chou, Kathryn Williams, Agnetha Faltskog, Jennifer Lopez, Shania Twain, Gabrielle, Madonna, Dido, Michael Bolton, Richard Marx, Adu, Ou De Yang, Jay Chou, Kathryn Williams, Agnetha Faltskog, Jennifer Lopez, Shania Twain, Gabrielle, Madonna, Dido	12	44100 Hz, 16-bit, στερεοφωνικά
TrainDB – SingerID (Nwe & Li, 2008)			-
DevelopmentDB – SingerID (Nwe & Li, 2008)			-
TestDB – SingerID (Nwe & Li, 2008)			-
(Li & Wang, Μάιος 2007)	-	-	16000 Hz, 16-bit
(Σοφιανός, et al., 2010)	Brian & D. Byrne, Sevara Nazarkhan, FeeIM, Radiohead, Nine Inch Nails, Los De Abajo, Anjelique Kidjo, Peter Gabrial, Brian & D. Byrne, Sevara Nazarkhan	-	-
(Sridhar & Geetha, 2008)	Alangudi Ranganathaiyer, M.S. Subbulaksumi K.J. Yesudas Nithyashree Mahadevan, Jayshree, Sowmya, …,	60	-
(Chanrungutai & Ratanamahatana, 2008)	-	10	16000 Hz, 16-bit
"artist20" (Shirali-Shahreza & Shirali-Shahreza, 2009)	-	20	-
"uspop2002" (Berenzweig, et al., 2004)	112, 311, 3_doors_down, 3lw, aaliyah, abba, paula_abdul, ac_dc, ace_of_base, bryan_adams, aerosmith, christina_aguilera, a_ha, air_supply, alice_in_chains, all_saints, alphaville, jessica_andrews, marc_anthony, fiona_apple, aqua, rick_astley, backstreet_boys, bad_brains, bad_company, bangles, barenaked_ladies, basement_jaxx, bbmak, beach_boys, beastie_boys, beatles, beck, bee_gees, lou_bega, pat_benatar, ben_folds_five, big_star, black_sabbath, blackstreet, blessid_union_of_souls, blind_melon, blink_182, blondie, bloodhound_gang, blood_sweat_tears, blur, bon_jovi, boston, david_bowie, toni_braxton, garth_brooks, busta_rhymes, cake, cardigans, mariah_carey, belinda_carlisle, aaron_carter, deana_carter,	400	22050 Hz, 128 kbps, μονοφωνικά, mp3

