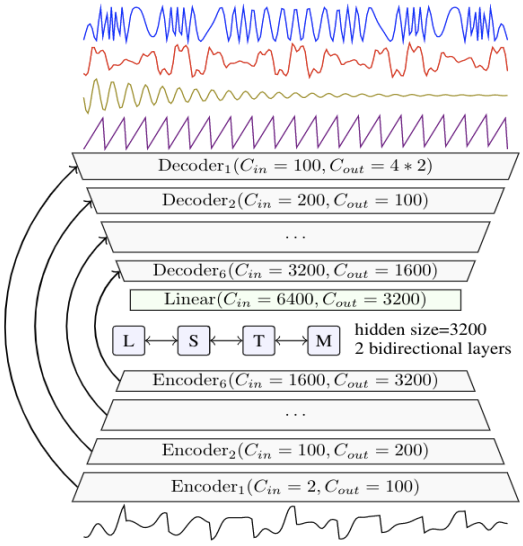
ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

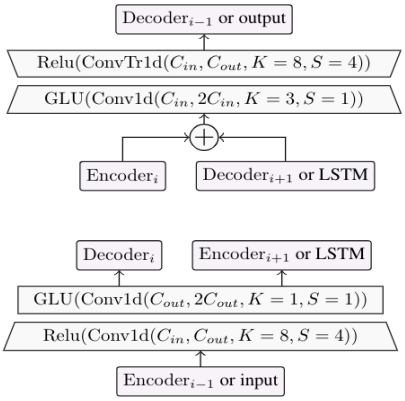
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία

ΔΙΑΧΩΡΙΣΜΟΣ ΜΟΥΣΙΚΩΝ ΣΗΜΑΤΩΝ ΜΕ ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΟΣ





Ονοματεπώνυμο: Μεθενίτης Δημήτριος

ΑΜ: 46094

Επιβλέπων: Βουλόδημος Αθανάσιος

ΑΘΗΝΑ-ΑΙΓΑΛΕΩ

ΟΚΤΩΒΡΙΟΣ 2021

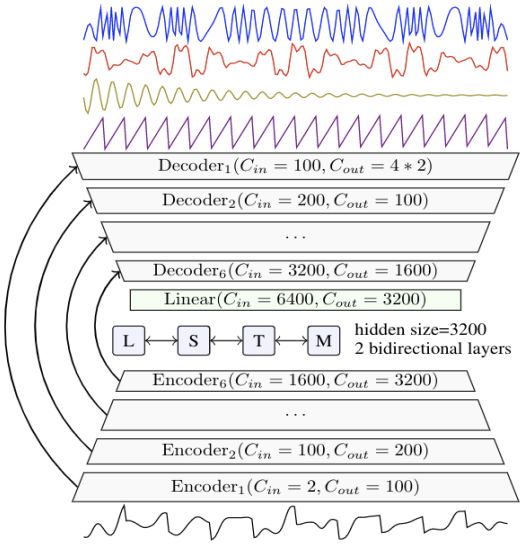
UNIVERSITY OF WEST ATTICA

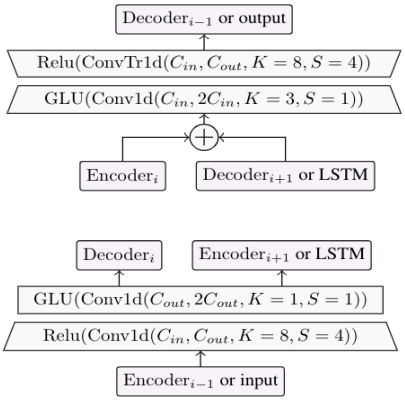
SCHOOL OF ENGINEERING

DEPARTMENT OF INFORMATICS AND COMPUTER ENGINEERING

Diploma Thesis

MUSIC SOURCE SEPARATION USING MACHINE LEARNING AND SIGNAL PROCESSING TECHNIQUES





Name and Surname: Methenitis Dimitrios

Registration Number: 46094

Supervisor: Voulodimos Athanasios

Athens-Aigaleo

October 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΗΜΙΟΥΡΓΙΑ ΜΟΥΣΙΚΩΝ ΣΗΜΑΤΩΝ ΜΕ ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΟΣ

