View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Thessaly Institutional Repository

Πανεπιστήμιο Θεσσαλίας

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Ανίχνευση Ήχων Πτηνών

Bird Audio Detection

Διπλωματική Εργασία από

Παπαγάτσια Μαρία

Παραδίδεται στο
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
για την ολοκλήρωση των απαιτήσεων του

Διπλώματος

του

Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών

Βόλος, Θεσσαλία, Οκτώβριος, 2019

Πανεπιστήμιο Θεσσαλίας

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τα μέλη της επιτροπής, πιστοποιούν ότι έχουν διαβάσει τη διπλωματική που παρουσίασε η Παπαγάτσια Μαρία και ότι είναι απόλυτα επαρκής ως προς το πεδίο εφαρμογής και την ποιότητα ως μερική απαίτηση για το δίπλωμα του Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών.

Επιβλέπων
Ποταμιάνος Γεράσιμος
Αναπληρωτής Καθηγητής

Συνεπιβλέπων
Μπέλλας Νικόλαος
Αναπληρωτής Καθηγητής

Συνεπιβλέπων
Κοράκης Αθανάσιος
Αναπληρωτής Καθηγητής

Βόλος, Θεσσαλία, Οκτώβριος, 2019

Δήλωση Συγγραφής

Εγώ, η Παπαγάτσια Μαρία, δηλώνω ότι αυτή η διπλωματική με τίτλο, 'Ανίχνευση Ήχων Πτηνών' και το έργο που παρουσιάζεται σε αυτή είναι δική μου. Βεβαιώνω ότι:

- Αυτό το έργο έγινε εξ ολοκλήρου ή κυρίως κατά την υποψηφιότητα για πτυχίο σε αυτό το Πανεπιστήμιο.
- Όπου οποιοδήποτε μέρος της παρούσας διπλωματικής έχει προηγουμένως υποβληθεί για πτυχίο ή οποιοδήποτε άλλο προσόν σε αυτό το Πανεπιστήμιο ή σε οποιοδήποτε άλλο ίδρυμα, αυτό έχει δηλωθεί με σαφήνεια.
- Όπου έχω συμβουλευθεί τη δημοσιευμένη δουλειά των άλλων, αυτό πάντα αποδίδεται σαφώς.
- Όπου έχω αναφέρει το έργο άλλων, η πηγή δίνεται πάντοτε. Με εξαίρεση τέτοιες παραθέσεις, αυτή η διπλωματική είναι εξ ολοκλήρου δική μου δουλειά.
- Έχω αναγνωρίσει όλες τις κύριες πηγές βοήθειας.
- Όπου η εργασία βασίζεται σε δουλειά που πραγματοποίησα από κοινού με άλλους, έχω καταστήσει σαφές τι ακριβώς έπραξαν οι άλλοι και αυτό που έχω συνεισφέρει η ίδια.

Παπαγάτσια Μαρία
Βόλος, Θεσσαλία, Οκτώβριος, 2019

©2019 by Παπαγάτσια Μαρία
All Rights Reserved

Ευχαριστίες

Αρχικά θα ήθελα να εκφράσω την ευγνωμοσύνη μου στον καθηγητή κ. Ποταμιάνο Γεράσιμο για την επίβλεψη αυτής της διπλωματικής εργασίας και για όλη τη βοήθεια που μου προσέφερε.

Επιπλέον, θα ήθελα να ευχαριστήσω όλους τους καθηγητές του τμήματος για την καθοδήγηση και τις γνώσεις που μου χάρισαν αυτά τα 5 χρόνια των σπουδών μου.

Στη συνέχεια, θα ήθελα να ευχαριστήσω τους αγαπημένους μου φίλους, και ιδιαίτερα τον Μιχαήλ, για την υποστήριξη και τις πολύτιμες συμβουλές τους κατά τη διάρκεια αυτής της διπλωματικής εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα τους γονείς μου και τον αδερφό μου για την υλική, ηθική συμπαράσταση και την αγάπη που μου προσέφεραν όλα αυτά τα χρόνια.

Ανίχνευση Ήχων Πτηνών

της

Παπαγάτσια Μαρία

Περίληψη

Η ακουστική παρακολούθηση των πουλιών είναι ένα εξαιρετικά σημαντικό κομμάτι για πολλά περιβαλλοντολογικά προγράμματα. Το έργο που καθορίζει την παρουσία ή απουσία ενός πουλιού σε ένα αρχείο ήχου ονομάζεται Ανίχνευση Ήχων Πουλιών (Bird Audio Detection - BAD). Η δυσκολία του BAD έγκειται στην αυξημένη παρουσία θορύβου λόγω περιβαλλοντολογικών συνθηκών (αέρας, βροχή, ήχοι ζώων) και στη γενίκευση των μοντέλων σε πολλαπλές συνθήκες ηχογράφησης. Στην παρούσα διπλωματική προτείνουμε την εξερεύνηση διαφόρων τύπων ακουστικών χαρακτηριστικών στοχεύοντας στην αποδοτικότερη αναπαράσταση μελωδικών ήχων όπως αυτούς των πουλιών. Ως αποτέλεσμα, επιλέγουμε να εργαστούμε με χαρακτηριστικά της κατηγορίας Chroma τα οποία παρουσιάζουν ανθεκτικότητα στις αλλαγές του ρυθμού και συνδέονται άμεσα με τη μουσική πτυχή της αρμονίας. Πιο συγκεκριμένα, τα τρία είδη χαρακτηριστικών που δοκιμάστηκαν βασίζονται στο μετασχηματισμό Fourier, το μετασχηματισμό Constant-Q και τη στατιστική ανάλυση της κατανομής ενέργειας στις ζώνες χρώματος. Για την αντιμετώπιση του θορύβου εφαρμόστηκε ένα υψιπερατό φίλτρο στο αρχικό σήμα όπως επίσης και κανονικοποίηση και αντιστοίχιση στην κανονική κατανομή στα διανύσματα χαρακτηριστικών. Οι ταξινομητές που χρησιμοποιήθηκαν είναι τα Τυχαία Δέντρα Αποφάσεων (Random forests) και οι Μηχανές Διανυσματικής Στήριξης (Support Vector Machines - SVMs). Η ευρωστία των χαρακτηριστικών Chroma CENS και Chroma CQT αναδείχθηκε μετά την εφαρμογή του φίλτρου. Συγκεκριμένα, για τον ταξινομητή SVM λάβαμε 98.56% στη μετρική Area Under Curve (AUC) για τα χαρακτηριστικά Chroma CENS και 96.99% για τα Chroma CQT.

Bird Audio Detection

by

Papagatsia Maria

Abstract

Acoustic monitoring of birds is a very important part of many environmental projects. The task that determines the presence or absence of a bird in an audio file is referred to as Bird Audio Detection (BAD). The difficulty of BAD lies in the increased presence of noise due to environmental conditions (wind, rain, animal sounds) and the generalization of the models to multiple recording conditions. In this thesis we explore various types of features aiming at the most efficient representation of the melodic sounds of birds. As a result of this, we choose to work with features of the Chroma class that are resistant to timbre changes and are directly linked to the musical aspect of harmony. Especially, the three types of Chroma features that were tested are Chroma features based on Fourier transform, Chroma features generated by the Constant-Q transform, and statistical analysis of energy distribution in Chroma bands. Concerning noise reduction, a high-pass filter was applied to the original signal as well as normalization and mapping to the normal distribution in the feature vectors. Random forests and Support Vector Machines (SVMs) classifiers were used in this study. The robustness of the Chroma CENS and Chroma CQT features was demonstrated after the application of the filter. Specifically, for the SVM classifier we obtained 98.56% Area Under Curve (AUC) for the Chroma CENS features and 96.99% for Chroma CQT.

Κατάλογος συντομογραφιών

AUC Area Under the ROC. 4, 5, 19, 20

BAD Bird Audio Detection, Ανίχνευση Ήχων Πουλιών. 1, 3

CENS Chroma Energy Normalized. 11

CNN Convolutional Neural Network. 4

CQT Constant-Q Transform. 10

DCASE Detection and Classification of Acoustic Sound Events. 1–3, 5

DNN Deep Neural Network. 2

GMM Gaussian Mixture Model. 2

HMM Hidden Markov Model. 2

MFCC Mel-frequency Cepstral Coefficients. 2, 4, 7, 8

ROC Receiver Operating Characteristic. 19

STFT Short-Time Fourier Transform. 9

SVM Support Vector Machine. 2, 4, 21

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Αναγνώριση Ήχων Πουλιών

Η ανίχνευση ήχων πουλιών είναι ένα έργο μεγάλης σημασίας για την παρακολούθηση του περιβάλλοντος και της βιοποικιλότητας. Οι κλιματικές αλλαγές και οι αλλαγές στη διαχείριση της γης οδήγησαν σε απότομη πτώση των πληθυσμών των πτηνών, και τα επόμενα χρόνια η κατανομή και ο αριθμός τους θα συνεχίσει να υποφέρει. Η παραδοσιακή παρακολούθηση των πτηνών αφορά χειρωνακτικές μεθόδους, συνήθως με συμμετοχή εθελοντών, και γίνεται πρακτικά αδύνατη αν λάβουμε υπόψιν μας τις μορφολογικά δύσκολες περιοχές. Επιπλέον, θεωρείται ότι πολλά είδη πτηνών εντοπίζονται ευκολότερα από τον ήχο παρά από την εικόνα τους ενισχύοντας έτσι τη συμβολή των σύγχρονων συσκευών ηχογράφησης στη διαδικασία της παρακολούθησης [1],[5]. Προκειμένου να ενθαρρυνθεί η αυτοματοποίηση όχι μόνο της εγγραφής αλλά και της φάσης ανίχνευσης, δημιουργήθηκε το έργο της Ανίχνευσης Ήχων Πουλιών (Bird Audio Detection - BAD) [6]. Σχεδιάστηκε για να εμπνεύσει τους συμμετέχοντες να χρησιμοποιήσουν τις τελευταίες τεχνολογικές μεθόδους για να επιλύσουν το πρόβλημα με κύριο στόχο τη γενίκευση σε νέες συνθήκες.

Το αντικείμενο του έργου είναι να προσδιοριστεί η αρχή και το τέλος της κλήσης πουλιών καθώς και τα είδη τους. Αρχικά, είναι απαραίτητο να δημιουργηθεί ένα προκαταρκτικό σύστημα που να μπορεί να διακρίνει απλώς την παρουσία και την απουσία κλήσεων πουλιών σε ένα αρχείο ήχου. Επιπρόσθετα, το μοντέλο θα πρέπει να είναι σε θέση να χειρίζεται το θόρυβο που μπορεί να προκύψει από ομιλία, μηχανήματα και φυσικά φαινόμενα (όπως ο άνεμος και η βροχή) καθώς και να εντοπίζει διαφορετικών ειδών κλήσεις πουλιών. Αυτός ο παρασκηνιακός θόρυβος είναι πολύ συνηθισμένος στις ηχογραφήσεις εξωτερικών χώρων και πολύ εύκολα μπορεί να συγχέεται με κλήσεις πτηνών χαμηλής έντασης [7],[8]. Οι προαναφερθείσες προκλήσεις παρεμβαίνουν στην ισχυρή απόδοση των προσεγγίσεων της Μηχανικής και Βαθιάς Μάθησης, στις οποίες τα δεδομένα δοκιμών (test set) και αξιολόγησης (evaluation set) αντλούνται από την ίδια κατανομή με τα δεδομένα εκπαίδευσης (train set) [1]. Για το λόγο αυτό, πραγματοποιήθηκαν αρκετοί διαγωνισμοί ανίχνευσης ήχου, όπως το LifeCLEF Bird Identification Task [9] και ο διαγωνισμός Bird Audio Detection του DCASE [5], με έμφαση στην επίτευξη υψηλών επιδόσεων σε συνθήκες με άγνωστα είδη πτηνών και ακουστικά περιβάλλοντα. Το DCASE παρέχει ένα ειδικό θέμα για την Ανίχνευση Ήχων Πουλιών, το οποίο διεξήχθη δύο χρόνια, και είναι το θέμα που παρουσιάζεται σε αυτή τη διπλωματική.

1.2 Περιεχόμενο Διπλωματικής

Στη βιβλιογραφία μπορεί κανείς να βρει αρκετές τεχνικές και μεθόδους για την ακουστική μοντελοποίηση και ανίχνευση. Ιδιαίτερα, η βαθιά εκμάθηση είναι ένα εξελισσόμενο και πολλά υποσχόμενο πεδίο για τις εργασίες ανίχνευσης και ταξινόμησης, όπως η αναγνώριση ομιλίας [10] και η ταξινόμηση εικόνας [11], και προσφέρει ελκυστικά αποτελέσματα καθώς το ποσό των δεδομένων αυξάνεται [12]. Επιπλέον, πολλές παραδοσιακές τεχνικές όπως τα Γκαουσιανά Μοντέλα Μίξης (Gaussian Mixture Models - GMM) και τα Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models - HMM) εφαρμόζονται ακόμη και προσφέρουν ικανοποιητικά αποτελέσματα. Η ακουστική μοντελοποίηση απαιτεί επίσης κατάλληλα χαρακτηριστικά και κάθε τεχνική συνδέεται συνήθως με ένα συγκεκριμένο τύπο χαρακτηριστικών. Για παράδειγμα, τα Βαθιά Νευρωνικά Δίκτυα (DNNs) χρησιμοποιούν Mel-Spectrograms ή filterbank energies [13], [6] και τα GMMs έχουν καλή απόδοση με τα Mel-Frequency Cepstral Coefficients (MFCC) [14].

Σε αυτή τη διπλωματική, επιχειρούμε να διερευνήσουμε διαφορετικά χαρακτηριστικά γνωρίσματα, όπως τα Chroma χαρακτηριστικά, με στόχο να εξετάσουμε την καλύτερη αναπαράσταση των ακουστικών σημάτων στο πεδίο των συχνοτήτων. Για την ταξινόμηση υιοθετήσαμε παραδοσιακούς ταξινομητές όπως οι Μηχανές Διανυσματικής Υποστήριξης (SVM) και τον ταξινομητή Τυχαίων Δέντρων Αποφάσεων (Random Forest).

1.3 Σχεδιάγραμμα Διπλωματικής

Η υπόλοιπη διπλωματική οργανώνεται ως εξής:

- Στο Κεφάλαιο 2 παρουσιάζουμε τα σχετικά ευρήματα που σχετίζονται με το θέμα μας καθώς και λεπτομέρειες για τον διαγωνισμό DCASE.
- Στο Κεφάλαιο 3 περιγράφουμε τις ιδιότητες των διάφορων τύπων χαρακτηριστικών που δοκιμάσαμε καθώς και πρόσθετες μεθόδους επεξεργασίας σήματος και χαρακτηριστικών.
- Στο Κεφάλαιο 4 αναφέρουμε συνοπτικά τη θεωρία των ταξινομητών που χρησιμοποιήσαμε στην εργασία μας.
- Στο Κεφάλαιο 5 αναλύουμε την επίδραση των τεχνικών που εφαρμόσαμε και συγκρίνουμε τα αποτελέσματα που λάβαμε για κάθε τύπο χαρακτηριστικών και ταξινομητών.
- Τέλος, στο Κεφάλαιο 6 παρέχουμε ένα συμπέρασμα και κάποιες σκέψεις για μελλοντικό έργο.

Κεφάλαιο 2

Σχετική Βιβλιογραφία

2.1 Οι διαγωνισμοί του DCASE

Το DCASE (Detection and Classification of Acoustic Sound Events) [15] είναι μια κοινότητα που υποστηρίζει ερευνητές με ακαδημαϊκό και βιομηχανικό υπόβαθρο για να διερευνήσουν και να μοιραστούν τις ιδέες και τις απόψεις τους σχετικά με την ταξινόμηση και ανίχνευση περιβαλλοντικών ήχων. Είναι μια δραστήρια και ταχέως αναπτυσσόμενη ερευνητική κοινότητα που κέρδισε ένα ειδικό τμήμα σε διεθνή συνέδρια επεξεργασίας σήματος όπως το EUSIPCO.

Ο ετήσιος διαγωνισμός του DCASE αποτελείται από πολλαπλές θεματικές ενότητες. Οι ενότητες του 2018 ήταν οι εξής:

- Ενότητα 1: Ταξινόμηση Ακουστικών Σκηνών (Acoustic scene classification).
- Ενότητα 2: Γενικού σκοπού σήμανση ήχου του περιεχομένου του Freesound με ετικέτες του AudioSet (General-purpose audio tagging of Freesound content with AudioSet labels).
- Ενότητα 3: Ανίχνευση Ήχων Πουλιών (Bird Audio Detection).
- Ενότητα 4: Μεγάλης κλίμακας ημι-επιβλεπόμενη ανίχνευση ηχητικών συμβάντων σε οικιακά περιβάλλοντα (Large-scale weakly labeled semi-supervised sound event detection in domestic environments)
- Ενότητα 5: Παρακολούθηση εγχώριων δραστηριοτήτων με ακουστική πολλαπλών καναλιών (Monitoring of domestic activities based on multi-channel acoustics).