	nick_cave_and_the_bad_seeds, tracy_chapman, cheap_trick, chemical_brothers, cher, kenny_chesney, chic, chicago, chumbawamba, eric_clapton, clash, coal_chamber, joe_cocker, coldplay, collective_soul, phil_collins, coolio, alice_cooper, corrs, elvis_costello, counting_crows, cranberries, cream, creedence_clearwater_revival, christopher_cross, crowded_house, sheryl_crow, culture_beat, culture_club, cure, cypress_hill, billy_ray_cyrus, daft_punk, dangelo, craig_david, miles_davis, alice_deejay, deep_purple, def_leppard, deftones, john_denver, depeche_mode, neil_diamond, dido, ani_difranco, celine_dion, dire_straits, disturbed, dixie_chicks, dmx, doors, nick_drake, dr_dre, duran_duran, bob_dylan, eiffel_65, missy_elliot, enigma, en_vogue, enya, erasure, melissa_etheridge, eurythmics, evan_and_jaron, everclear, everlast, everly_brothers, everything_but_the_girl, extreme, lara_fabian, faith_no_more, fastball, fatboy_slim, filter, fine_young_cannibals, finger_eleven, fleetwood_mac, foo_fighters, foreigner, aretha_franklin, fuel, fugees, nelly_furtado, gabrielle, peter_gabriel, garbage, marvin_gaye, genesis, get_up_kids, kenny_g, goldfinger, goo_goo_dolls, gorillaz, al_green, green_day, pj_harvey, wade_hayes, heart, jimi_hendrix_experience, dru_hill, faith_hill, lauryn_hill, house_of_pain, whitney_houston, human_league, ice_cube, billy_idol, enrique_iglesias, natalie_imbruglia, incubus, inxs, iron_maiden, chris_isaak, alan_jackson, janet_jackson, michael_jackson, jamiroquai, ja_rule, wyclef_jean, joe, billy_joel, elton_john, olivia_newton_john, janis_joplin, montell_jordan, juvenile, kansas, kc_and_the_sunshine_band, toby_keith, kid_rock, kiss, mark_knopfler, korn, lenny_kravitz, la_bouche, cyndi_lauper, led_zeppelin, john_lennon, annie_lennox, les_rythmes_digitales, huey_lewis_and_the_news, lfo, lifehouse, lil_bow_wow, limp_bizkit, linkin_park, live, ll_cool_j, kenny_loggins, lucy_pearl, ludacris, luniz, lynyrd_skynyrd, madison_avenue, madonna, marilyn_manson, bob_marley, ricky_martin, richard_marx, dave_matthews_band, edwin_mccain, tim_mcgraw, sarah_mclachlan, don_mclean, me_first_and_the_gimme_gimmes, melanie_c, men_at_work, metallica, george_michael, bette_midler, steve_miller_band, milli_vanilli, moby, moody_blues, morcheeba, debelah_morgan, alanis_morissette, van_morrison, mr_big, mudvayne, samantha_mumba, muse, mxpx, mya, mystikal, nazareth, nelly, new_found_glory, new_order, new_radicals, next, stevie_nicks, nine_days, nine_inch_nails, nirvana, no_doubt, nsync, oasis, offspring, mike_oldfield, oleander, roy_orbison, orgy, ozzy_osbourne, o_town, our_lady_peace, outkast, jennifer_paige, robert_palmer, papa_roach, pennywise, pet_shop_boys, tom_petty, liz_phair, pink, pink_floyd, placebo, poison, police, portishead, presidents_of_the_united_states_of_america, elvis_presley, pretenders, prince, procol_harum, prodigy, propellerheads, queen, queensryche, radiohead, gerry_rafferty, rage_against_the_machine, rammstein, rancid, rednex, rem, reo_speedwagon, lionel_richie, leann_rimes, r_kelly, rolling_stones, roxette, run_dmc, sade, santana, savage_garden, s_club_7, scorpions, seal, neil_sedaka, selena, semisonic, seven_mary_three, ron_sexsmith, shaggy, simon_and_garfunkel, carly_simon, paul_simon, simple_minds, jessica_simpson, frank_sinatra, sisqo, sixpence_none_the_richer, skid_row, sly_and_the_family_stone, smashing_pumpkins, smash_mouth, sneaker_pimps, soft_cell, soul_asylum, soundgarden, spandau_ballet, britney_spears, spice_girls, spin_doctors, spineshank, bruce_springsteen, staind, steppenwolf, stereophonics, cat_stevens, rod_stewart, sting, stone_temple_pilots, stroke_9, styx, sublime, sugar, sugar_ray, donna_summer, supertramp, survivor, keith_sweat, matthew_sweet, talking_heads, tears_for_fears, temple_of_the_dog, tesla, texas, third_eye_blind, thompson_twins, tlc, tonic, tool, toto, tricky, tina_turner, shania_twain, twista, bonnie_tyler, u2, ub40, ugly_kid_joe, uriah_heep, usher, paul_van_dyk, van_halen, vanilla_ice, bobby_vee, velvet_underground, vengaboys, vertical_horizon, verve, violent_femmes, war, weezer, westlife, wham, wheatus, whiskeytown, barry_white, white_zombie, brian_wilson, wilson_phillips, steve_winwood, stevie_wonder, gary_wright, wu_tang_clan, xzibit, neil_young, zz_top		
CAL500 Expansion (CAL500exp) (Wang, et al., 2014)	–	67	22050 Hz, μονοφωνικά, mp3

Βιβλιογραφία

Abdulla, W. H., 2002. *Auditory Based Feature Vectors for Speech Recognition Systems.* s.l., WSEAS Press, pp. 231-236.

Aertsen, A. M. H. J., Johannesma, P. I. M. & Hermes, D. J., 1980. *Spectro-Temporal Receptive Fields of Auditory Neurons in the Grassfrog.* s.l.:s.n.

Akeroyd, M. A., Moore, B. C. J. & Moore, G. A., 2001. *Melody recognition using three types of dichotic-pitch stimulus.* s.l.:J. Acoust. Soc. Amer..

Anon., n.d. *Vocal.* [Ηλεκτρονικό]
Available at: https://www.vocal.com/noise-reduction/perceptual-noise-reduction/
[Πρόσβαση 16 Μάιος 2020].