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

Η πτυχιακή/διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή:

Α/α	ΟΝΟΜΑ ΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ/ΙΔΙΟΤΗΤΑ	ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ
1	Αθανάσιος Βουλόδημος	Επίκουρος Καθηγητής	Athanasios Voulodimos Digitally signed by Athanasios Voulodimos Date: 2021.11.11 00:05:05 +02'00'
2	Γεώργιος Μπαρδής	Επίκουρος Καθηγητής	Georgios Bardis Digitally signed by Georgios Bardis Date: 2021.11.11 19:28:24 +02'00'
3	Ευτύχιος Πρωτοπαπαδάκης	Υπότροφος ΕΣΠΑ	Eftychios Protopapadakis Digitally signed by Eftychios Protopapadakis Date: 2021.11.11 11:37:27 +02'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος/η Μεθενίτης Δημήτριος του Θεμιστοκλή, με αριθμό μητρώου 71346094 φοιτητής/τρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο/Η Δηλών/ούσα

Μεθενίτης Δημήριος

(Υπογραφή)

*** Ονοματεπώνυμο /Ιδιότητα**

Ψηφιακή Υπογραφή Επιβλέποντα

Περίληψη

Η διπλωματική μας εργασίας έχει ως αντικείμενο το διαχωρισμό μουσικών σημάτων με τη χρήση τεχνικών μηχανικής μάθησης. Πιο συγκεκριμένα σκοπός μας είναι να συγκρίνουμε και να αξιολογήσουμε μοντέλα μηχανικής μάθησης τα οποία μπορούν να λύσουν το πρόβλημα του διαχωρισμού ενός μουσικού σήματος. Αρχικά αναλύουμε το πρόβλημα αυτό το οποίο σήμερα έχει κερδίσει σε μεγάλο βαθμό το ενδιαφέρον της επιστημονικής κοινότητας. Αναφέρουμε έννοιες όπως το cocktail party problem και το Independent component analysis. Στη συνέχεια αναφερόμαστε στα τέσσερα μοντέλα τα οποία θα τρέξουμε για να πάρουμε αποτελέσματα και να μπορούμε να τα αξιολογήσουμε. Τα μοντέλα αυτά είναι τα Demucs, OpenUnmix, D3Net και Spleeter. Αφού τελειώσουμε με το θεωρητικό μέρος θα περάσουμε στο πειραματικό μέρος όπου θα συγκρίνουμε τα μοντέλα μας. Αρχικά χρησιμοποιώντας ως δεδομένα το MusDb βγάζουμε τέσσερις μετρικές οι οποίες είναι τα SDR, SIR, SAR και ISR. Μετά κάνουμε διαχωρισμό σε τέσσερα μουσικά κομμάτια της επιλογής μας. Ο διαχωρισμός γίνεται σε τέσσερα μέρη τα οποία είναι τα drums, bass, vocals και other. Στη συνέχεια συγκρίνουμε τις τιμές των μετρικών και το διαχωρισμό των μουσικών κομματιών ώστε να περάσουμε στην τελευταία ενότητα της εργασίας μας που είναι τα συμπεράσματα μας. Ως τελικό συμπέρασμα έχουμε ότι το μοντέλο Demucs είναι το καλύτερο σε απόδοση όμως κάθε ένα από τα τέσσερα μοντέλα μπορεί να λύσει το πρόβλημα του music separation αναλόγως τα ζητούμενα της εργασίας.

Abstract

The subject of our diploma thesis is the music source separation using machine learning techniques. More specifically, our goal is to compare and evaluate machine learning models which can solve the problem of signal separation. We first analyze the theory behind the problem of signal separation which today haw greatly gained the interest of the scientific community. We mention concepts such as the cocktail party problem and the Independent component analysis. Furthermore, we refer to four models which we will run to get results and be able to evaluate them. These models are Demucs, OpenUnmix, D3Net and Spleeter. After we finish with the theoretical part we will move on to the experimental part where we will compare our models. Initially using MusDb as our data we get four metrics which are SDR, SIR, SAR and ISR. Next, we separate four songs of our choice. The separation is made into four parts which are the drums, bass, vocals and other. Then we compare the values of the metrics and the separation of the songs in order to move on to the last section of our work which is our conclusion. As our final conclusion we have that the Demucs model is the best in performance but each of the four models can solve the problem of music separation depending on the requirements of the work.