Το έργο που παρουσιάζεται σε αυτή τη διατριβή αποτελεί μέρος της ενότητας 3. Η κοινότητα αποφάσισε να επεκτείνει την αρχική έκδοση του διαγωνισμού Ανίχνευσης Ήχων Πουλιών του 2016-2017 με σκοπό την περαιτέρω διερεύνηση σχετικών μεθόδων.

2.2 Ο πρώτος διαγωνισμός Ανίχνευσης Ήχων Πουλιών

Ο πρώτος διαγωνισμός BAD πραγματοποιήθηκε το 2016-2017 και ο στόχος ήταν

να αναπτυχθεί ένα σύστημα που να μπορεί να εντοπίσει την παρουσία ή απουσία ήχων πουλιών σε αρχεία ήχου. Τα σύνολα δεδομένων που δόθηκαν αποτελούνται από αρχεία ήχου διάρκειας 10 δευτερολέπτων και η έξοδος του ταξινομητή προτάθηκε να βρίσκεται στο διάστημα [0,1] αντί του δυαδικού «0» ή «1». Η μετρική που χρησιμοποιήθηκε για την απόδοση ταξινόμησης ήταν η Area Under the ROC Curve (AUC) [5].

Για την ανάπτυξη των μεθόδων δόθηκαν δύο σύνολα δεδομένων και οι ετικέτες τους. Ένα τρίτο dataset κρατήθηκε κρυφό μέχρι και την τελευταία μέρα του διαγωνισμού έτσι ώστε να αξιολογηθούν τα συστήματα σε άγνωστες συνθήκες. Η παρακάτω εικόνα συνοψίζει τα αποτελέσματα και της τεχνικές που χρησιμοποιήθηκαν από κάθε ομάδα στον διαγωνισμό καθώς και τα αποτελέσματα τους.

User	Preview Score	Final Score	Noise reduction	Features	Data augmentation	Classifier	Domain adaptation	Ensembling
DJBul	88.9 %	88.7 %	Mean-subtraction, silence/click trimming	Mel-spectrogram	time shift, freq shift	CNN	Pseudo-labelling	Model averaging
cakir	88.3 %	88.5 %	no	log Mel-band energies	no	CRNN	no	no (for strongest submission); Model averaging
tpoel	88.6 %	88.2 %	Did not help	Mel-spectrogram	freq shift, time shift, time-reversal	CNN-DenseNet	Pseudo-labelling	Multi-epoch, Model averaging(2point)
MarioElias	88.5 %	88.1 %	no	Segment-probabilities, openSMILE spectrograms(segmented)	noise, same-class combining, time shift, freq shift	CNN, ExtraTreesRegressor	no	Model averaging (over 2 diverse methods)
adavanne	88.2 %	88.1 %	no	log Mel-band energies, dominant-frequency	block mixing, test mixing	CRNN	Test mixing	no
Elias	88.0 %	88.0 %	Signal-noise segmentation	Spectrogram (signal/noise segmented)	noise, same-class combining, time shift, freq shift	CNN	no	Model averaging
kdrossos	86.1 %	85.8 %						
toan	85.9 %	85.6 %						
tiri	85.8 %	85.3 %						
RocketTeam	84.4 %	84.6 %	no	MFCC, Mel spectrogram	no	CNN	no	Max across an MFCC net and a Mel spec net
Mario	85.0 %	84.2 %	no	Segment-probabilities, openSMILE	no	ExtraTreesRegressor	no	Multi-model
kenders2000	85.0 %	83.9 %	no	Gammatone spectrogram	time shift, freq shift	CNN	no	no
sparrow	84.6 %	83.7 %	no	Mel-spectrogram	freq shift, noise	CNN	no	no
jensen	83.9 %	83.0 %	no	Mel-spectrogram	audio combining	CNN	no	no
kardi.piczak	83.6 %	82.5 %	no	spectrogram	noise, audio combining, time shift, permute segments	CNN	no	Model averaging
trucek	82.7 %	81.4 %	no	spectrogram	no	CNN	no	no
ding	82.3 %	80.8 %						
prometheus	78.7 %	80.3 %						
MMyll	80.8 %	79.7 %						
qiuqiangkong	80.8 %	79.1 %	OM-LSA, medianSub	Mel filterbank, spectrogram	no	CNN-joint-detector-classifier	no	no
WildcatOne	80.9 %	78.3 %	no	rawaudio, temporal-avg-pyramid	cropping	CNN	no	Averaging over multi augmentations of each datum
lelfan	79.9 %	78.3 %						
wilkinsong	78.0 %	77.9 %						
IIT_Mandi	77.2 %	75.2 %	cepstral normalisation, short-time feature warping	MFCC	no	GMM+PSK+SVM	no	no
friedrichvogel	74.4 %	74.5 %						
ncbin	73.5 %	72.4 %						
vmchikos	70.7 %	71.4 %						
NewDogOldTricks	70.5 %	71.2 %						
NewDogOldTricks	71.9 %	70.8 %						
kalimi	73.1 %	70.4 %						

Σχήμα 2.1: Σύνοψη των μεθόδων και των αποτελεσμάτων του DCASE BAD 2017 για κάθε ομάδα. Λήφθηκε από [1].

Όπως δείχνουν τα αποτελέσματα, οι περισσότεροι από τους συμμετέχοντες υιοθέτησαν μεθόδους βασισμένες σε Συνελικτικά Νευρωνικά Δίκτυα (CNN) σε συνδυασμό με Mel-Spectograms [6],[12],[16] ή F-BANKs [17],[13]. Επιπροσθέτως, ένας μικρός αριθμός ομάδων εργάστηκε με συμβατικά MFCC χαρακτηριστικά χρησιμοποιώντας Archetypal Analysis και SVMs [18],[8] , όπως επίσης και άλλου είδους χαρακτηριστικά και ταξινομητές [19],[20],[21] ακόμη και τεχνικές Matrix Factorization [7].

Όλες οι παραπάνω ομάδες εξέτασαν αρκετές τεχνικές αύξησης δεδομένων (Data Augmentation) [6], μείωσης θορύβου [18] και προσαρμογής τομέα (Domain Adaptation) [13] με στόχο την επιπλέον γενίκευση των μοντέλων.

Το Ευρωπαϊκό Συνέδριο Επεξεργασίας Σήματος (EUSIPCO) 2017 παρείχε ένα ειδικό

τμήμα στο οποίο οι συμμετέχοντες υπέβαλαν το έργο τους. Οκτώ δημοσιεύσεις του διαγωνισμού έγιναν αποδεκτές και παρουσιάστηκαν μαζί με επιπλέον δημοσιεύσεις, η δουλειά των οποίων αναφέρθηκε παραπάνω [1].

2.3 DCASE 2018: Ενότητα 3

Μετά το πέρας του διαγωνισμού του 2017, η κοινότητα του DCASE θέλησε να επεκτείνει το έργο με σκοπό τη γενίκευση σε νέες συνθήκες. Προσπάθησαν να ενθαρρύνουν τις ομάδες να αναπτύξουν συστήματα που μπορούν εκ φύσεως να γενικεύουν σε όλες τις συνθήκες (συμπεριλαμβανομένων των συνθηκών που δεν εμφανίζονται στα δεδομένα εκπαίδευσης) ή που μπορούν να προσαρμοστούν σε νέα σύνολα δεδομένων (Domain Adaptation). Το ισχυρότερο σύστημα του 2017 [6] επιλέχθηκε ως το βασικό (baseline) σύστημα και η πλειοψηφία των ομάδων εξέτασε τεχνικές πάνω σε αυτό.

Στο σύνολο δεδομένων που χρησιμοποιήθηκαν για ανάπτυξη προστέθηκε ένα ακόμη σύνολο δεδομένων, δηλαδή συνολικά προσφέρθηκαν 3. Με αυτό τον τρόπο, οι διαγωνιζόμενοι είχαν την ευκαιρία να εξερευνήσουν τη συμπεριφορά του συστήματος τους σε διαφορετικές συνθήκες. Επιπλέον, δόθηκε η δυνατότητα χρήσης εξωτερικών συνόλων δεδομένων για τη διαδικασία της ανάπτυξης, τα οποία όμως έπρεπε να επισημανθούν δημόσια, έτσι ώστε όλοι να έχουν την ίδια ευκαιρία να τα χρησιμοποιήσουν. Τρία υποσύνολα δεδομένων κρατήθηκαν κρυφά για τη διαδικασία της αξιολόγησης των συστημάτων. Το ένα από αυτά ήταν κομμάτι του συνόλου ανάπτυξης και τα υπόλοιπα αγνώστων συνθηκών.

Η προτεινόμενη διαδικασία αξιολόγησης διέφερε από την κλασσική. Στις παραδοσιακές εργασίες Μηχανικής Μάθησης η διαδικασία αξιολόγησης (cross-validation) διαιρεί το σύνολο δεδομένων σε κομμάτια (folds) τα οποία χρησιμοποιούνται τόσο για ανάπτυξη όσο και για αξιολόγηση. Συνεπώς, τα δεδομένα αξιολόγησης του συστήματος (evaluation data) διατηρούν τις ίδιες συνθήκες με αυτά που χρησιμοποιήθηκαν για ανάπτυξη. Η παραπάνω κλασσική διαδικασία δεν εξυπηρετεί τις ανάγκες του προβλήματος της γενίκευσης, όποτε και προτάθηκε μια παραλλαγή αυτής. Συγκεκριμένα, πρόκειται για μία stratified 3-way cross-validation διαδικασία αξιολόγησης όπου σε κάθε επανάληψη δύο σύνολα δεδομένων χρησιμοποιούνται για ανάπτυξη και ένα για αξιολόγηση. Ως μετρική αξιολόγησης της απόδοσης ορίζεται ξανά η AUC η οποία θα υπολογίζεται ξεχωριστά για κάθε σύνολο δεδομένων που αξιολογεί το σύστημα. Το συνολικό αποτέλεσμα του συστήματος λαμβάνεται υπολογίζοντας την αρμονική μέση τιμή (harmonic mean) των AUC του κάθε συνόλου δεδομένων. Η παραπάνω εναλλακτική διαδικασία δεν επηρεάζεται από τον αριθμό των δεδομένων του κάθε συνόλου αφού υπολογίζει τη συνολική απόδοση με ένα σταθμισμένο τρόπο. Επιπλέον, επιτρέπει τη διερεύνηση του ορίου ανίχνευσης για κάθε σύνολο δεδομένων ξεχωριστά. Όλα τα παραπάνω κάνουν την προτεινόμενη διαδικασία να πλεονεκτεί σε σχέση με την «απλή».

Τα αποτελέσματα του διαγωνισμού δεν παρουσίασαν σημαντική βελτίωση παρόλο που εξετάστηκαν αρκετές μέθοδοι. Συγκεκριμένα, η ομάδα με τις υψηλότερες βαθμολογίες χρησιμοποίησε το ισχυρότερο σύστημα του περασμένου έτους, συμπεριλαμβανομένων πολλών μεθόδων όπως η ενίσχυση σήματος και μια τεχνική προσαρμογής τομέα. Η βελτίωση στο αποτέλεσμα προήλθε από τον συνδυασμό μεμονωμένων μεθόδων που διερευνήθηκαν [22]. Παράλληλα, μία ενδιαφέρουσα προσέγγιση με Capsule Networks [23] διακρίθηκε και

βραβεύτηκε από τους κριτές και μια διαφορετική τεχνική προσαρμογής τομέα [24] έλαβε ειδική μνεία.

Όπως παρατηρήθηκε από τους διοργανωτές, παρά τη διετή έρευνα πάνω σε μεθόδους γενίκευσης κυρίως σε νευρωνικά δίκτυα, υπάρχει ακόμη αρκετή διαφορά στην ακρίβεια των μοντέλων σε δεδομένα ίδιων συνθηκών και σε εκείνα που διαφέρουν. Για αυτό τον λόγο αναφέρθηκε η εκδοχή εξερεύνησης υψηλότερης ανάλυσης χαρακτηριστικών, όπως για παράδειγμα δεδομένα κυματομορφών και άλλα [1]. Με αυτή την παρατήρηση, στην παρούσα διπλωματική επιχειρούμε να ερευνήσουμε χαρακτηριστικά που κανένα σύστημα του διαγωνισμού δεν έχει συμπεριλάβει έως τώρα.

Κεφάλαιο 3

Μέθοδοι

Σε αυτή την ενότητα εξηγούμε εν συντομία το θεωρητικό μέρος των χαρακτηριστικών που χρησιμοποιήσαμε και πώς τα προσαρμόσαμε στο πρόβλημά μας. Υπάρχουν πολλά εργαλεία για την παραγωγή χαρακτηριστικών, είτε βιβλιοθήκες γλωσσών προγραμματισμού είτε άλλα λογισμικά ανοιχτού κώδικα. Δεδομένου ότι η Python είναι η γλώσσα προγραμματισμού που επιλέξαμε για την ανάπτυξη του κώδικα μας, χρησιμοποιήσαμε από τα προτεινόμενα πακέτα τη βιβλιοθήκη Librosa [25] η οποία είναι μία ευρέως χρησιμοποιούμενη βιβλιοθήκη για την εξαγωγή χαρακτηριστικών ήχου. Επιπλέον, εργαστήκαμε και με το OpenSmile Toolkit [26] αφού είναι ένα αρκετά γνωστό εργαλείο επεξεργασίας ήχου. Εφόσον λοιπόν το Opensmile έχει αναπτυχθεί από ειδικούς στον χώρο ερευνητές, μελετήσαμε τη συμπεριφορά των χαρακτηριστικών με τις ρυθμίσεις που εκείνοι προτείνουν.

3.1 Χαρακτηριστικά

3.1.1 Mel-Frequency Cepstral Coefficients

Τα MFCC επιλέγονται συχνά για εργασίες ταξινόμησης ήχου [27]. Η διαδικασία παραγωγής τους ξεκινά δημιουργώντας n πλαίσια (frames) από το αρχικό σήμα στα οποία εν συνεχεία εφαρμόζεται ένα παράθυρο για να εξομαλύνει τις ασυνέχειες μεταξύ τους. Έπειτα λαμβάνεται ο μετασχηματισμός Fourier του σήματος και το προκύπτον φάσμα μετατρέπεται στην κλίμακα Mel με τη βοήθεια ενός Mel φίλτρου. Κατόπιν, υπολογίζεται ο λογάριθμος κάθε Mel φάσματος και εφαρμόζεται ο διακριτός μετασχηματισμός συνημιτόνου (DCT) στο προηγούμενο αποτέλεσμα. Η κλίμακα Mel υπολογίζεται από

$$Mel(f) = 2595 log_{10}(1 + \frac{f}{700}) \tag{3.1}$$

και θεωρείται ότι προσεγγίζει το ανθρώπινο ακουστικό σύστημα καλύτερα από τις γραμμικά διαχωρισμένες ζώνες συχνοτήτων του κανονικού cepstrum [28]. Ως cepstrum ορίζουμε το φάσμα που προκύπτει από την εφαρμογή του διακριτού μετασχηματισμού συνημιτόνου στο λογαριθμικό φάσμα του αρχικού σήματος. Επιπλέον, είναι ευρέως γνωστό ότι η εφαρμογή του DCT παράγει ασυσχέτιστα δεδομένα, πράγμα που κάνει πιο εύκολη τη μοντελοποίηση τους από αλγορίθμους Μηχανικής Μάθησης.

Στα παραπάνω χαρακτηριστικά συνήθως προστίθενται οι συντελεστές delta, delta-delta οι οποίοι προσπαθούν να αναπαραστήσουν τις χρονικές αλλαγές στο σήμα.

3.1.1.1 OpenSmile MFCC

Το εργαλείο Opensmile προσφέρει διάφορα αρχεία ρυθμίσεων για την παραγωγή χαρακτηριστικών. Ιδιαίτερα, επιλέγουμε να εργαστούμε με το αρχείο "MFCC12_E_D_A_Z.conf" στο οποίο υπολογίζονται 13 MFCC από 26 Mel-frequency Bands. Το μέγεθος των frames είναι 25ms και δειγματοληπτούνται με ρυθμό 10ms χρησιμοποιώντας ένα Hamming window. Ο συντελεστής μηδενικής τάξης αντικαθίσταται από τη λογαριθμική ενέργεια του σήματος και επιπλέον προστίθενται 13 delta και 13 delta-delta συντελεστές στα προϋπολογισμένα MFCC. Στη συνέχεια, τα προκύπτοντα χαρακτηριστικά κανονικοποιούνται [26].