Atal, S. B., 1974. *Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification..* s.l.:The Journal of the Acoustical Society of America.

Atame, S., Therese, P. S. S. & Gedam, P. M., 2015. *A Survey On: Continuous Voice Recognition Techniques.* s.l.:International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).

Bartsch, M. & Wakefield, G., 2004. *Singing Voice Identification Using Spectral Envelope Estimation.* s.l.:IEEE Transactions, Speech and Audio Processing.

Berenzweig, A. & Ellis, D., 2001. *Locating singing voice segments within music signals.* s.l.:2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics.

Berenzweig, A., Ellis, D. & Lawrence, S., 2002. *Using voice segments to improve artist classification of music.* Espoo, Φινλανδία: AES 22nd Int. Conf..

Berenzweig, A., Logan, B., Ellis, D. & Whitman, B., 2004. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal,* Ιούνιος, Issue 28 (14pp), pp. 63-76.

Bouckaert, R. R. και συν., 2013. *WEKA Manual for Version 3-7-8.* Hamilton(New Zealand): s.n.

Bregman, A. S., 1990. *Auditory Scene Analysis.* Cambridge(MA): MIT Press.

Brown, G. J. & Wang, D. L., n.d. *Separation of speech by computational auditory scene analysis.* New York: Springer.

Burgess, C. J. C., 1998. *A tutorial on support vector machines for pattern recognition..* s.l.:Data Mining and Knowledge Discovery.

Burred, J. J. & Sikora, T., 2007. *Monaural source separation from musical mixtures based on time-frequency timbre models.* s.l.:Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007).

Byrd, D. & Crawford, T., 2002. *Problems of music information retrieval in the real world.* s.l.:Inform, Process, Manage.

Cai, W., Li, Q. & Guan, X., 2011. *Automatic singer identification based on auditory features.* Shanghai, s.n., pp. 1624-1628.

Campbell Jr, J. P., 1997. *Speaker recognition: A tutorial.* s.l.:Proc. IEEE.

Cano, P., Koppenberger, M. & Wack, N., 2005. *Content-based music audio recommendation.* Singapore: Proceedings of the 13th Annual ACM international Conference on Multimedia (MULTIMEDIA '05), Hilton.

Chang, P., 2009. *Pitch Oriented Automatic Singer Identification in Pop Music.* s.l., s.n., pp. 161-166.

Chanrungutai, A. & Ratanamahatana, C. A., 2008. *Singing Voice Separation in Mono-Channel Music.* s.l., IEEE Xplore.

Christianini, N. & Shawe-Taylor, J., 2000. *An introduction to support Vector Machines: and other kernel-based learning methods..* Νέα Υόρκη(NY): Cambridge University Press.

D. Godsmark, D. & Brown, G. J., 1999. *A blackborad architecture for computational auditory scene analysis.* s.l.:Speech Commun..

Davis, S. B. & Mermelstein, P., 1980. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* s.l.:IEEE Trans. Acoust., Speech, Signal Processing.

Deshmukh, S. & Bhirud, S. G., 2012. *A Hybrid Selection Method of Audio Descriptors for Singer Identification in North Indian Classical Music.* s.l.:IEEE.

Deshmukh, S. & Bhirud, S. G., 2014. A Novel Method to Identify Audio Descriptors, Useful in Gender Identification from North Indian Classical Music Vocal. 5(2), pp. 1139-1143.

Divenyi, P., επιμ., 2005. *Speech Separation by Humans and Machines.* Norwell(MA): Kluwer Academic.

Durey, A. S. & Clements, M. A., 2002. *Features for melody spotting using hidden Markov models.* Orlando(FL): s.n.

Ellis, D., Berenzweig, A. & Whitman, B., 2003. *The "uspop2002" Pop Music data set.* s.l., s.n.

Ellis, D. P. W., 2007. *Classifying Music Audio with Timbral and Chroma Features.* s.l.:Austrian Computer Society (OCG).

Erickson, M. L., Handel, S. & Perry, S., 2001. *Discrimination functions: Can they be used to classify singing voices.* s.l.:J. Voice.

Eronen, A., 2003. *Musical instrument recognition using ICA-based transform of featyres and discriminatively trained HMMs.* Παρίσι: Proc. 7th Int. Symp. Signal Processing Applications.