Περιεχόμενα

ΕΙΣΑΓΩΓΗ

Αντικείμενο της διπλωματικής εργασίας

Το κύριο θέμα της εργασίας μας είναι ο διαχωρισμός μουσικών σημάτων με τη βοήθεια της μηχανικής μάθησης και της επεξεργασίας σήματος. Ο διαχωρισμός μουσικών σημάτων μπορεί να βοηθήσει στην παραγωγή ενός νέου τραγουδιού ή μουσικού κομματιού. Πιο συγκεκριμένα, απομονώνουμε διάφορα μουσικά όργανα και στη συνέχεια αναμειγνύοντας τα μπορούμε να έχουμε το επιθυμητό μας αποτέλεσμα. Οι συνιστώσες που μπορεί να έχουμε είναι η κιθάρα, η φωνή, το μπάσο, τα ντραμς και άλλα. Τα τελευταία χρόνια το πρόβλημα του διαχωρισμού μουσικών σημάτων και του τυφλού διαχωρισμού σημάτων έχει προσελκύσει το ενδιαφέρον της ερευνητικής κοινότητας. Για το λόγο αυτό έχουν προταθεί από πολλές ομάδες ανθρώπων διάφορες τεχνικές μηχανικής μάθησης για την επίλυση του προβλήματος αυτού. Λόγω του μεγάλου ενδιαφέροντος που παρουσιάζει το πρόβλημα αυτό θα ασχοληθούμε και εμείς βγάζοντας τα δικά μας αποτελέσματα και συμπεράσματα.

Σκοπός και Στόχοι

Σκοπός της εργασίας μας είναι η μελέτη του προβλήματος του διαχωρισμού μουσικών σημάτων και του τυφλού διαχωρισμού σημάτων. Πιο συγκεκριμένα, θα συγκρίνουμε διάφορα συστήματα διαχωρισμού σημάτων με χρήση τεχνικών μηχανικής μάθησης και επεξεργασίας σήματος. Στόχος μας είναι αρχικά να αναλύσουμε επιμέρους προβλήματα του κεντρικού μας προβλήματος όπως το cocktail party problem και το independent component analysis. Στη συνέχεια θα περάσουμε στο κύριο στόχο της εργασίας που θα είναι να βρούμε ένα σύστημα του οποίου τα αποτελέσματα θα μπορούν να επιλύουν το κεντρικό μας πρόβλημα.

Μεθοδολογία

Όπως αναφέραμε και προηγουμένως θα συγκρίνουμε διάφορα συστήματα διαχωρισμού σημάτων με χρήση τεχνικών μηχανικής μάθησης και επεξεργασίας σήματος. Τα μοντέλα αυτά έχουν γραφτεί σε γλώσσα Python. Εμείς θα τρέξουμε τα διάφορα συστήματα και τις διάφορες τεχνικές που παρουσιάζει το καθένα από αυτά. Προσθέτοντας διάφορες δικές μας μετρικές θα μπορέσουμε να κάνουμε τη σύγκριση και να βγάλουμε το συμπέρασμα για το ποιο σύστημα και ποια τεχνική αποδίδουν καλύτερα. Στην εργασίας μας θα χρησιμοποιήσουμε τέσσερα διαφορετικά μοντέλα που βασίζονται στη μηχανική μάθηση. Τα μοντέλα αυτά έχουν ονόματα όπως

demucs, *OpenUnmix*, *D3Net* και *Spleeter*. Το dataset που θα χρησιμοποιήσουμε και θα εφαρμοστεί στα μοντέλα μας θα είναι ένα δημόσια διαθέσιμο dataset το MusDB.