3.1.1.2 Librosa MFCC

Η βιβλιοθήκη Librosa διαθέτει μία προκαθορισμένη συνάρτηση για τον υπολογισμό των MFCC χαρακτηριστικών. Με βάση αυτή, λάβαμε 13 χαρακτηριστικά MFCC τροφοδοτώντας το ακουστικό σήμα στην αντίστοιχη συνάρτηση. Το αποτέλεσμα υπολογίζεται με FFT σε Hanning Window μήκους 2048 και hop length ίσο με 512, έτσι ώστε να έχουμε 75% επικάλυψη. Ο ρυθμός δειγματοληψίας ορίστηκε σε 22.05 kHz με στόχο τη μείωση της ποσότητας των δεδομένων χωρίς όμως να χάνουμε πολύ πληροφορία. Επιπλέον, προστέθηκαν οι συντελεστές delta και delta-delta δίνοντας ως αποτέλεσμα ένα διάνυσμα χαρακτηριστικών διάστασης 39.

3.1.2 Chroma Features

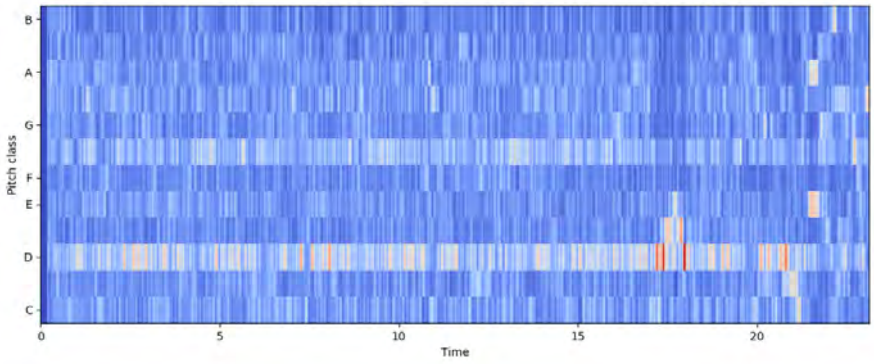
Στη θεωρία της μουσικής, μια οκτάβα είναι το διάστημα μεταξύ δύο μουσικών φθόγγων με ένα από αυτά να έχει τη διπλάσια συχνότητα από το άλλο. Η αντίληψη του ανθρώπου για τους φθόγγους ή διαφορετικά «εντάσεις ήχων» (pitches) είναι περιοδική. Ως αποτέλεσμα, οι «εντάσεις» (pitces) που χωρίζονται από μια οκτάβα ανήκουν στην ίδια κατηγορία (pitch class) και αντιπροσωπεύονται από το ίδιο χρώμα. Με αυτή τη συμβολική αναπαράσταση, χαρακτηρίζουμε ένα pitch από το χρώμα του και το ύψος του. Αναφερόμενοι στη Δυτική Μουσική Σημειογραφία και την ισοσταθμισμένη κλίμακα, κάθε χρώμα αντιστοιχεί σε μία από τις δώδεκα κλάσεις $\{C, C\sharp, D, D\sharp, E, F, F\sharp, G, G\sharp, A, A\sharp, B\}$. Το διάνυσμα χαρακτηριστικών Chroma που προκύπτει αποτελείται από διανύσματα που αντιπροσωπεύουν τον τρόπο με τον οποίο κατανέμεται η ενέργεια σε αυτές τις δώδεκα κλάσεις χρωμάτων [29].

Τα χαρακτηριστικά Chroma είναι ευρέως γνωστό ότι αναγνωρίζουν τα pitces που διαφέρουν κατά μια οκτάβα. Ως pitch ορίζεται η δυνατότητα του ανθρώπου να αντιλαμβάνεται έναν ήχο ως υψηλό ή χαμηλό. Συνεπώς, χρησιμοποιούνται για την ανάλυση και την επεξεργασία μουσικών δεδομένων, λόγω του ότι είναι ανθεκτικά στις αλλαγές του τέμπο και συνδέονται άμεσα με τη μουσική πτυχή της αρμονίας. Αυτή η προοπτική μας ώθησε να διερευνήσουμε τις επιδόσεις τους στους ήχους των πουλιών, το οποία είναι μελωδικά δεδομένα και πιθανότατα θα αναπαρίστανται ιδανικά από τα χαρακτηριστικά χρώματος.

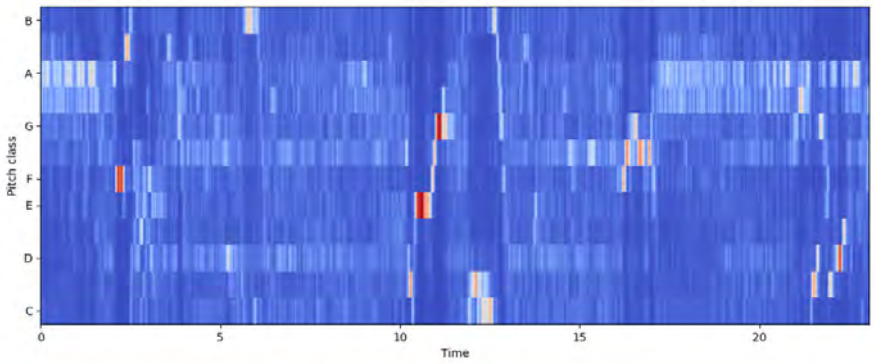
Υπάρχουν αρκετές εφαρμογές για την εξαγωγή τω Chroma χαρακτηριστικών. Η βιβλιοθήκη Librosa παρέχει τρεις διαφορετικές υλοποιήσεις, η μια εκ των οποίων βασίζεται στον Σύντομο Μετασχηματισμό Fourier (STFT), μια άλλη στον Μετασχηματισμό Constant-Q (CQT) και η τελευταία στον CQT με ένα επιπλέον βήμα κανονικοποίησης (CENS). Επιλέξαμε να χρησιμοποιήσουμε τους τελευταίους δύο και την υλοποίηση που παρέχει το εργαλείο OpenSmile.

3.1.2.1 OpenSmile Chromagram

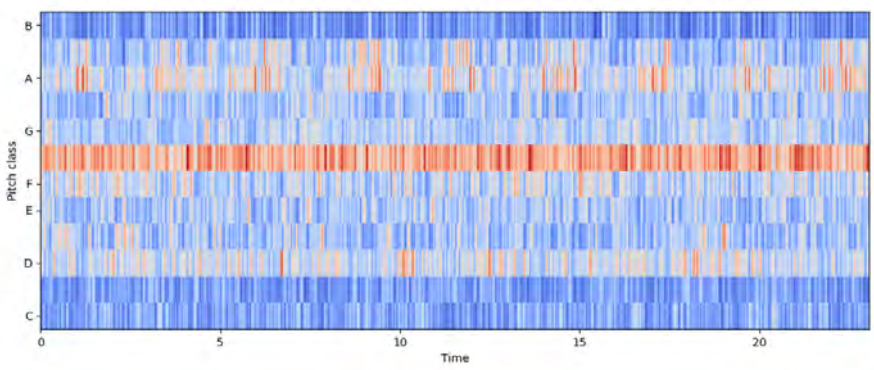
Το αρχείο ρυθμίσεων που παρέχει το OpenSmile εξάγει 12 χαρακτηριστικά Chroma από ένα short-time FFT spectrogram με ένα παράθυρο Gauss μήκους 50ms και ρυθμό δειγματοληψίας 10 ms. Εφαρμόζοντας τριγωνικά φίλτρα, το φασματογράφημα κλιμακώνεται σε ημιτονική κλίμακα παράγοντας το τελικό αποτέλεσμα [26]. Παρακάτω φαίνονται τα χαρακτηριστικά Chroma του OpenSmile Toolkit.



(α′) warblrb10k : 0d30dbf2-dfdc-4e46-bce3.wav



(β′) ff1010bird : 182740.wav



(γ′) BirdVox-DCASE-20k : c1dc30de-1cbd-454e-a565-b1d929fecf73.wav

Σχήμα 3.1: Χαρακτηριστικά Chroma του Opensmile Toolkit

3.1.2.2 Librosa Constant-Q Chromagram

Βασισμένη στο [30] η βιβλιοθήκη Librosa υλοποιεί την εξαγωγή Chroma χαρακτηριστικών με χρήση του Constant-Q μετασχηματισμού. Ο μετασχηματισμός Constant-Q μετατρέπει ένα σήμα από το πεδίο του χρόνου στο πεδίο της συχνότητας διαχωρίζοντας γεωμετρικά τις συχνότητες και διατηρώντας τους συντελεστές Q ίσους. Δηλαδή για τις συχνότητες ισχύει [30]:

$$f_k = f_1 2^{\frac{k-1}{B}} \tag{3.2}$$

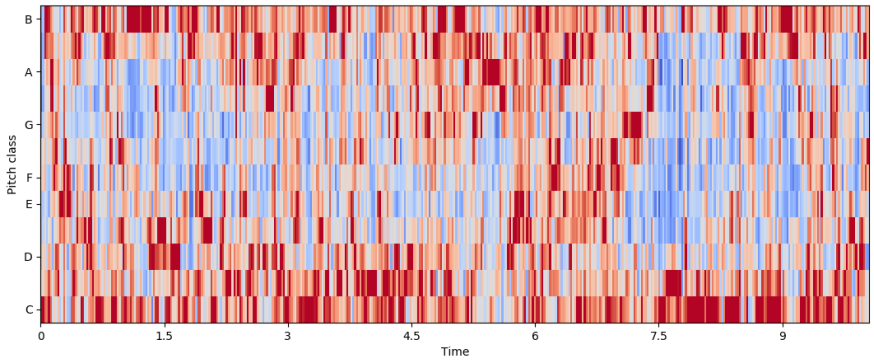
όπου το $k = 1, 2, ...K$ αναπαριστά τα διαστήματα συχνοτήτων του CQT, f_1 είναι η κεντρική συχνότητα του χαμηλότερου διαστήματος συχνότητας (frequency bin) και το B καθορίζει τον αριθμό των «κάδων» ανά οκτάβα. Επιπλέον, οι συντελεστές Q δίνονται από:

$$Q = \frac{f_k}{\Delta f_k} \tag{3.3}$$

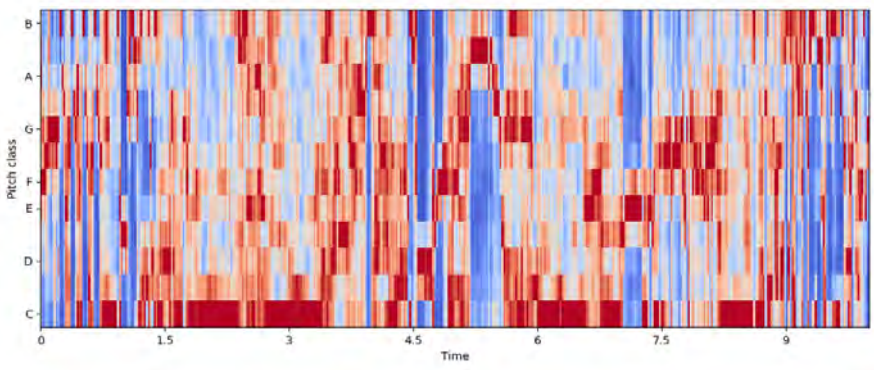
και εξ′ ορισμού διατηρούνται σταθεροί για τον CQT.

Ο παραπάνω μετασχηματισμός έχει ως αποτέλεσμα την καλύτερη ανάλυση συχνότητας των χαμηλών συχνοτήτων και την αποδοτικότερη ανάλυση χρόνου των υψηλών συχνοτήτων. Η πρώτη απόπειρα χρήσης του CQT [31] για μουσικά δεδομένα δεν κέντρισε το ενδιαφέρον της ερευνητικής κοινότητας αφού υπέφερε από βασικά προβλήματα όπως η πολυπλοκότητα και η έλλειψη αντίστροφου μετασχηματισμού. Παρ′ όλα αυτά, τα επόμενα χρόνια προτάθηκαν λύσεις στα παραπάνω εμπόδια και δημιουργήθηκε μια βιβλιοθήκη σε Matlab η οποία υπολόγιζε αποδοτικότερα τον συγκεκριμένο μετασχηματισμό και υποστήριζε τον αντίστροφο του με αρκετή ακρίβεια [30].

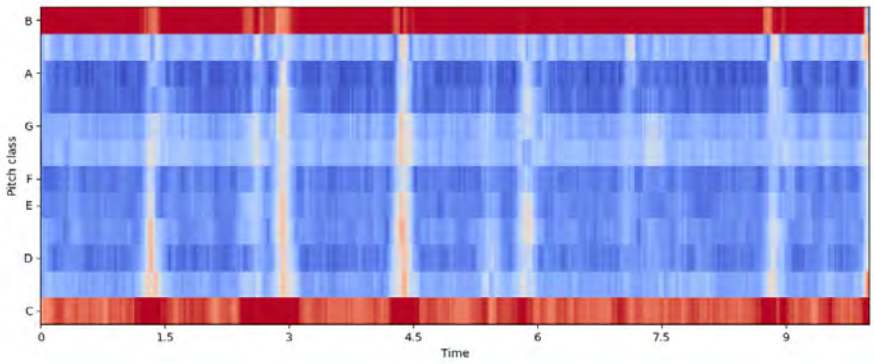
Για την εξαγωγή χαρακτηριστικών με χρήση της Python χρησιμοποιήσαμε τις προεπιλεγμένες παραμέτρους και στην επόμενη εικόνα παρουσιάζουμε τη μορφή αυτών των χαρακτηριστικών. Η εν λόγω υλοποίηση εφαρμόζει κανονικοποίηση των δεδομένων ανά στήλη διαιρώντας κάθε στοιχείο της στήλης με τη μέγιστη τιμή της.



(α′) warblrb10k : 0d30dbf2-dfdc-4e46-bce3.wav



(β′) ff1010bird : 182740.wav



(γ′) BirdVox-DCASE-20k : c1dc30de-1cbd-454e-a565-b1d929fecf73.wav

Σχήμα 3.2: Χαρακτηριστικά CQT της βιβλιοθήκης Librosa

3.1.2.3 Librosa Chroma Energy Normalized (CENS)

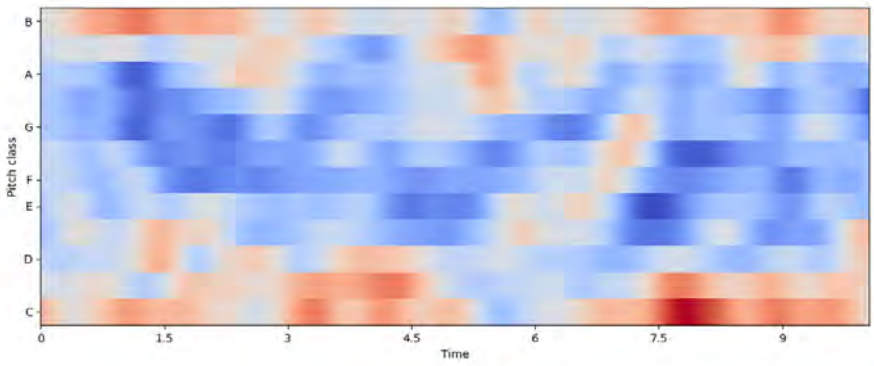
Τα CENS χαρακτηριστικά παράγονται από τη στατιστική ανάλυση της κατανομής της ενέργειας στις ζώνες χρώματος. Αποτελούν ισχυρά ηχητικά χαρακτηριστικά μιας και είναι σταθερά σε δυναμικές αλλαγές, σε αλλαγές στο τέμπο (trimbe) και στην προφορά, δηλαδή

στον τρόπο με τον οποίο ακούγεται μία νότα.