Esmaili, S., Krishnan, S. & Raahemifar, K., Μάιος 2004. *Content based audio classification and retrieval using joint time – frequency analysis.* Montreal, Καναδάς: IEEE International Conference on Acoustics, Speech, and Signal Processing.

Ezzaidi, H., Bahoura, M. & Rouat, J., 2010. *Singer and music discrimination based threshold in polyphonic music.* Λούξορ, DBLP.

Freund, Y. & Schapire, R. E., 1997. *A decision-theoretic generalization of on-line learning and an application to boosting..* s.l.:J. Comput. Syst. Sci..

Fujihara, H. & Goto, M., 2010. *A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre similarity-based music information retrieval.* s.l., s.n., pp. 638-648.

Fujihara, H. και συν., 2005. *Singer identification based on accompaniment sound reduction and reliable frame selection.* s.l.:Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005).

Furui, S., 1997. *Recent advances in speaker recognition.* s.l.:Pattern Recognit. Lett..

Goto, M., 2004. *A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals.* s.l.:Speech Commun..

Goto, M., Saitou, T., Nakano, T. & Fujihara, H., 2010. *Singing information processing based on singing voice modeling.* s.l.:IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP).

Gouyon, F. & Herrera, P., 2001. *Exploration of techniques for automatic labeling of audio drum tracks instruments.* s.l., s.n.

Grgic, M., Delac, K. & Ghanbari, M., 2009. *Recent Advances in Multimedia Signal Processing and Communications.* Berlin(Heidelberg): Springer-Verlag.

Handel, S. & Erickson, M. L., 2001. *A rule of thumb: The bandwidth for timbre invariance is one octave.* s.l.:Music Percept.

Herrera, P., Amatriain, X., Batlle, E. & Serra, X., 2000. *Toward instrument segmentation for music content description: A critical review of instrument classification techniques.* Plymouth(MA): Proc. 1st Int. Symp. Music Information Retrieval.

Holzapfel, A. & Στυλιανού, Γ., n.d. *Singer Identification in Rembetiko Music.* Κρήτη: Institute of Computer Science, FORTH, and Multimedia Informatics Laboratory, Dep. of Computer Science, University of Crete.

Hsu, J. L., Liu, C. C. & Chen, A. L. P., 2001. *Discovering nontrivial repeating patters in music data.* s.l.:IEEE Trans. Multimedia.

Hu, G. & Wang, D. L., Σεπ. 2004. *Monaural speech segregation based on pitch tracking and amplitude modulation.* s.l.:IEEE Trans. Neural Netw..

Hunt, E. B., Marin, J. & Stone, P. J., 1966. *Experiments in induction.* Νέα Υόρκη: Academic Press.

Hu, Y. & Liu, G., 2013. *Automatic singer identification using missing feature methods.* s.l.:'ICME' , IEEE Computer Society.

Imai, S. & Abe, Y., 1979. *Spectral envelope extraction by improved cepstral method.* s.l.:Electron. and Commun. in Japan.

Jeong, I.-Y. & Lee, K., 2014. *Vocal Separation from Monaural Music Using Temporal/Spectral Continuity and Sparsity Constraints.* s.l., s.n., pp. 1197-1200.

Jiang, D.-N.και συν., n.d. *Music Type Classification By Spectral Contrast Feature.* s.l., Department of Computer Science and Technology, Tsinghua University.

Karaman, E., 2009. *A Comparison of Different Classification Systems for Automatic Singer Identification.* s.l.:İZMİR.

Kelly, T., n.d. *Music Artist Identification Using Linear Temporal Pyramid Matching.* s.l.:s.n.

Kim, Y. E. & Whitman, B., 2002. *Singer identification in popular music recordings using voice coding features.* Παρίσι, Γαλλία: Proc. ISMIR 2002: 3rd Int. Conf. Music Information Retrieval.

Kim, Y., Williamson, D. & Pilli, S., 2006. *Understanding and Quantifying the "Album Effect" in Artist Identification.* Βικτώρια(CA): s.n.

Kohavi, R., 1995. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* s.l.:Appears in the International Joint Conference on Artificial Intelligence (IJCAI).

Kopparapu, S. K. & Laxminarayana, M., 2010. *Choice of Mel filter bank in computing MFCC of a resampled speech.* Κουάλα Λουμπούρ: IEEE.

Kroher, N. & Gómez, E., 2014. *Automatic singer identification for improvisational styles based on vibrato, timbre and statistical performance descriptors.* s.l.:Ann Arbor, MI: Michigan Publishing, University of Michigan Library.

Kumar, R. C. P. & Suguna, S., 2015. *Analysis of Mel Based Features for Audio Retrieval.* s.l.:ARPN Journal of Engineering and Applied Sciences.

Lagrange, M., Ozerov, A. & Vincent, E., 2012. *Robust Singer Identification in Polyphonic Music Using Melody Enhancement and Uncertainty-Based Learning.* Porto, s.n., pp. 565-600.

Leinweber, D., 2007. *Stupid Data Miner Tricks: Overfitting the S&P 500.* s.l.:s.n.

Li, S. Z., 2000. *Content-based audio classification and retrieval using the nearest feature line method.* s.l.:IEEE Trans. Speech Audio Processing.

Liu, C.-C. & Huang, C.-S., 2002. *A singer identification technique for content-based classification of MP3 music objects.* McLean(VA): Proc. Conf. Information and Knowledge Management.

Li, Y. & Wang, D. L., 2005. *Detecting pitch of singing voice in polyphonic audio.* s.l.:Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.

Li, Y. & Wang, D., Μάιος 2007. *Separation of Singing Voice From Music Accompaniment for Monaural Recordings.* s.l.:IEEE Transactions On Audio, Speech, and Language Processing.

Lu, L., Zhang, H. I. & Li, S. Z., Απρ. 2003. *Content based audio classification and segmentation by using support vector machines.* s.l.:Multimedia Systems 8 (6).

Maazouri, F. & Bahi, H., n.d. *Singing Voice Classification in Commercial Music.* s.l., s.n.

Maddage, N. C., Xu, C., Kankanhalli, M. S. & Shaoin, X., 2004. *Content-based music structure analysis with applications to music semantics understanding.* s.l.:Proc. 12th Annu. ACM Int. Conf. Multimedia.

Maddage, N. C., Xu, C., Kankanhalli, M. S. & Shao, X., 2004. *Content-Based Music Structure Analysis with Applications to Music Semantics Understanding.* s.l.:Proceedings of 12th Annual ACM International Conference on Multimedia.

Maddage, N. C., Xu, C. & Wang, Y., 2004. *Singer identification based on vocal and instrumental models.* s.l.:Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference.

Mammone, R., Zhang, X. & Ramachandran, R. P., 1996. *Robust speaker recognition: A feature based approach.* s.l.:Signal Process Mag..

Mandel, M. I. & Ellis, D. P. W., 2005. *Song-level features and support vector machines for music classification.* Λονδίνο: Proceedings of International Conference on Music Information Retrieval (ISMIR 2005).

Mandel, M. I., Poliner, G. E. & and D. P. W. Ellis, D. P. W., 2006. *Support vector machine active learning for music retrieval.* s.l.:Multimedia systems.

Mandel, M. I., Poliner, G. E. & Ellis, D. P. W., 2006. *Support vector machine active learning for music retrieval.* s.l.:Springer-Verlag.

Martin, K. D. & Kim, K. Y. E., 1998. *Musical instrument identification: A pattern-recognition approach.* s.l., s.n.

Martin, K., Screirer, E. & Vercoe, B., 1998. *Music content analysis through models of audition.* Bristol, Ηνωμένο Βασίλειο: Proc. ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications.

McHugh, M. L., n.d. *Interrater reliability: the kappa statistic.* Ζάγκρεμπ: Biochem Med.

Mellinger, D. K., 1991. *Event formation and separation in musical sound.* Stanford(CA): Ph.D. dissertation, Dept. Comput. Sci., Stanford University.

Mellody, M., 2001. *Signal Analysis of the Female Singing Voice: Features for Perceptual Singer Identity.* Ann Arbor: Ph.D. dissertation.

Mellody, M., Herseth, F. & Wakefield, G. H., 2001. *Modal distribution analysis, synthesis, and perception of a soprano's sung vowels.* s.l.:J. Voice.

Mellody, M. & Wakefield, G. H., 2000. *Signal analysis of the singing voice: Low-order representations of singer identity.* Βερολίνο, Γερμανία: Proc. Int. Computer Music Conf. 2000.

Meron, Y. & Hirose, K., 1998. *Separation of singing and piano sounds.* s.l.:Proc. 5th Int. Conf. Spoken Lang. Process. (ICSLP 98).