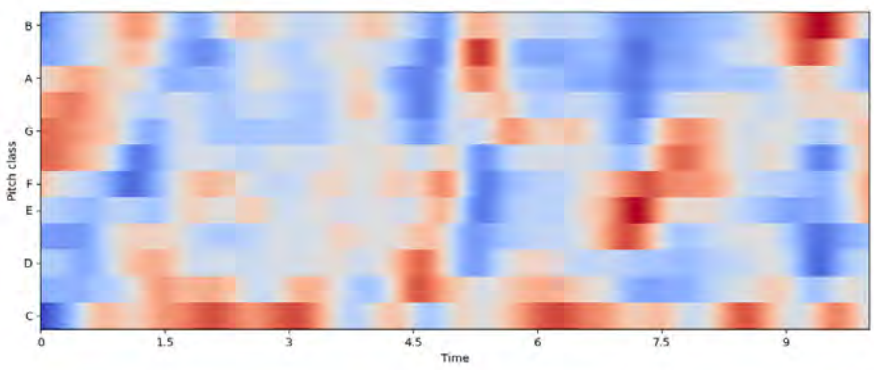
Σύμφωνα με το [29] το πρώτο βήμα για την παραγωγή των χαρακτηριστικών CENS αφορά την αναπαράσταση κατανομής της ενέργειας η οποία λαμβάνεται από την κανονικοποίηση του κάθε Chroma διανύσματος με τη χρήση της $\ell 1$ νόρμας . Επιπλέον, διαλέγοντας κατάλληλα λογαριθμικά όρια ποσοτικοποιούνται τα διανύσματα χαρακτηριστικών και εισάγεται κάποια λογαριθμική συμπίεση. Στη συνέχεια, εφαρμόζοντας ένα παράθυρο μήκους w τα διανύσματα εξομαλύνονται επιπλέον και μειώνονται με δειγματοληψία (downsampling) κατά ένα παράγοντα d. Το τελικό αποτέλεσμα διαμορφώνεται με επιπλέον κανονικοποίηση χρησιμοποιώντας την $\ell 2$ νόρμα. Η βιβλιοθήκη Librosa υλοποιεί με την ίδια σειρά τα βήματα για την παραγωγή των χαρακτηριστικών CENS με εξαίρεση του βήματος της υποδειγματολειψίας (downsampling).

Η φύση των CENS και η σταθερότητα τους στις δυναμικές αλλαγές τα καθιστά κατάλληλα για εφαρμογές όπως η αντιστοίχιση (audio matching) [32] και η ανάκτηση ήχου (audio retrieval). Επιπλέον, τα εν λόγω χαρακτηριστικά διανύσματα αναπαριστούν επαρκώς μελωδικά σήματα αφού συνδέονται στενά με την αρμονική εξέλιξη τέτοιου είδους σημάτων. Το γεγονός ότι το υπόβαθρό τους είναι τα χαρακτηριστικά Chroma τα κάνει ανθεκτικά στις αλλαγές στο τέμπο και η κανονικοποίηση που εφαρμόζεται τα σταθεροποιεί ως προς τις δυναμικές αλλαγές. Τέλος, η χρήση λογαριθμικών ορίων ενέργειας καθιστά τα CENS ανθεκτικά στο θόρυβο που μπορεί να προκύψει από λάθος νότες και η χρήση παραθύρων αντισταθμίζει τις διαφορές που μπορεί να προκύψουν από όμοιες ομάδες νοτών [32].

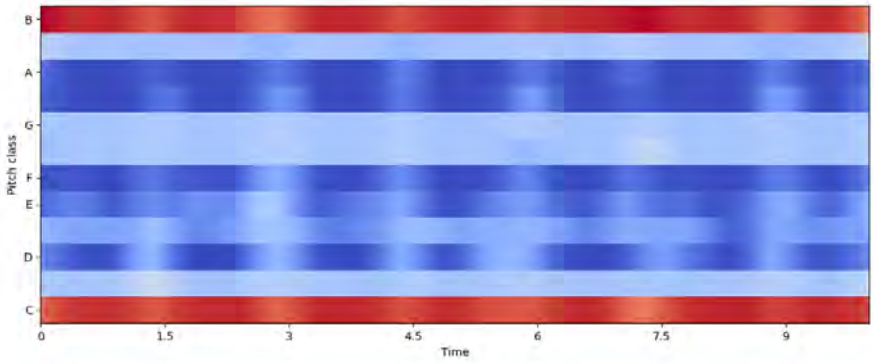
Όπως στην προηγούμενη υλοποίηση έτσι και εδώ χρησιμοποιήσαμε τις προεπιλεγμένες παραμέτρους της συνάρτησης τις βιβλιοθήκης Librosa και η μορφή των χαρακτηριστικών φαίνεται στο σχήμα 3.3.



(α′) warblrb10k : 0d30dbf2-dfdc-4e46-bce3.wav



(β′) ff1010bird : 182740.wav



(γ′) BirdVox-DCASE-20k : c1dc30de-1cbd-454e-a565-b1d929fecf73.wav

Σχήμα 3.3: Χαρακτηριστικά CENS της βιβλιοθήκης Librosa

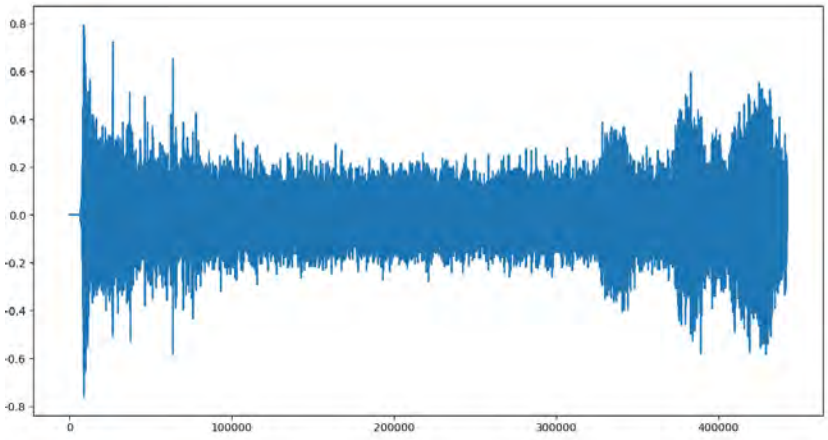
3.2 Επιπλέον Μέθοδοι

3.2.1 Μείωση Θορύβου

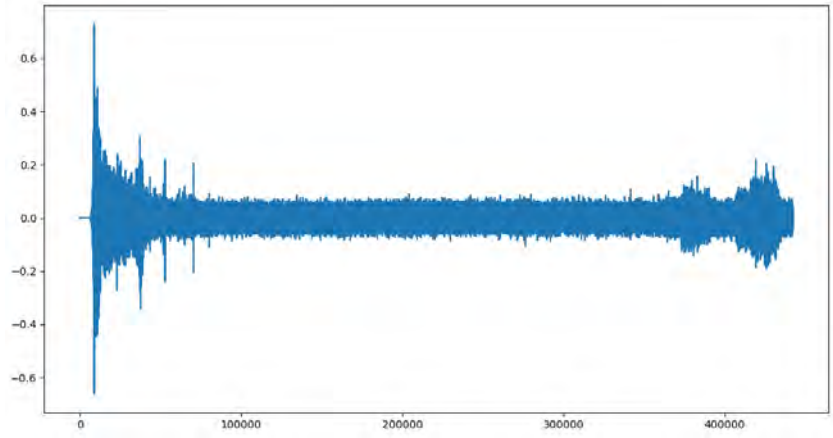
Η μείωση του θορύβου είναι μια ενδιαφέρουσα και συνηθισμένη τεχνική για τα αρχεία ήχου και συνήθως εφαρμόζεται χρησιμοποιώντας ένα σταθερό όριο θορύβου. Όσον αφορά ηχογραφήσεις που αντλούνται από φυσικά περιβάλλοντα όπως μία πόλη ή ένα δάσος, η έννοια του θορύβου περιλαμβάνει και ήχους που προέρχονται από διάφορα ζώα, τη βροχή και τον αέρα. Συγκεκριμένα, τέτοιου είδους θόρυβος συναντάται συχνά στα δεδομένα του διαγωνισμού και αξίζει να εξερευνηθούν τεχνικές οι οποίες τον αναγνωρίζουν και τον απομακρύνουν.

Μια τεχνική αντιμετώπισης του θορύβου είναι η εφαρμογή ενός φίλτρου. Από τη θεωρία Επεξεργασίας Ψηφιακών Σημάτων γνωρίζουμε ότι υπάρχουν διαφόρων ειδών φίλτρα, όπως τα Chebyshev, τα Butterworth, τα Elliptic filters και άλλα. Επιπρόσθετα, τα φίλτρα διαχωρίζονται σε Low-pass , High-pass, Band-pass filters και άλλα ανάλογα με τις ζώνες συχνοτήτων που επιτρέπουν. Θα εστιάσουμε στα υψιπερατά φίλτρα και ιδιαίτερα σε ένα Butterworth High Pass φίλτρο δεύτερης τάξης.

Συγκεκριμένα, στο [33] υποστηρίζεται ότι εφαρμόζοντας ένα υψιπερατό φίλτρο βελτιώνουμε το σήμα ενισχύοντας τις συχνότητες που σχετίζονται με ήχους πουλιών. Ιδιαίτερα εφαρμόσαμε σε κάθε σήμα εισόδου ένα Butterworth High Pass φίλτρο δεύτερης τάξης με συχνότητα αποκοπής ίση με 2kHz. Παρακάτω (σχήματα 3.4,3.5,3.6) φαίνονται, μετά από τυχαία επιλογή, αρχεία ήχου που περιέχουν ήχους πουλιών πριν και μετά το φίλτρο. Όπως είναι φανερό, ξαφνικές αλλαγές και αιχμές (peaks) του αρχικού σήματος διατηρούνται και ενισχύονται μετά την εφαρμογή του φίλτρου.

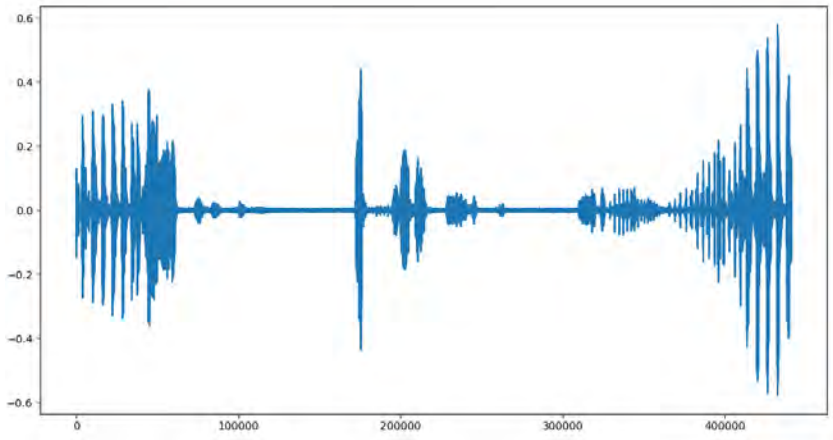


(α′) Αρχικό σήμα

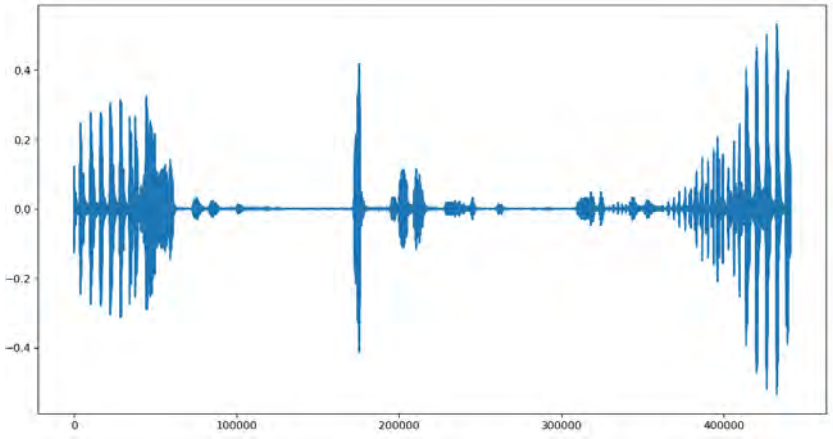


(β′) Μετά το φίλτρο

Σχήμα 3.4: Επίδραση υψιπερατού φίλτρου στο αρχείο ήχου 0d30dbf2-dfdc-4e46-bce3.wav το οποίο ανήκει στο warblrb10k dataset.

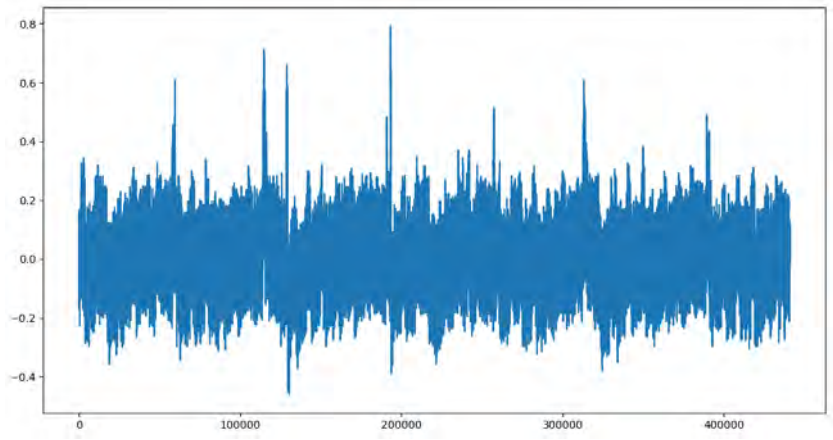


(α′) Αρχικό σήμα

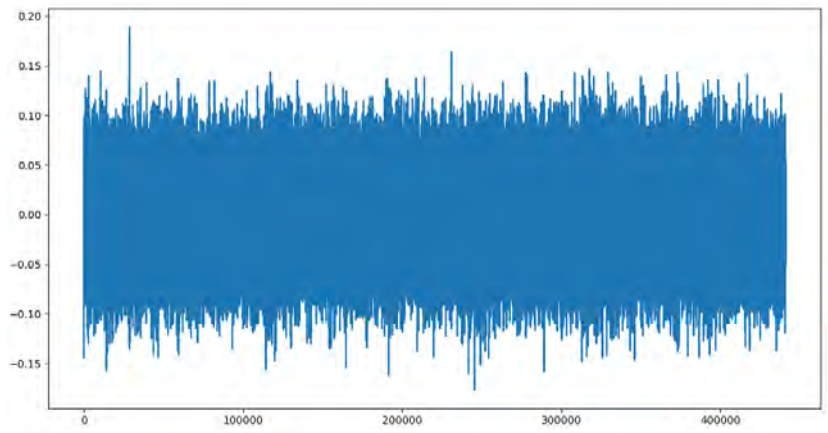


(β′) Μετά το φίλτρο

Σχήμα 3.5: Επίδραση υψιπερατού φίλτρου στο αρχείο ήχου: 182740.wav το οποίο ανήκει στο ff1010bird dataset.



(α′) Αρχικό σήμα



(β′) Μετά το φίλτρο

Σχήμα 3.6: Επίδραση υψιπερατού φίλτρου στο αρχείο ήχου: c1dc30de-1cbd-454e-a565-b1d929fecf73.wav το οποίο ανήκει στο BirdVox-DCASE-20k dataset.

3.2.2 Κανονικοποίηση Χαρακτηριστικών

Η κανονικοποίηση (normalization) είναι μια διαδικασία που εφαρμόζεται συχνά ως πρωταρχικό βήμα στην επεξεργασία δεδομένων για τους αλγορίθμους Μηχανικής Μάθησης. Η συμβολή της είναι ιδιαίτερα σημαντική ειδικά για δεδομένα των οποίων οι τιμές διαφέρουν αρκετά. Συνήθως τα δεδομένα που υφίστανται κανονικοποίηση κλιμακώνονται στο διάστημα [0,1] ή [-1,1] ανάλογα με το είδος και τις ανάγκες του προβλήματος.