Mesaros, A., Virtanen, T. & Klapuri, A., 2007. *Singer identification in polyphonic music using vocal separation and pattern recognition methods.* s.l.:ISMIR.

Monzo, J. L., 1998. *JustMusic: A New Harmony - Representing Pitch as Prime Series..* s.l.:Joseph L. Monzo.

Nikam, H. W., 5 Δεκ. 2013. *Evaluating Audio Descriptors For Timbre Analysis In Singer Identification Process and Information Technology Research (IJCSEITR).* s.l.:International Journal of Computer Science Engineering.

Nwe, T. L. & Li, H., 2007. *Exploring vibrato-motivated acoustic features for singer identification.* s.l.:IEEE Trans. Audio, Speech, and Language Processing.

Nwe, T. L. & Li, H., 2008. *On Fusion of Timbre-Motivated Features for Singing Voice Detection and Singer Identification.* Las Vegas, s.n., pp. 2225-2228.

Nwe, T. L. & Wang, Y., 2004. *Automatic Detection of Vocal Segments in Popular Songs.* s.l.:Proceedings of 5th International Conference on Music Information Retrieval (ISMIR).

Ozerov, A., Philippe, P., Gribonval, R. & Bimbot, F., Οκτ. 2005. *One microphone singing voice separation using source-adapted models.* s.l.:IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.

Patil, H. A., Radadia, P. & Basu, T. K., n.d. *Combining Evidences from Mel Cepstral Features and Cepstral Mean Subtracted Features for Singer Identification.* s.l.:Asian Language Processing (IALP), 2012 International Conference.

Peeters, G., 2004. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project.* s.l.:CUIDADO I.S.T. Project Report.

Powers, D. M. W., 2007. *Evaluation: From Precision, Recall and F-Factor.* Αδελαΐδα: School of Informatics and Engineering, Flinders University, .

Quinlan, J. R., 1979. *Discovering rules by induction from large collections of examples..* Εδινβούργο: Edinburgh University Press.

Quinlan, J. R., 1993. *C4.5: Programs for machine learning.* San Mateo: Morgan Kaufman Publishers.

Rao, V., Ramakrishnan, S. & Rao, P., 2009. *Singing voice detection in polyphonic music using predominant pitch.* s.l.:INTERSPEECH ’09.

Regnier, L. & Peeters, G., 2011. *Combining Classifications Based on Local and Global Features: Application to Singer Identification.* s.l.:Audio Effects, Digital Audio.

Reynolds, D. A. & Rose, R. C., 1995. *Robust text-independent speaker identification using Gaussian mixture speaker models.* s.l.:IEEE Trans. Speech Audio Process.

Röbel, A., 2005. *Efficients spectral envelope estimation and its application to pitch shifting and envelope preservation.* s.l.:DAF’x.

Rocamora, M. & Herrera, P., n.d. *Comparing audio descriptors for singing voice detection in music audio files.* s.l.:Comisión Sectorial de Investigación Científica, UdelaR.

Rosenthal, D. F. & Okuno, H. G. επιμ., 1998. *Computational Auditory Scene Analysis.* Mahwah(NJ): s.n.

Scheirer, E. & Slaney, M., 1997. *Construction and evaluation of a robust multifeature speech/music discriminator.* Μόναχο, Γερμανία: Proc. ICASSP.

Shen, J., Shepherd, J., Cui, B. & Tan, K., 2006. *HSI: A Novel Framework for Efficient Automated Singer Identification in Large Music Databases..* Atlanta: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06).

Shikano, K., 1986. *Evaluation of LPC spectral matching measures for phonetic unit recognition..* s.l.:Technical Report CMU-CS-96-108, CMU, Computer Science Department.

Shirali-Shahreza, S. & Shirali-Shahreza, M. H., 2009. *Fast and Scalable System For Automatic Artist Identification.* s.l.:IEEE.

Shixiong, C., Qin, G. & Huijun, J., 2008. *Gammatone filter bank to simulate the characteristics of the human basilar membrane.* s.l.:J. Tsinghua Univ (Sci & Tech).

Shruti & Chhabra, B., Νοέμβριος 2015. *An Approach for Singer Identification Technique Using Artificial Neural Network.* s.l.:International Journal of Advanced Research in Computer Science and Software Engineering.