Όσον αφορά το πρόβλημα μας, οι διαφορετικές συσκευές που χρησιμοποιήθηκαν για την εγγραφή των αρχείων ήχου επηρεάζουν τις τιμές των δεδομένων μιας και κάθε σύνολο δεδομένων θα παρουσιάζει διαφορές στη μέση τιμή και την τυπική απόκλιση. Έτσι σύμφωνα με το [18], αν αφαιρέσουμε από το διάνυσμα των χαρακτηριστικών τη μέση τιμή και διαιρέσουμε με την τυπική απόκλιση τότε θα μετριάσουμε την επιρροή που έχουν οι διαφορετικές συσκευές στα διανύσματα χαρακτηριστικών. Συνεπώς, για ένα διάνυσμα χαρακτηριστικών $X = \{x_1, x_2, .., x_n\}$ με i διανύσματα χαρακτηριστικών και n frames το καθένα, υπολογίζουμε για κάθε αρχείο ξεχωριστά τη μέση τιμή (μ) και τη διασπορά (σ^2) και το κανονικοποιημένο διάνυσμα χαρακτηριστικών προκύπτει από:

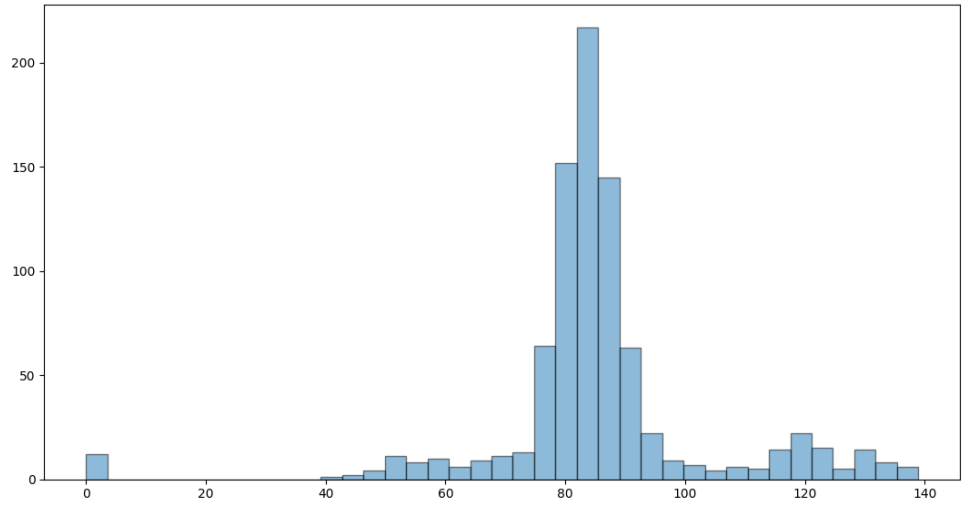
$$Y_n(i) = \frac{x_n(i) - \mu(i)}{\sigma(i)} \tag{3.4}$$

Ως επιπρόσθετο βήμα για να αντιμετωπιστούν τα προβλήματα που προκαλούν οι διαφορετικές συσκευές στα δεδομένα εφαρμόζουμε την αντιστοίχιση τους σε μία κανονική μορφή. Στην προσέγγισή μας, μετασχηματίσαμε τα δεδομένα έτσι ώστε να ακολουθούν την κανονική κατανομή στοχεύοντας στην εξάλειψη επιπλέον θορύβου στα αρχεία. Για τον εν λόγω μετασχηματισμό χρησιμοποιήσαμε τη συνάρτηση QuantileTransformer του πακέτου signal.preprocessing της Python. Η ζητούμενη κανονικοποιημένη μορφή παράγεται από τη συνάρτηση:

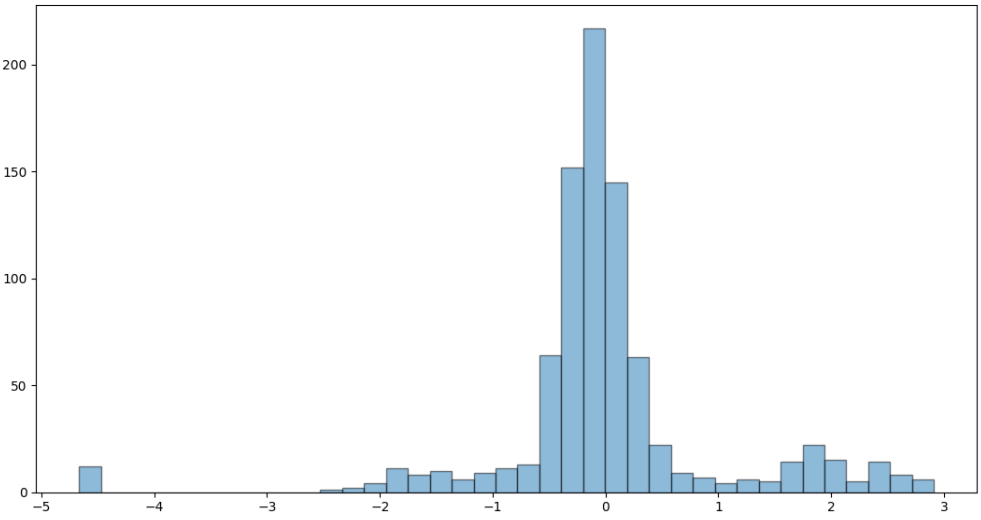
$$G^{-1}(F(X)) \tag{3.5}$$

όπου F είναι η Αθροιστική Συνάρτηση Κατανομής (Cumulative Distribution Function) των χαρακτηριστικών και G^{-1} η ποσοτική συνάρτηση της επιθυμητής κατανομής εξόδου

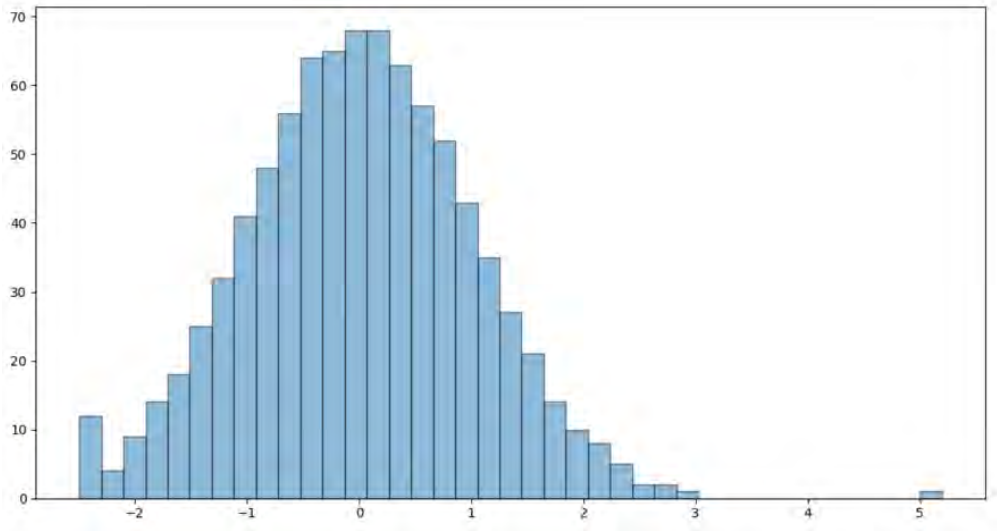
G, όπου στη συγκεκριμένη περίπτωση είναι η κανονική κατανομή. Στο σχήμα 3.7 φαίνεται συντελεστής πρώτης τάξης στην αρχική του μορφή (3.7α), μετά την κανονικοποίηση (3.7β) και μετά την αντιστοίχιση του στην κανονική κατανομή (3.7γ).



(α΄) Αρχική μορφή



(β΄) Κανονικοποιημένη μορφή



(γ′) Μορφή μετασχηματισμένη στην κανονική κατανομή

Σχήμα 3.7: Πρώτης τάξης συντελεστής των MFCC χαρακτηριστικών κατά τα στάδια της κανονικοποίησης. Αρχείο ήχου warblrb10k : 0d30dbf2-dfdc-4e46-bce3.wav

3.3 Δεδομένα

Τα δεδομένα που παρείχε ο διαγωνισμός προερχόταν από πολλαπλές πηγές όπως συσκευές απομακρυσμένης παρακολούθησης οι οποίες παθητικά συλλέγουν δεδομένα και κινητές συσκευές όπως για παράδειγμα smartphones τα οποία συμβάλουν στην ενεργή συλλογή. Όπως είναι λογικό, οι διαφορές στα ακουστικά δεδομένα αφορούν τόσο την αναλογία υπάρχει πουλί ή όχι στην ηχογράφηση, όσο και στο μέσο που χρησιμοποιήθηκε για αυτή. Αυτή η επιλογή πολλαπλών πηγών, στοχεύει στην καλύτερη γενίκευση σε νέες συνθήκες παρέχοντας στον υποψήφιους διαγωνιζόμενους αρκετά δεδομένα τόσο για ανάπτυξη όσο και για τον έλεγχο των αλγορίθμων [1]. Τα δεδομένα που χρησιμοποιήθηκαν στην ανάπτυξη της δικής μας προσέγγισης είναι τα εξής:

- **Warblr dataset** : Αποτελεί ένα εκ των δύο συνόλων δεδομένων που αντλήθηκαν από πολλαπλές πηγές. Ανήκει στο Warblr project του Ηνωμένου Βασιλείου το οποίο παρέχει μια εφαρμογή για smartphones η οποία από 10 δευτερολέπτων αρχείων ήχου ταξινομεί τα είδη των πουλιών. Το σύνολο απαρτίζεται από 8,000 αρχεία τα οποία συλλέχθηκαν ενεργά από συσκευές κινητής τηλεφωνίας και πολύ πιθανόν να περιέχουν ομιλία, θόρυβο προερχόμενο από καιρικές συνθήκες όπως βροχή και αέρας αλλά και μιμήσεις ήχων πουλιών από ανθρώπους. Οι συνθήκες των ηχογραφήσεων αφορούν κυρίως πρωϊνά και σαββατοκύριακα σε όλες τις περιοχές του Ηνωμένου Βασιλείου και όλες τις εποχές του χρόνου [1].

- **Freefield1010 dataset** [34] : Είναι το δεύτερο σύνολο δεδομένων που περιέχει ηχογραφήσεις από πολλαπλές πηγές. Είναι μέρος του προγράμματος FreeSound και συνίσταται από 7,690 αρχεία ήχου. Οι συνθήκες των ηχογραφήσεων είναι πολύ

διαφορετικές από αυτές του Warblr dataset αφού είναι από παγκόσμια εμβέλεια και οι συσκευές παρακολούθησης είναι συνήθως μη καταγεγραμμένες, είτε μέσης είτε υψηλής ποιότητας.

- **BirdVox-DCASE-20k dataset** [35] : Το συγκεκριμένο σύνολο δεδομένων ανήκει στην κατηγορία δεδομένων που αντλήθηκαν από ελεγχόμενες συσκευές. Έξι αισθητήρες τοποθετημένοι στο Ίθακα της Νέας Υόρκης τον μήνα Σεπτέμβριο έλαβαν τα δεδομένα εκ των οποίων το 50,09% περιέχει τουλάχιστον ένα πουλί. Το σύνολο αποτελείται από 20,000 αρχεία ήχου των 10 δευτερολέπτων και ανήκει στον BirdVox project.

Για κάθε ένα από τα παραπάνω σύνολα δεδομένων δόθηκε ένα αρχείο .csv με τις ετικέτες των ηχογραφήσεων. Το εν λόγω αρχείο περιείχε το id, την ετικέτα («0», δεν περιέχει πουλί και «1», περιέχει πουλί) και την πηγή (dataset) στο οποίο ανήκει η ηχογράφηση. Οι ετικέτες για το κάθε αρχείο ήχου δόθηκαν χειροκίνητα οπότε και είναι λογικό να υπάρχει κάποιο μικρό ποσοστό λάθους. Το BirdVox-DCASE-20k dataset λέγεται ότι έχει 99.5% ακρίβεια στις ετικέτες και τα υπόλοιπα δύο 96.7% [1].

Για τους σκοπούς τις εργασίας μας αντλήσαμε το 15% του κάθε συνόλου με τυχαία επιλογή. Από αυτό, περίπου το 20% των αρχείων χρησιμοποιήθηκε για την αξιολόγηση του συστήματος. Στον πίνακα 3.1 φαίνεται ο αριθμός των δεδομένων που αντλήσαμε από κάθε σύνολο δεδομένων για εκπαίδευση και αξιολόγηση.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k
Εκπαίδευση	807	840	2100
Δοκιμή	346	360	900
Σύνολο	1054	1200	3000

Πίνακας 3.1: Ποσό δεδομένων που αντλήθηκε από κάθε dataset.

3.4 Μετρική Αξιολόγησης

Στόχος του διαγωνισμού είναι η σχεδίαση ενός γενικού συστήματος όπου θα διακρίνει με ακρίβεια την παρουσία ή απουσία πουλιών σε ένα αρχείο ήχου. Επιπλέον, εφόσον το θέμα του διαγωνισμού δεν αφορούσε συγκεκριμένη εφαρμογή, οι δημιουργοί θέλησαν να εστιάσουν στο κόστος που δημιουργείται από τις ψευδώς θετικές προς τις ψευδώς αρνητικές προβλέψεις (false positive/ false negative). Για τον λόγο αυτό, η έξοδος/πρόβλεψη του συστήματος προτάθηκε να είναι πραγματική τιμή στο διάστημα [0,1] και η μετρική απόδοσης να είναι η Area Under the ROC Curve (AUC).

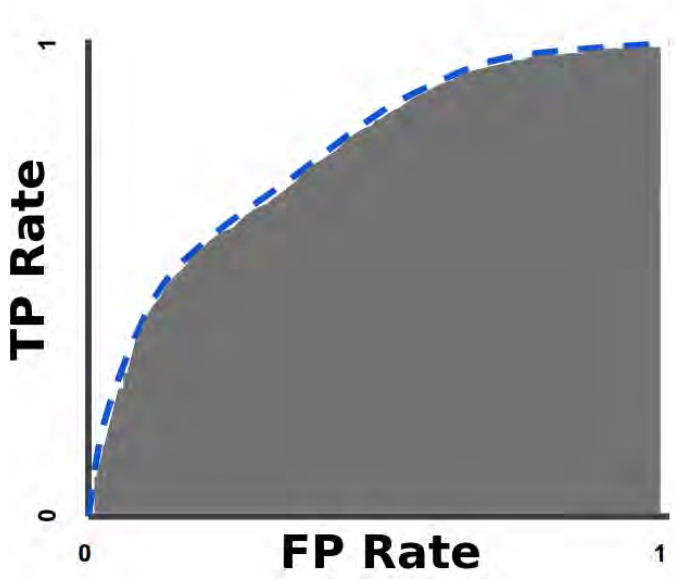
Η AUC μετράει το κατά πόσο ένας ταξινομητής διαχωρίζει επαρκώς δύο κλάσεις. Είναι βασισμένη στη ROC Curve η οποία σχεδιάζεται σύμφωνα με το ποσοστό των σωστά θετικών προβλέψεων TPR ως προς το ποσοστό των λανθασμένα θετικών FPR όπου [36]:

$$TPR = \frac{TruePositive}{TruePositive + FalseNegative} \tag{3.6}$$

$$FPR = \frac{FalsePositive}{TrueNegative + FalsePositive} \tag{3.7}$$

Είναι γνωστό ότι η AUC εφαρμόζει πολλαπλά όρια (thresholds) σε ταξινομητές των οποίων η έξοδος είναι πραγματικός αριθμός. Επιπλέον, κρίνεται κατάλληλη για μη ισορροπημένα σύνολα δεδομένων εφόσον η ανομοιομορφία στα στοιχεία των κλάσεων δεν επιδρά αρνητικά στο αποτέλεσμα. Όλα τα παραπάνω χαρακτηριστικά διακρίνουν την AUC από άλλα μέτρα αξιολόγησης της απόδοσης του ταξινομητή με αποτέλεσμα να επιλέγεται συχνά σε τέτοιου είδους προβλήματα.

Μια τυπική αναπαράσταση της AUC φαίνεται παρακάτω όπου το υποκείμενο ποσοστό λαμβάνεται υπολογίζοντας την περιοχή κάτω από τη ROC καμπύλη (διακεκομμένη μπλε γραμμή).



Σχήμα 3.8: Τυπική αναπαράσταση της AUC. Λήφθηκε από [2].

Κεφάλαιο 4

Ταξινομητές

Στο παρόν κεφάλαιο παραθέτουμε εν συντομία το θεωρητικό υπόβαθρο των ταξινομητών που επιλέξαμε για την εργασία μας. Για την υλοποίηση των μοντέλων επιλέξαμε τη γλώσσα προγραμματισμού Python μιας και χρησιμοποιείται ευρέως για ανάπτυξη αλγορίθμων Μηχανικής Μάθησης. Επιπλέον, παρέχει ένα ευρύ φάσμα πακέτων επεξεργασίας δεδομένων και υλοποιημένων αλγορίθμων Μηχανικής Μάθησης.

4.1 Μηχανές Διανυσματικής Υποστήριξης

Οι Μηχανές Διανυσματικής Υποστήριξης (SVM) αποτελούν κομμάτι των μεθόδων μάθησης υπό επίβλεψη. Χρησιμοποιούνται σε προβλήματα ταξινόμησης, παλινδρόμησης και ανίχνευσης ακραίων τιμών και η κύρια ιδέα τους είναι ο γραμμικός διαχωρισμός ενός συνόλου δεδομένων τα οποία αρχικά έχουν χαρτογραφηθεί σε έναν προεπιλεγμένο χώρο Z. Τα δεδομένα χαρτογραφούνται συνήθως με κάποια μη γραμμική συνάρτηση και η επιφάνεια που τα διαχωρίζει στον χώρο Z στοχεύει στην υψηλή ικανότητα γενίκευσης του δικτύου [3]. Η συνάρτηση που παράγει το μέγιστο περιθώριο μεταξύ δύο κλάσεων ορίζεται ως το βέλτιστο υπερεπίπεδο που διαχωρίζει τις δύο κλάσεις όπως φαίνεται και στο σχήμα 4.1. Ένας μικρός αριθμός δεδομένων επιλέγεται ως διανύσματα υποστήριξης, και αυτά είναι αρκετά για να ορίζουν το παραπάνω υπερεπίπεδο.