Sridhar, R. & Geetha, T. V., 2008. *Music Information Retrieval Of Carnatic Songs Based On Carnatic Music Singer Identification.* s.l.:IEEE.

Stevens, S. S., 1975. *Psychophysics.* s.l., Transaction Publishers.

Streich, S., 2007. *Music complexity: a multi-faceted description of audio content.* Βαρκελώνη(UPF): s.n.

Sundberg, J., 1987. *The Science of the Singing Voice..* DeKalb: Northern Illinois Univ. Press.

Titze, I. R., 1994. *Principles of Voice Production.* Englewood Cliffs(N.J.): Prentice Hall.

Tsai, W. H. & Lee, H.-C., 2012. *Automatic Singer Identification Based on Speech-Derived Models.* Hong Kong: Proc. International Conference on Advancements in Information Technology.

Tsai, W. H. & Lin, H. P., 2011. *Background Music Removal Based on Cepstrum Transformation for Popular Singer Identification.* s.l.:IEEE Trans. Audio, Speech, Lang. Process.

Tsai, W. H., Wang, H. W. & Rodgers, D., 2003. *Automatic singer identification of popular music recordings via estimation and modeling of solo vocal signal.* Γενεύη: Proc. 8th Eur. Conf. Speech Communication and Technology.

Tsai, W.-H. & Lin, H.-P., Ιούλιος 2010. *Popular singer identification based on cepstrum transformation.* s.l.:IEEE International Conference on Multimedia and Expo (ICME), 2010.

Tsai, W.-H. & Wang, H. m., 2006. *Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals.* s.l.:IEEE Transaction on Speech and Audio Processing, to appear.

Turk, M., 1991. Engenfaces for recognition. *Journal of cognitive neuroscience,* 3(1), pp. 71-86.

Turnbull, D., Barrington, L., Torres, D. & Lanckriet, G., 2008. *Semantic annotation and retrieval of music and sound effects.* s.l.:IEEE Transactions.

Tzanetakis, G., n.d. *Manipulation, Analysis and Retrieval system for audio signals.* s.l.:PhD Thesis..

Vapnik, V., 1998. *Statistical learning theory.* s.l.:Wiley.

Wang, A. L.-C., 1994. *Instantaneous and frequency-warped signal processing techniques for auditory source separation.* Stanford, CA: Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ..

Wang, S.-Y., Wang, J.-C., Yang, Y.-H. & Wang, H.-M., 2014. *Towards Time-varying Music Auto-tagging Based on CAL500 Expansion.* Βαρκελώνη: IEEE International Conference on Multimedia and Expo.

Warren, J., 2008. *How does the brain process music?.* s.l.:s.n.

Whitman, B., Flake, G. & Lawrence, S., 2001. *Artist detection in music with minnowmatch.* Falmouth(MA): Proc. 2001 IEEE Workshop on Neural Networks for Signal Processing.

Wold, E., T., B., Keislar, D. & Wheaton, J., 1996. *Content-based classification, search, and retrieval of audio.* s.l.:IEEE Multimedia.

Wu, X. και συν., 2007. *Top 10 algorithms in data mining.* s.l.:Springer-Verlag London Limited.

Xu, C. και συν., 2003. *Musical genre classification using support vector machines.* Hong Kong: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing.

Y.E. Kim, D. W. a. S. P., 2006. *Towards quantifying the album effect in artist identification.* s.l.:ISMIR. Citeseer.

Zhang, T., 2003. *Authomatic singer identification.* s.l.:ICME.

Zhang, T., Ιούλιος 2003. *System and method for automatic singer identification.* s.l.:IEEE Int. Conf. Multimedia and Expo..

Zhong, Z. και συν., 2017. *An Effective Way to Improve YouTube-8M Classification Accuracy in Google Cloud Platform.* s.l.:s.n.

Παναγιώτου, Β. & Μητιανούδης, Ν., n.d. *PCA Summarization for Audio Song Identification.* Ξάνθη: Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Δημοκρίτειο Πανεπιστήμιο Θράκης.

Σοφιανός, Σ., Ariyaeeinia, A. & Polfreman, R., 2010. *Towards effective singing voice extraction from stereophonic recordings.* Ντάλας(TX): IEEE.

Τζανετάκης, Γ. & Cook, P., 2002. *Musical genre classification of audio signals.* s.l.:IEEE Trans. Speech Audio Processing.