Έστω ότι x_i για $i = 1, ..., N$ είναι τα διανύσματα εκπαίδευσης του συνόλου X και $g(x) = w^T x + w_0$ το υπερεπίπεδο διαχωρισμού. Για δύο κλάσεις ω_1, ω_2 όπου

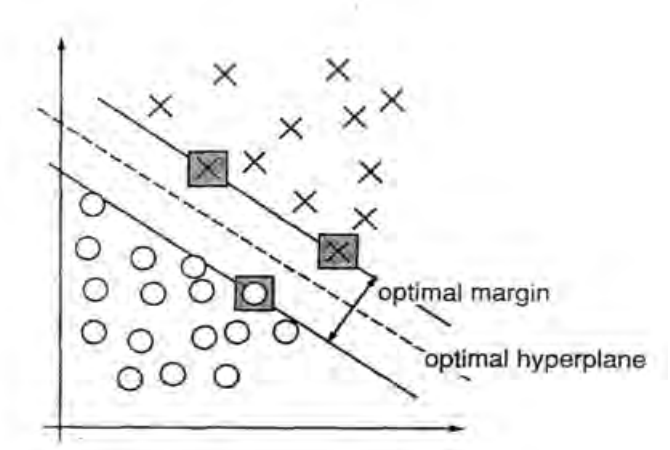
$$g(x) = +1 \text{ για } \omega_1$$
$$\text{και}$$
$$g(x) = -1 \text{ για } \omega_2$$

η απόσταση ενός σημείου x_i από το υπερεπίπεδο διαχωρισμού ορίζεται ως

$$z_x = \frac{|g(x)|}{\|w\|}$$

και το μέγιστο περιθώριο που ορίζει το βέλτιστο υπερεπίπεδο υπολογίζεται ως

$$\frac{2}{\|w\|}$$



Σχήμα 4.1: Διαχωρίσιμο πρόβλημα στον δισδιάστατο χώρο. Τα γκρι διανύσματα υποστήριξης δημιουργούν το μέγιστο δυνατό περιθώριο μεταξύ των δύο κλάσεων [3].

Επιθυμούμε να ελαχιστοποιήσουμε τη συνάρτηση

$$J(w) = \frac{1}{2}\|w\|^2 \tag{4.1}$$

υπό τους ακόλουθους περιορισμούς

$$y_i(w^T x_i + w_0) \geq 1, i = 1, ..., N \tag{4.2}$$

$$y_i = 1, x_i \in \omega_1 \tag{4.3}$$

$$y_i = -1, x_i \in \omega_2 \tag{4.4}$$

Το παραπάνω πρόβλημα ελαχιστοποιεί τη νόρμα $\|w\|$ και μεγιστοποιεί το περιθώριο $\frac{2}{\|w\|}$ συνεπώς είναι ένα τετραγωνικό πρόβλημα βελτιστοποίησης με γραμμικούς περιορισμούς ανισοτήτων. Με χρήση της Lagrange συνάρτησης με συντελεστές λ_i και της θεωρίας βελτιστοποίησης κυρτών προβλημάτων έχουμε:

$$\text{Lagrangian: } \mathcal{L}(w, w_0, \lambda) = \frac{1}{2}w^T w - \sum_{i=1}^{N} \lambda_i [y_i(w^T x_i + w_0) - 1] \tag{4.5}$$

$$\sum_{i=i}^{N} \lambda_i y_i x_i = w \tag{4.6}$$

$$\sum_{i=i}^{N} \lambda_i y_i = 0 \tag{4.7}$$

όπου λύνοντας την εξίσωση 4.5 λαμβάνουμε το υπερεπίπεδο w (εξίσωση 4.6) το οποίο είναι γραμμικός συνδυασμός από $N_1 < N$ διανύσματα υποστήριξης για τα οποία ισχύει $\lambda_i \neq 0$.

Με βάση το παραπάνω πρόβλημα το αντίστοιχο δυϊκό του είναι:

$$\begin{aligned} \underset{\lambda}{\text{maximize}} \quad & \Big(\sum_{i=i}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j \Big) \\ \text{subject to} \quad & \sum_{i=i}^{N} \lambda_i y_i = 0 \\ & \lambda \geq 0 \end{aligned} \tag{4.8}$$

από το οποίο υπολογίζονται οι βέλτιστοι πολλαπλασιαστές Lagrange έτσι ώστε να ληφθεί το βέλτιστο υπερεπίπεδο από την εξίσωση 4.6. Για περισσότερες λεπτομέρειες σχετικά με την επίλυση του παραπάνω προβλήματος αναφερόμαστε στο [3].

Αν και οι Μηχανές Διανυσματικής Υποστήριξης αρχικά προτάθηκαν για γραμμική ταξινόμηση με κάποιες διαφοροποιήσεις μπορούν να εκτελέσουν και μη γραμμική ταξινόμηση. Επιλέγοντας κατάλληλες συναρτήσεις πυρήνα αντιστοιχίζουν τα χαρακτηριστικά από τον χώρο εισόδου σε υψηλών διαστάσεων χώρους στους οποίους είναι δυνατή η γραμμική ταξινόμηση. Συνηθισμένες συναρτήσεις πυρήνα είναι οι:

- Γκαουσιανή Ακτινική Συνάρτηση Βάσης (Gaussian radial basis function)
- Πολυωνυμική Συνάρτηση (Polynomial)
- Υπερβολική Εφαπτομένη (Hyperbolic tangent)

Με την εφαρμογή των παραπάνω συναρτήσεων πυρήνα τις οποίες συμβολίζουμε ως $K(x, x^{'})$, το πρόβλημα βελτιστοποίησης της εξίσωσης (4.8) γίνεται [37]:

$$\begin{aligned} \underset{\lambda}{\text{maximize}} \quad & \Big(\sum_{i=i}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j K(x, x^{'}) \Big) \\ \text{subject to} \quad & \sum_{i=i}^{N} \lambda_i y_i = 0 \\ & \lambda \geq 0 \end{aligned} \tag{4.9}$$

και η λύση του με βάση την εξίσωση (4.6) είναι [37]:

$$w = \sum_{i=i}^{N} \lambda_i y_i K(x, x^{'}) \tag{4.10}$$

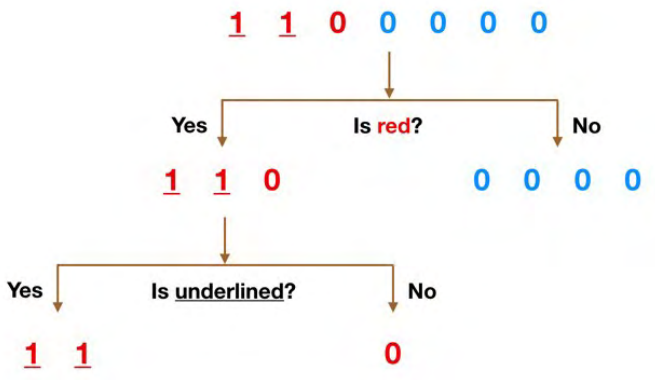
4.1.1 Λεπτομέρειες Υλοποίησης SVM

Για την ανάπτυξη του SVM μοντέλου επιλέξαμε την υλοποίηση της βιβλιοθήκης Sklearn [38] η οποία βασίζεται στη βιβλιοθήκη Libsvm. Συγκεκριμένα χρησιμοποιήσαμε την κλάση SVC σε συνδυασμό με τον RBF kernel και τις προκαθορισμένες ρυθμίσεις της βιβλιοθήκης.

4.2 Τυχαία Δέντρα Αποφάσεων

Τα Τυχαία Δέντρα αποφάσεων (Random Forests) είναι μία τεχνική μάθησης όπου αποτελείται από έναν αριθμό δέντρων αποφάσεων (Decision Trees) τα οποία λειτουργούν σαν σύνολο. Χρησιμοποιείται τόσο σε προβλήματα ταξινόμησης όσο και σε παλινδρόμησης. Στα προβλήματα ταξινόμησης κάθε ένα από τα δέντρα απόφασης που αποτελούν το Random Forest κάνει μία πρόβλεψη και η τελική απόφαση προέρχεται από τον μέσο όρο των μοναδικών μοντέλων-δέντρων. Τα Random Forests είναι μια παραλλαγή της μεθόδου bagging η οποία είναι μια τεχνική που προσπαθεί να ελαχιστοποιήσει την απόκλιση μιας συνάρτησης πρόβλεψης με το να υπολογίζει κατά μέσο όρο τα μοναδικά μοντέλα-δέντρα [37].

Όπως αναφέρθηκε, τα Τυχαία Δέντρα αποφάσεων αποτελούνται από πολλά Δέντρα Αποφάσεων, γι αυτό θα αναφέρουμε περιληπτικά τη βασική ιδέα των τελευταίων. Συγκεκριμένα, ένα Δέντρο Απόφασης είναι ένα μοντέλο όπου κάθε κόμβος κατευθύνει τα δεδομένα εισόδου προς έναν κόμβο-φύλλο το οποίο αντιστοιχεί σε μία ετικέτα. Η κατεύθυνση του κάθε κόμβου ορίζεται από τον διαχωρισμό του χώρου εισόδου είτε από ένα προκαθορισμένο σύνολο ερωτήσεων είτε με βάση τα χαρακτηριστικά των δεδομένων εισόδου [39]. Ας υποθέσουμε το παράδειγμα τις εικόνας 4.2 όπου τα δεδομένα εισόδου αποτελούνται από μία αλληλουχία κόκκινων και μπλε αριθμών 0 ή 1. Η πρώτη ερώτηση αφορά το χρώμα το οποίο είναι ένα χαρακτηριστικό που εκτελεί τον πρώτο διαχωρισμό με σαφήνεια. Στη συνέχεια, εφόσον απαντηθεί «Ναι» στην ερώτηση «Είναι κόκκινο;», ορίζεται η δεύτερη που αφορά την υπογράμμιση των στοιχείων. Αυτή είναι και η τελευταία ερώτηση αφού τα δεδομένα εισόδου έχουν διαχωριστεί σε κάθε φύλλο με επιτυχία.

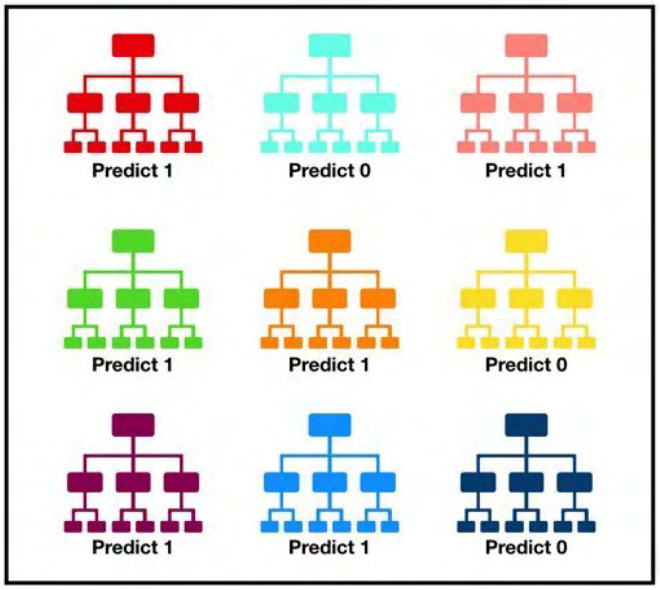


Σχήμα 4.2: Παράδειγμα Δέντρου Απόφασης. Λήφθηκε από [4].

Είναι γνωστό ότι τα Δέντρα Αποφάσεων υποφέρουν από το πρόβλημα της υπερμοντελοποίησης (overfitting) και για να αντιμετωπιστεί αυτό εφαρμόστηκαν μέθοδοι συνόλων. Τέτοιου είδους τεχνικές δημιουργούν μοντέλα τα οποία αποτελούνται από ένα σύνολο από Δέντρα Αποφάσεων που λειτουργούν συλλογικά. Μια συνηθισμένη τεχνική που ανήκει σε αυτή την κατηγορία είναι η Boosted Trees με την οποία σταδιακά χτίζεται το συνολικό μοντέλο από ασθενείς ταξινομητές και εκμεταλλευόμενοι τα λανθασμένα ταξινομημένα δεδομένα. Επιπλέον, ειδική περίπτωση μεθόδων συνόλων αποτελούν τα Τυχαία Δέντρα Αποφάσεων τα οποία ανήκουν στην κατηγορία μεθόδων Bagging ή Bootstrap aggregated. Οι συγκεκριμένες τεχνικές λαμβάνουν ένα ποσοστό των δεδομένων σε κάθε βήμα και χτίζουν

ένα Δέντρο Αποφάσεων. Η επιλογή των δεδομένων γίνεται με αντικατάσταση, δηλαδή ένα στοιχείο μπορεί να επιλεχθεί περισσότερες από μια φορές στα μεμονωμένα υποσύνολα [37].

Σύμφωνα με την παραπάνω θεωρία, στα Τυχαία Δέντρα Αποφάσεων το κάθε Δέντρο Αποφάσεων κάνει μια πρόβλεψη καταγράφοντας πολύπλοκες δομές των δεδομένων εισόδου. Για τα προβλήματα ταξινόμησης, η τελική απόφαση του Τυχαίου Δέντρου Αποφάσεων υπολογίζεται από την πλειοψηφία των μεμονωμένων προβλέψεων. Η παρακάτω εικόνα παρουσιάζει ένα Τυχαίο Δέντρο Αποφάσεων όπου κάθε Δέντρο Απόφασης έχει προβλέψει είτε την κλάση «1» είτε την κλάση «0». Συγκεκριμένα υπάρχουν 6 προβλέψεις για την κλάση «1» και 3 για την κλάση «0», οπότε και η πρόβλεψη του Τυχαίου Δέντρου Αποφάσεων με βάση την πλειοψηφία είναι η κλάση «1».



Σχήμα 4.3: Παράδειγμα Τυχαίου Δέντρου Αποφάσεων. Λήφθηκε από [4].

4.2.1 Υλοποίηση Τυχαίου Δέντρου Αποφάσεων

Η βιβλιοθήκη Sklearn παρέχει ένα σύνολο υλοποιημένων μεθόδων συνόλου. Σε αυτή την κατηγορία υλοποιείται και ο ταξινομητής Τυχαίων Δέντρων Αποφάσεων από την κλάση RandomForestClassifier. Με βάση αυτή και τις προτεινόμενες ρυθμίσεις, χτίσαμε ένα Τυχαίο Δέντρο Αποφάσεων με 300 εκτιμητές - Δέντρα Αποφάσεων. Επιπλέον, θέσαμε την παράμετρο bootstrap=True έτσι ώστε η επιλογή των δειγμάτων να γίνεται με αντικατάσταση.

Κεφάλαιο 5

Συγκρίσεις και Αποτελέσματα

Σε αυτό το κεφάλαιο θα παρουσιάσουμε τα αποτελέσματα της Ανίχνευσης Πουλιών στα αρχεία ήχου. Όπως αναφέραμε, εξερευνήσαμε διαφορετικών ειδών χαρακτηριστικά χρησιμοποιώντας δύο πολύ γνωστούς ταξινομητές. Παρακάτω θα αναλυθούν αποτελέσματα και θα επισημανθεί το ποσοστό κατά το οποίο κάθε τεχνική βελτιώνει ή όχι το μοντέλο. Τα υποκεφάλαια αφορούν τους ταξινομητές και σε κάθε ένα από αυτά θα παρουσιαστούν τα αποτελέσματα των τεχνικών και των χαρακτηριστικών που δοκιμάστηκαν.

5.1 Ταξινομητής SVM

Ως baseline σύστημα θα θεωρήσουμε τον ταξινομητή και τα χαρακτηριστικά με μόνη προεπεξεργασία την κανονικοποίηση των δεδομένων όπως αναφέρθηκε στην ενότητα 3.2.2. Με βάση αυτό θα αναλύσουμε την συμπεριφορά των επιλεγμένων μεθόδων παρουσιάζοντας τα αποτελέσματα σε κάθε περίπτωση. Υπενθυμίζουμε ότι τα χαρακτηριστικά CENS, Chroma CQT και OpenSmile MFCC εφαρμόζουν διαφορετική κανονικοποίηση όπως αυτή κρίθηκε από τους κατασκευαστές των αντίστοιχων μεθόδων. Συγκεκριμένα για τα χαρακτηριστικά CQT δοκιμάστηκαν και άλλες μορφές κανονικοποίησης όπως $\ell 2$, $\ell 1$ νόρμες χωρίς βέβαια να υπάρχει βελτίωση στο αποτέλεσμα.

Εξετάζοντας όλα τα χαρακτηριστικά όπως αυτά ορίστηκαν στην ενότητα 3.4, η AUC όπως και το Harmonic Mean κυμαίνονται μεταξύ των τιμών 47-60%. Ο πίνακας 5.1 παρουσιάζει τα αποτελέσματα για κάθε τύπο χαρακτηριστικών. Παρά τις ιδιότητες των χαρακτηριστικών που περιγράψαμε στην ενότητα 3.1, κανένα από αυτά δεν επιφέρει ικανοποιητικά αποτελέσματα χωρίς περαιτέρω προεπεξεργασία. Συγκεκριμένα για τα χαρακτηριστικά που ανήκουν στην κατηγορία Chroma τα ποσοστά τις AUC είναι ιδιαίτερα χαμηλά. Θα χρειαστεί λοιπόν επιπλέον εξερεύνηση για να διαπιστωθεί εάν αποτελούν κατάλληλα χαρακτηριστικά για την αναπαράσταση ήχων πουλιών.

Όσον αφορά την απόδοση των χαρακτηριστικών του OpenSmile Toolkit παρατηρούμε ότι αυτά δεν ανταποκρίνονται καλά με τον επιλεγμένο ταξινομητή. Για αυτό τον λόγο θα προσπαθήσουμε να εφαρμόσουμε κάποιες από τις μεθόδους προεπεξεργασίας που επιλέξαμε όπου αυτό είναι δυνατό. Επιπλέον, παρόλο που τα MFCC χαρακτηριστικά είναι ειδικά σχεδιασμένα και επεξεργασμένα, δεν παρουσίασαν κάποιο ενδιαφέρον στα αποτελέσματα τις παρούσας διπλωματικής για αυτό τον λόγο δε θα αναλυθούν περαιτέρω.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA CQT	Train	Train	Test	52.1	50.04
	Train	Test	Train	49.66	
	Test	Train	Train	51.42	
CHROMA CENS	Train	Train	Test	54.19	50.03
	Train	Test	Train	48.23	
	Test	Train	Train	48,13	
MFCC	Train	Train	Test	48.84	52.06
	Train	Test	Train	49.07	
	Test	Train	Train	59.63	
OPENSMILE CHROMA	Train	Train	Test	56.88	49.74
	Train	Test	Train	46.42	
	Test	Train	Train	47.18	
OPENSMILE MFCC	Train	Train	Test	54.29	55.86
	Train	Test	Train	53.61	
	Test	Train	Train	60.13	

Πίνακας 5.1: Αποτελέσματα των SVM ταξινομητών που ορίστηκαν ως baseline συστήματα σε συνδυασμό με τα υπό εξερεύνηση χαρακτηριστικά.

5.1.1 Εφαρμογή φίλτρου

Η χρήση του φίλτρου εφαρμόστηκε με σκοπό να ενισχύσει τις συχνότητες που αφορούν ήχους πουλιών και να αποκόψει χαμηλές συχνότητες οι οποίες εισήγαγαν κάποια μετατόπιση και συνεπώς θόρυβο στο σήμα. Αξιοσημείωτη είναι η βελτίωση που επιφέρει το εν λόγω φίλτρο τόσο στα Chroma όσο και στα MFCC χαρακτηριστικά. Συγκεκριμένα στα πρώτα παρατηρείται 45% βελτίωση στην AUC και περίπου 38% στα MFCC. Συνεπώς, το φίλτρο είναι μια αρκετά καλή τεχνική προεπεξεργασίας των δεδομένων αφού από τα ευρήματα δείχνει να βελτιώνει την ταξινόμηση ακόμη και σε συνθήκες άγνωστες των συνθηκών εκπαίδευσης.

Στο σημείο αυτό είναι φανερή η βελτίωση της απόδοσης των χαρακτηριστικών της κατηγορίας Chroma. Ιδιαίτερα μόλις αφαιρέθηκαν στοιχεία θορύβου από τα σήματα που περιέχουν ήχους πουλιών, το ποσοστό της AUC αυξήθηκε κατακόρυφα προσεγγίζοντας την ιδανική ταξινόμηση.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA CQT	Train	Train	Test	92.94	96.99
	Train	Test	Train	98.64	
	Test	Train	Train	99.66	

Πίνακας 5.2: Αποτελέσματα της εφαρμογής φίλτρου στα Chroma CQT χαρακτηριστικά.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA CENS	Train	Train	Test	97.28	98.56
	Train	Test	Train	99.49	
	Test	Train	Train	98.93	

Πίνακας 5.3: Αποτελέσματα της εφαρμογής φίλτρου στα CENS χαρακτηριστικά.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
MFCC	Train	Train	Test	99.75	90.3
	Train	Test	Train	77.74	
	Test	Train	Train	96.76	

Πίνακας 5.4: Αποτελέσματα της εφαρμογής φίλτρου στα MFCC χαρακτηριστικά.

5.1.2 Αντιστοίχιση σε Κανονική Μορφή

Η αντιστοίχιση των δεδομένων στην κανονική μορφή είναι μία τεχνική που εμπνευστήκαμε από τη δημοσίευση [18]. Σύμφωνα με αυτή, η εν λόγω προεπεξεργασία βοηθά στη μείωση του θορύβου και της επίδρασης του καναλιού στο σήμα. Παρακάτω φαίνονται τα αποτελέσματα της αντιστοίχισης στην κανονική μορφή. Η επίδραση της τεχνικής είναι θετική τόσο στα MFCC (Librosa) όσο και στα Chroma (OpenSmile) χαρακτηριστικά. Ωστόσο, στα Chroma CQT δε φέρει βελτίωση αντιθέτως μειώνει το ποσοστό της AUC. Συγκεκριμένα για τα δύο πρώτα έχουμε περίπου 0.3% και 1.2% βελτίωση αντίστοιχα και στα Chroma CQT περίπου 0.5% μείωση.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
MFCC	Train	Train	Test	49.35	52.27
	Train	Test	Train	49.07	
	Test	Train	Train	59.65	

Πίνακας 5.5: Αποτελέσματα της αντιστοίχισης στην Κανονική Κατανομή στα MFCC χαρακτηριστικά.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA CQT	Train	Train	Test	52.99	49.46
	Train	Test	Train	46.61	
	Test	Train	Train	49.19	

Πίνακας 5.6: Αποτελέσματα της αντιστοίχισης στην Κανονική Κατανομή στα Chroma CQT χαρακτηριστικά.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
OPENSMILE CHROMA	Train	Train	Test	52.03	50.97
	Train	Test	Train	50.48	
	Test	Train	Train	50.42	

Πίνακας 5.7: Αποτελέσματα της αντιστοίχισης στην Κανονική Κατανομή στα Chroma χαρακτηριστικά του OpenSmile Toolkit.

5.1.3 Συνδυαστική Υλοποίηση

Σε αυτή την υποενότητα θα παρουσιάσουμε τα αποτελέσματα του SVM ταξινομητή και των χαρακτηριστικών όπου έχουν υποστεί με την σειρά την παρακάτω επεξεργασία:

- Εφαρμογή υψιπερατού φίλτρου στο σήμα χρόνου.
- Εξαγωγή χαρακτηριστικών.
- Κανονικοποίηση (Zero Mean, Unit Variance)
- Μετατροπή τιμών ώστε να ακολουθούν την κανονική κατανομή.

Τα αποτελέσματα Chroma του Opensmile Toolkit παραλείπονται αφού η μόνη επεξεργασία που εφαρμόσαμε αναφέρεται στην ενότητα 5.2.1. Παρόμοια, για τα CENS χαρακτηριστικά η τελική επίδραση της προεπεξεργασίας τους παρατίθεται στον πίνακα 5.3.

Όπως περιγράψαμε και στις προηγούμενες υποενότητες η κάθε μία τεχνική προεπεξεργασίας επιφέρει κυρίως βελτίωση στο αποτέλεσμα του μοντέλου. Παρόλα αυτά, η συνδυαστική εφαρμογή τους επιφέρει μείωση τόσο στα Chroma CQT όσο και στα MFCC στα οποία η AUC βελτιωνόταν σε κάθε περίπτωση. Συγκεκριμένα παρατηρούμε περίπου 20% μείωση στα αποτελέσματα των χαρακτηριστικών Chroma CQT και περίπου 1% στα MFCC σε σύγκριση με τη μέχρι τώρα καλύτερη επίδοση που λάβαμε από την εφαρμογή του φίλτρου.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA CQT	Train	Train	Test	46.04	70.08
	Train	Test	Train	94.52	
	Test	Train	Train	95.12	
MFCC	Train	Train	Test	99.25	89.20
	Train	Test	Train	75.28	
	Test	Train	Train	97.35	

Πίνακας 5.8: Αποτελέσματα των SVM ταξινομητών μετά από συνδυασμό των μεθόδων προεπεξεργασίας.

5.2 Ταξινομητής Random Forest

Όπως και με τον SVM έτσι και με τον Random Forest ταξινομητή παρουσιάζουμε τα αποτελέσματα του baseline συστήματος. Φανερά η απόδοση του Random Forest υστερεί αυτής του SVM ταξινομητή.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA CQT	Train	Train	Test	50.67	47.02
	Train	Test	Train	42.24	
	Test	Train	Train	49.05	
CHROMA CENS	Train	Train	Test	49.54	47.53
	Train	Test	Train	45.62	
	Test	Train	Train	47.60	
MFCC	Train	Train	Test	48.78	50.45
	Train	Test	Train	53.11	
	Test	Train	Train	49.67	
OPENSMILE CHROMA	Train	Train	Test	52.27	47.32
	Train	Test	Train	42.63	
	Test	Train	Train	48.08	
OPENSMILE MFCC	Train	Train	Test	56.7	49.92
	Train	Test	Train	45.18	
	Test	Train	Train	49.21	

Πίνακας 5.9: Harmonic Mean των ορισμένων ως baseline Random Forest ταξινομητών.

5.2.1 Εφαρμογή φίλτρου

Η εφαρμογή του φίλτρου στα δεδομένα ήχου για τον Random Forest ταξινομητή έχει παρόμοια επίδραση με αυτή του SVM. Συγκεκριμένα παρατηρείται 41% αύξηση της AUC για τα χαρακτηριστικά MFCC και περίπου 43% για τα χαρακτηριστικά Chroma CQT και CENS. Και σε αυτή την περίπτωση ταξινομητή διαπιστώνεται η δραματική αύξηση της AUC σε συνθήκες άγνωστες των συνθηκών εκπαίδευσης.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA CQT	Train	Train	Test	77.6	90.25
	Train	Test	Train	97.75	
	Test	Train	Train	98.77	

Πίνακας 5.10: Αποτελέσματα από την εφαρμογή φίλτρου στα Chroma CQT χαρακτηριστικά.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
MFCC	Train	Train	Test	99.66	91.73
	Train	Test	Train	82.38	
	Test	Train	Train	94.95	

Πίνακας 5.11: Αποτελέσματα από την εφαρμογή φίλτρου στα MFCC χαρακτηριστικά.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA CENS	Train	Train	Test	82.58	92.43
	Train	Test	Train	98.47	
	Test	Train	Train	98.12	

Πίνακας 5.12: Αποτελέσματα από την εφαρμογή φίλτρου στα CENS χαρακτηριστικά.

5.2.2 Αντιστοίχιση σε Κανονική Μορφή

Η αντιστοίχηση των χαρακτηριστικών στην κανονική μορφή έφερε θετικά αποτελέσματα σε συνδυασμό με τον ταξινομητή Random Forest. Συγκεκριμένα, βελτίωσε κατά 1-3% το Harmonic Mean του μοντέλου για τα Chroma CQT και MFCC χαρακτηριστικά αντίστοιχα και 4% στα χαρακτηριστικά OpenSmile Chroma. Συνολικά, η τεχνική αποδίδει καλύτερα σε συνδυασμό με τον ταξινομητή Random Forest αφού μεμονωμένα αυξάνει το ποσοστό της AUC του κάθε σεναρίου εκπαίδευσης (2 datasets για εκπαίδευση και 1 για αξιολόγηση) κατά 1-7 %.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
MFCC	Train	Train	Test	49.74	51.7
	Train	Test	Train	48.9	
	Test	Train	Train	57.24	

Πίνακας 5.13: Αποτελέσματα αντιστοίχισης στην Κανονική Κατανομή στα MFCC χαρακτηριστικά.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA CQT	Train	Train	Test	51.83	50.18
	Train	Test	Train	50.1	
	Test	Train	Train	48.74	

Πίνακας 5.14: Αποτελέσματα αντιστοίχισης στην Κανονική Κατανομή στα Chroma CQT χαρακτηριστικά.

	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
OPENSMILE	Train	Train	Test	52.42	
CHROMA	Train	Test	Train	51.48	51.02
	Test	Train	Train	49.25	

Πίνακας 5.15: Αποτελέσματα αντιστοίχισης στην Κανονική Κατανομή στα Chroma χαρακτηριστικά του OpenSmile Toolkit.

5.2.3 Συνδυασμός μεθόδων

Ο συνδυασμός των μεθόδων προεπεξεργασίας που επιλέξαμε να ερευνήσουμε δεν έφερε θετικά αποτελέσματα στα μοντέλα ταξινομητών. Βασισμένοι στην καλύτερη επίδοση που παρουσιάζεται έως τώρα στην ενότητα 5.2.1 (εφαρμογή φίλτρου) η μείωση είναι φανερή. Συγκεκριμένα, όπως στον ταξινομητή SVM έτσι και στον Random Forest τα ποσοστά της AUC μειώθηκαν κατά 1-20% για τα Chroma CQT, ενώ για τα MFCC παρατηρούμε αύξηση κατά 1% στη μετρική AUC. Η παραπάνω μείωση φαίνεται και στο ποσοστό του Harmonic Mean του μοντέλου για κάθε τύπο χαρακτηριστικών όπου η μείωση αντιστοιχίζεται περίπου στο 13%.

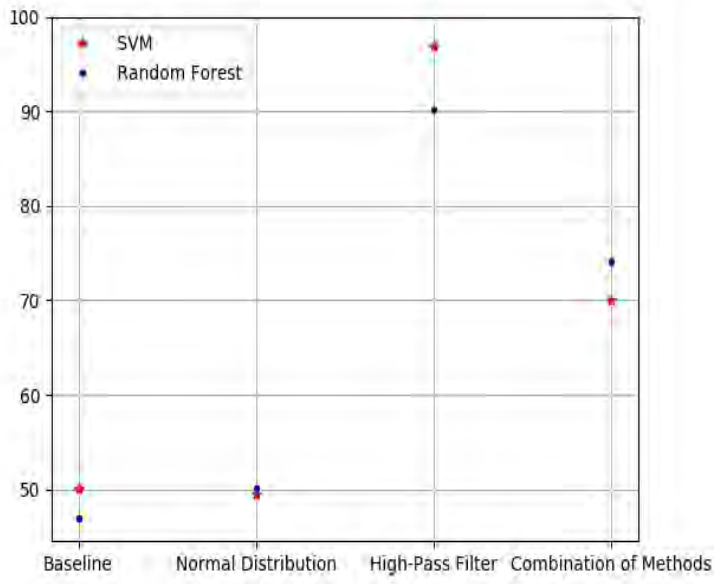
	Freefield1010	Warblrb10k	BirdVox-DCASE-20k	AUC (%)	Harmonic Mean (%)
CHROMA	Train	Train	Test	55.54	
CQT	Train	Test	Train	94.9	74.19
	Test	Train	Train	84.09	
	Train	Train	Test	99.52	
MFCC	Train	Test	Train	83.78	92.27
	Test	Train	Train	94.99	

Πίνακας 5.16: Αποτελέσματα των Random Forest ταξινομητών μετά από συνδυασμό των μεθόδων προεπεξεργασίας.

5.3 Συμπεράσματα

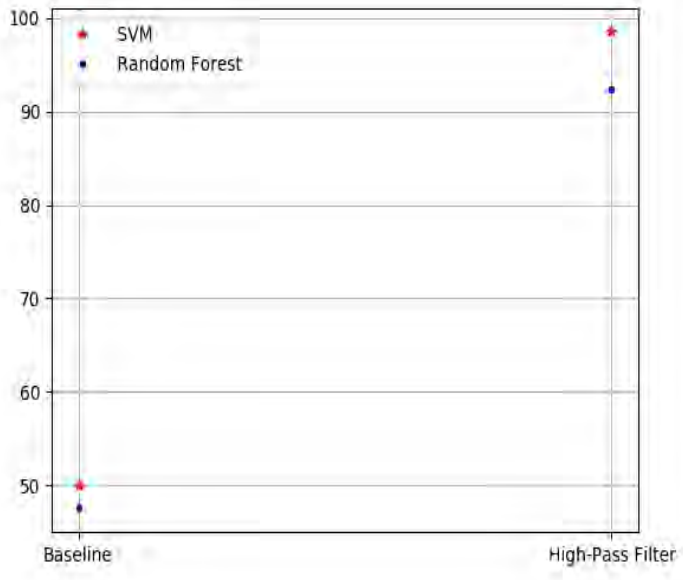
Λαμβάνοντας υπόψιν την παραπάνω ανάλυση και εστιάζοντας στην απόδοση που επιφέρουν τα χαρακτηριστικά Chroma στα μοντέλα των ταξινομητών, στις παρακάτω εικόνες συνοψίζουμε τα αποτελέσματα για την ευκολότερη σύγκρισή τους.

Αναλύοντας το σχήμα 5.1, δεν υπάρχει ξεκάθαρος διαχωρισμός για το ποιος ταξινομητής συμπεριφέρεται καλύτερα με τα χαρακτηριστικά Chroma CQT. Σίγουρα οποιαδήποτε τεχνική βελτιώνει το ποσοστό του Harmonic Mean καθώς και της AUC του κάθε σεναρίου εκπαίδευσης- αξιολόγησης. Παρ΄ όλα αυτά, λαμβάνοντας υπόψιν την καλύτερη απόδοση η οποία είναι αποτέλεσμα της εφαρμογής του φίλτρου και συγκρίνοντας τη με το Baseline σύστημα, ο ταξινομητής που υπερτερεί είναι ο SVM.



Σχήμα 5.1: Σύνοψη ποσοστών Harmonic Mean για τα χαρακτηριστικά Chroma CQT.

Σε αντίθεση με τα χαρακτηριστικά CQT, στα CENS εφαρμόστηκε μόνο το φίλτρο σαν στάδιο προεπεξεργασίας. Η επιλογή αυτή έγινε συνειδητά αφού τα συγκεκριμένα χαρακτηριστικά υφίστανται από τη φύση τους 2 στάδια κανονικοποίησης. Και σε αυτή την περίπτωση, ο ταξινομητής Random Forest υστερεί σε σχέση με τον SVM.



Σχήμα 5.2: Σύνοψη ποσοστών Harmonic Mean για τα χαρακτηριστικά CENS

Με βάση την παραπάνω έρευνα βλέπουμε το ποσοστό επιτυχίας των Chroma χαρακτηριστικών. Συγκεκριμένα, οι κατηγορίες Chroma CENS και Chroma CQT οι οποίες δεν έχουν χρησιμοποιηθεί ιδιαίτερα μέχρι τώρα σε προβλήματα ταξινόμησης, ανταποκρίνονται σε ικανοποιητικό βαθμό στο πρόβλημα της εργασίας μας. Η αρχική μας εκτίμηση για καλύτερη αναπαράσταση των ήχων των πουλιών μέσω των χαρακτηριστικών Chroma επιβεβαιώθηκε και έφερε ποσοστό επιτυχίας 96.99 - 98.56%. Ο ταξινομητής SVM ανταποκρίθηκε καλύτερα στην παρούσα εργασία χωρίς όμως να διαφέρει αρκετά από τα ποσοστά του Random Forest. Τα δύο αυτά είδη χαρακτηριστικών αντιμετώπισαν το πρόβλημα της γενίκευσης χωρίς την εφαρμογή ιδιαίτερα πολύπλοκων μεθόδων, παρά μόνο με την εφαρμογή φίλτρου για απομάκρυνση θορύβου.

Κεφάλαιο 6

Επίλογος και Μελλοντική Δουλειά

Η κοινότητα του DCASE δημιούργησε δύο διαγωνισμούς με κύριο θέμα την ανίχνευση πουλιών σε αρχεία ήχου. Και στις δύο εκδοχές του διαγωνισμού στόχος ήταν η γενίκευση των μοντέλων σε άγνωστες συνθήκες γι αυτό και δόθηκαν διαφορετικά σύνολα δεδομένων με διαφορετικές συνθήκες ηχογράφησης το κάθε ένα. Η προσπάθεια να βρεθεί το κατάλληλο σύστημα με τις κατάλληλες τεχνικές που θα γενίκευε επαρκώς, ανεξαρτήτων των συνθηκών της ηχογράφησης, ήταν αξιοσημείωτη αλλά δεν είχε τα ιδανικά αποτελέσματα. Το ποσοστό επιτυχίας των μοντέλων που προσπαθούν να γενικεύσουν σε άγνωστες συνθήκες είχε σημαντική απόκλιση από αυτών των οποίων οι συνθήκες ήταν ταιριαστές. Γι αυτό τον λόγο, προτάθηκε η εξερεύνηση διαφορετικών ειδών χαρακτηριστικών πράγμα που αποτέλεσε την ιδέα της παρούσας διπλωματικής.

Στην εργασία μας δοκιμάστηκαν διαφόρων ειδών χαρακτηριστικά ήχου σε συνδυασμό με δύο ευρέως γνωστούς ταξινομητές, τον SVM και τον Random Forest. Οι συγκρίσεις αφορούσαν τόσο τη συμπεριφορά των χαρακτηριστικών όσο και των δύο ταξινομητών μεταξύ τους. Επιπλέον, εφαρμόστηκαν μέθοδοι για την απομάκρυνση του θορύβου από τα αρχεία ήχου μέσω φίλτρου και μέσω αντιστοίχισης των τιμών στην ιδανικές κατανομές. Κύρια απαίτηση ήταν η εξερεύνηση νέων χαρακτηριστικών, ικανών να αναπαριστούν καλύτερα μελωδικούς ήχους. Σε αυτή την κατηγορία, ανήκουν τα χαρακτηριστικά Chroma τα οποία είναι στενά συνδεδεμένα με τις 12 κλάσεις $\{C, C\sharp, D, D\sharp, E, F, F\sharp, G, G\sharp, A, A\sharp, B\}$ και αναπαριστούν καλά μελωδικούς και αρμονικούς ήχους. Από τις τεχνικές μείωσης θορύβου μόνο η εφαρμογή του φίλτρου που επιλέχθηκε έφερε θετικά αποτελέσματα, και συγκεκριμένα έφερε 45-50% αύξηση στην μετρική AUC.

Η εξερεύνηση χαρακτηριστικών και μεθόδων απομάκρυνσης θορύβου μας οδήγησαν στην επιλογή των CENS χαρακτηριστικών και του SVM ταξινομητή. Η προεπεξεργασία των εν λόγω χαρακτηριστικών αφορά την εφαρμογή ενός υψιπερατού φίλτρου το οποίο παίζει καθοριστικό ρόλο στην απόδοση του μοντέλου. Συγκεκριμένα, λάβαμε Harmonic Mean ίσο με 98.56% το οποίο περιγράφει τον σταθμισμένο μέσο όρο των τριών διαφορετικών datasets τα οποία χρησιμοποιήθηκαν για αξιολόγηση όταν τα υπόλοιπα δύο χρησιμοποιούνταν για εκπαίδευση.

Σύμφωνα με τα παραπάνω αποτελέσματα, η καλή επίδοση των CENS και Chroma CQT χαρακτηριστικών παρουσιάζει αρκετό ενδιαφέρον για τους σκοπούς του προβλήματος μας. Μόνο η εφαρμογή του φίλτρου η οποία απομακρύνει στοιχεία θορύβου δίνει στα CENS χαρακτηριστικά τη δυνατότητα να επιτυγχάνουν υψηλά ποσοστά ακρίβειας ακόμη

και σε συνθήκες άγνωστες από αυτών της εκπαίδευσης. Για τους παραπάνω λόγους αρκετό ενδιαφέρον θα παρουσίαζε η εξερεύνηση των χαρακτηριστικών CENS και CQT με state of the art μεθόδους όπως τα Νευρωνικά Δίκτυα.

Παράρτημα Α

Λεπτομέρειες Υλοποίησης

Σε αυτό το παράρτημα αναφέρονται λεπτομέρειες που αφορούν την υλοποίηση της εργασίας μας.

Α.1 Βιβλιοθήκες

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε είναι η Python 3.6.8 και οι κύριες βιβλιοθήκες είναι οι librosa sklearn scipy numpy και pandas. Στον πίνακα Α.1 φαίνονται οι κύριες βιβλιοθήκες που χρησιμοποιήθηκαν σε αυτή τη διπλωματική καθώς και οι μέθοδοι που φάνηκαν χρήσιμες.

Βιβλιοθήκη	Μέθοδοι
librosa	feature.{mfcc,chroma_cens,chroma_cqt}
scipy	signal.butter
sklearn	preprocessing.{StandardScaler, QuantileTransformer}
numpy	
pandas	

Πίνακας Α.1: Οι βασικές βιβλιοθήκες που χρησιμοποιήθηκαν για την ανάπτυξη των μοντέλων και την επεξεργασία των δεδομένων.

Επιπλέον, άλλα εργαλεία που φάνηκαν χρήσιμα είναι το OpenSmile Toolkit έκδοση 2.3.0 από το οποίο εργαστήκαμε με τα εξής αρχεία ρυθμίσεων:

- MFCC: MFCC12_E_D_A_Z.conf
- Chroma STFT: chroma_fft.conf

Η μετατροπή των αρχείων .htk που παράγει το αρχείο ρυθμίσεων για τα MFCC χαρακτηριστικά έγινε με τη βοήθεια της μεθόδου HTKFile του αρχείου HTK.py [40].

Α.2 Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν είναι διαθέσιμα δημόσια στη σελίδα [41] του

διαγωνισμού. Συγκεκριμένα για τα τρία σύνολα δεδομένων ανάπτυξης παρέχεται και το αντίστοιχο αρχείο .csv το οποίο περιέχει τις ετικέτες των αρχείων ήχου.

Α.3 Δομή Κώδικα

Ο κώδικας αποτελείται από τρία βασικά αρχεία για τα οποία:

- feature_extraction.py: Περιέχει τις μεθόδους εξαγωγής των χαρακτηριστικών καθώς και τη μέθοδο dataset_split() η οποία πραγματοποιεί το subsampling των datasets.
- train_svm.py: Εκπαίδευση του SVM ταξινομητή εκτελώντας τη 3-way cross validation διαδικασία όπως αυτή ορίζεται στη συγκεκριμένη διπλωματική εργασία.
- train_rforest.py: Εκπαίδευση του Random Forest ταξινομητή εκτελώντας τη 3-way cross validation διαδικασία όπως αυτή ορίζεται στη συγκεκριμένη διπλωματική εργασία.

Επιπλέον, το αρχείο run.py εκτελεί μια ολοκληρωμένη διαδικασία για την εργασία μας. Η συγκεκριμένη διαδικασία περιλαμβάνει τη μορφοποίηση των datasets, την παραγωγή των χαρακτηριστικών και την εκπαίδευση των μοντέλων. Τα ορίσματα του αρχείου είναι τα εξής:

- -h : βοήθεια σχετικά με τα ορίσματα του αρχείου.
- -p : βασικό μονοπάτι στο οποίο θα αποθηκευτούν τα αρχεία εισόδου (datasets).
- -f : Επιθυμητά χαρακτηριστικά προς εξαγωγή.
- -c : Επιθυμητός ταξινομητής προς εκπαίδευση.

Για τη σωστή εκτέλεση του παραπάνω αρχείου θα πρέπει:

- Το μονοπάτι path που θα δοθεί σαν είσοδος να περιέχει τους φακέλους
  - audio/
  - labels/

  όπου ο πρώτος φάκελος θα περιέχει τα datasets με τα αρχεία ήχου και στον δεύτερο θα είναι αποθηκευμένα τα αρχεία csv με τις ετικέτες.
- Τα ονόματα των datasets θα πρέπει να είναι της μορφής :
  - /BirdVox-DCASE-20k
  - /warblrb10k
  - /ff1010bird

  και αντίστοιχα των αρχείων ετικετών με την κατάληξη .csv.
- διαθέσιμες επιλογές του -c:

  - opensmile_mfcc
  - opensmile_chroma
  - mfcc
  - cqt
  - cens

- διαθέσιμες επιλογές του -f:
  - svm
  - forest

A.4 Hardware

Η ανάπτυξη και η αξιολόγηση του συστήματος έγιναν σε επεξεργαστή Intel(R) Core(TM) i7-4510U CPU 2.00GHz και μνήμη RAM 8 GB.

Βιβλιογραφία

[1] Dan Stowell, Michael D Wood, Hanna Pamuła, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3):368–380, 2019.

[2] `https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc`.

[3] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[4] `https://towardsdatascience.com/understanding-random-forest-58381e0602d2`.

[5] Dan Stowell, Mike Wood, Yannis Stylianou, and Hervé Glotin. Bird detection in audio: a survey and a challenge. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016.

[6] Thomas Grill and Jan Schlüter. Two convolutional neural networks for bird detection in audio signals. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1764–1768. IEEE, 2017.

[7] Iwona Sobieraj, Qiuqiang Kong, and Mark D Plumbley. Masked non-negative matrix factorization for bird detection using weakly labeled data. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1769–1773. IEEE, 2017.

[8] Vinayak Abrol, Pulkit Sharma, Anshul Thakur, Padmanabhan Rajan, Aroor Dinesh Dileep, and Anil K Sao. Archetypal analysis based sparse convex sequence kernel for bird activity detection. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1774–1778. IEEE, 2017.

[9] Hervé Goëau, Hervé Glotin, Willem-Pier Vellinga, Robert Planqué, Andreas Rauber, and Alexis Joly. Lifeclef bird identification task 2014. Sheffield, United Kingdom, 2014.

[10] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545, 2014.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[12] Qiuqiang Kong, Yong Xu, and Mark D Plumbley. Joint detection and classification convolutional neural network on weakly labelled bird audio detection. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1749–1753. IEEE, 2017.

[13] Thomas Pellegrini. Densely connected CNNs for bird audio detection. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1734–1738. IEEE, 2017.

[14] Peter Jančovič and Münevver Köküer. Automatic detection and recognition of tonal bird sounds in noisy environments. *EURASIP Journal on Advances in Signal Processing*, 2011(1):982936, 2011.

[15] `http://dcase.community/`.

[16] Emre Cakir, Sharath Adavanne, Giambattista Parascandolo, Konstantinos Drossos, and Tuomas Virtanen. Convolutional recurrent neural networks for bird audio detection. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1744–1748. IEEE, 2017.

[17] Sharath Adavanne, Konstantinos Drossos, Emre Çakir, and Tuomas Virtanen. Stacked convolutional and recurrent neural networks for bird audio detection. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1729–1733. IEEE, 2017.

[18] Anshul Thakur, R Jyothi, Padmanabhan Rajan, and Aroor Dinesh Dileep. Rapid bird activity detection using probabilistic sequence kernels. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1754–1758. IEEE, 2017.

[19] Peter Jančovič and Münevver Köküer. Automatic detection of bird species from audio field recordings using HMM-based modelling of frequency tracks. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1779–1783. IEEE, 2017.

[20] Colm Or'Reilly, Münevver Kcöküer, Peter Jančović, Regan Drennan, and Naomi Harte. Automatic frequency feature extraction for bird species delimitation. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1959–1763. IEEE, 2017.

[21] Maria Sandsten and Johan Brynolfsson. Classification of bird song syllables using Wigner-Ville ambiguity function cross-terms. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1739–1743. IEEE, 2017.

[22] Sidrah Liaqat, Narjes Bozorg, Neenu Jose, Patrick Conrey, Antony Tamasi, and Michael T. Johnson. Domain tuning methods for bird audio detection. Technical report, DCASE2018 Challenge, September 2018.

[23] Fabio Vesperini, Leonardo Gabrielli, Emanuele Principi, and Stefano Squartini. A capsule neural networks based approach for bird audio detection. Technical report, DCASE2018 Challenge, September 2018.

[24] Franz Berger, William Freillinger, Paul Primus, and Wolfgang Reisinger. Bird audio detection - DCASE 2018. Technical report, DCASE2018 Challenge, September 2018.

[25] `https://github.com/librosa/librosa`.

[26] Florian Eyben, Martin Wöllmer, and Björn Schuller. OPENSMILE: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[27] Ooi Chia Ai, M Hariharan, Sazali Yaacob, and Lim Sin Chee. Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems with Applications*, 39(2):2157–2165, 2012.

[28] Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. HMM-based audio keyword generation. In *Pacific-Rim Conference on Multimedia*, pages 566–574. Springer, 2004.

[29] Meinard Müller and Sebastian Ewert. Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), 2011.*

[30] Christian Schörkhuber and Anssi Klapuri. Constant-q transform toolbox for music processing. In *7th Sound and Music Computing Conference, Barcelona, Spain*, 2010.

[31] Judith C Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[32] Meinard Müller, Frank Kurth, and Michael Clausen. Audio matching via chroma-based statistical features. In *ISMIR*, 2005.

[33] Mario Lasseck. Acoustic bird detection with deep convolutional neural networks. Technical report, DCASE2018 Challenge, September 2018.

[34] Dan Stowell and Mark D Plumbley. An open dataset for research on audio field recording archives: freefield1010. *arXiv:1309.5275*, 2013.

[35] Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Birdvox-full-night: a dataset and benchmark for avian flight call detection. In *Proc. IEEE ICASSP*, April 2018.

[36] Andrew P Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[37] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[39] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[40] `https://github.com/danijel3/PyHTK`.

[41] `http://dcase.community/challenge2018/task-bird-audio-detection`.