Καινοτομία

Η εργασία μας θα μπορούσε να χαρακτηριστεί καινοτόμα λόγω του θέματος και του προβλήματος με το οποίο ασχολείται. Το πρόβλημα του διαχωρισμού μουσικών σημάτων έχει προσελκύσει το ενδιαφέρον της ερευνητικής κοινότητας τα τελευταία χρόνια. Για το λόγο αυτό υπάρχουν πολλές προτάσεις από ομάδες με διάφορες τεχνικές μηχανικής μάθησης. Μία από τις προτάσεις αυτές ήρθε από το Facebook AI Research Lab παρουσιάζοντας μία λύση για το πρόβλημα του διαχωρισμού μουσικών σημάτων. Επίσης η μηχανική μάθηση είναι ένας τομέας της τεχνολογίας με τον οποίο ασχολούνται όλο και περισσότεροι προγραμματιστές και μηχανικοί. Αυτό το φαινόμενο συμβαίνει καθώς η μηχανική μάθηση μπορεί να εφαρμοστεί και να παρέχει προτάσεις για λύση πολλών προβλημάτων της καθημερινότητας. Καταλήγουμε λοιπόν στο συμπέρασμα ότι θα ασχοληθούμε με ένα σύγχρονο πρόβλημα το οποίο θα προσπαθήσουμε με την βοήθεια ενός καινοτόμου τρόπου όπως είναι η χρήση τεχνικών μηχανικής μάθησης.

Δομή

Στο 1ο σκέλος της εργασίας μας έχουμε την εισαγωγή στο θέμα και το πρόβλημα με το οποίο θα ασχοληθούμε. Το σκέλος αυτό ολοκληρώνεται με τη δομή της εργασίας μας. Στη συνέχεια έχουμε το θεωρητικό κομμάτι του διαχωρισμού μουσικού σήματος. Στο κεφάλαιο θα ασχοληθούμε με θέματα όπως το πολύ ενδιαφέρον cocktail party problem και το independent component analysis. Στο τέλος θα αναφερθούμε σε πιο μεγάλο βάθος στο πρόβλημα του Music source separation. Στη συνέχεια θα περάσουμε στο 2ο κεφάλαιο μας στο οποίο θα ασχοληθούμε με τις διάφορες τεχνικές μάθησης οι οποίες μπορούν να μας βοηθήσουν στη λύση του προβλήματος μας. Στην αρχή θα αναφερθούμε σε διάφορες γενικές τεχνικές οι οποίες θα μπορούσαν να είναι πολύ χρήσιμες. Στην συνέχεια θα παρουσιάσουμε τέσσερεις υποενότητες όπου η κάθε μια θα αντιστοιχεί σε ένα από τα μοντέλα με τα οποία θα ασχοληθούμε στην εργασία μας. Θα αναλύσουμε την αρχιτεκτονική και τις τεχνικές του κάθε μοντέλου. Μετά θα περάσουμε στο 3ο και πειραματικό κεφάλαιο της εργασίας μας. Στο κεφάλαιο αυτό θα τρέξουμε τα μοντέλα μας προσθέτοντας τους διάφορες μετρικές και εφαρμόζοντας πάνω τους το dataset μας. Με αυτό τον τρόπο θα μπορέσουμε να συγκρίνουμε και να αξιολογήσουμε τα μοντέλα και τις τεχνικές μηχανικής μάθησης που εφαρμόζουν. Μετά από το βήμα αυτό θα είμαστε σε θέση να παρουσιάσουμε τα αποτελέσματα και να εκφράσουμε τα συμπεράσματα

μας. Τα συμπεράσματα μας θα βασίζονται στο ποιο μοντέλο έχει την καλύτερη απόδοση και μπορεί να επιλύσει το πρόβλημα που είναι αντικείμενο της εργασίας μας. Η απόδοση θα βασίζεται στις μετρικές που θα προσθέσουμε εμείς στα μοντέλα αυτά. Αφού εκφράσουμε τα συμπεράσματα μας θα περάσουμε στον επίλογο της εργασίας. Τέλος θα έχουμε την βιβλιογραφία από όπου θα έχουμε αντλήσει όλες τις πληροφορίες για τα μοντέλα και το πρόβλημα μας. Αφού έχουμε ολοκληρώσει όλα αυτά τα βήματα θα φτάσουμε και στο τέλος της διπλωματικής μας εργασίας.

2. Διαχωρισμός Μουσικού σήματος

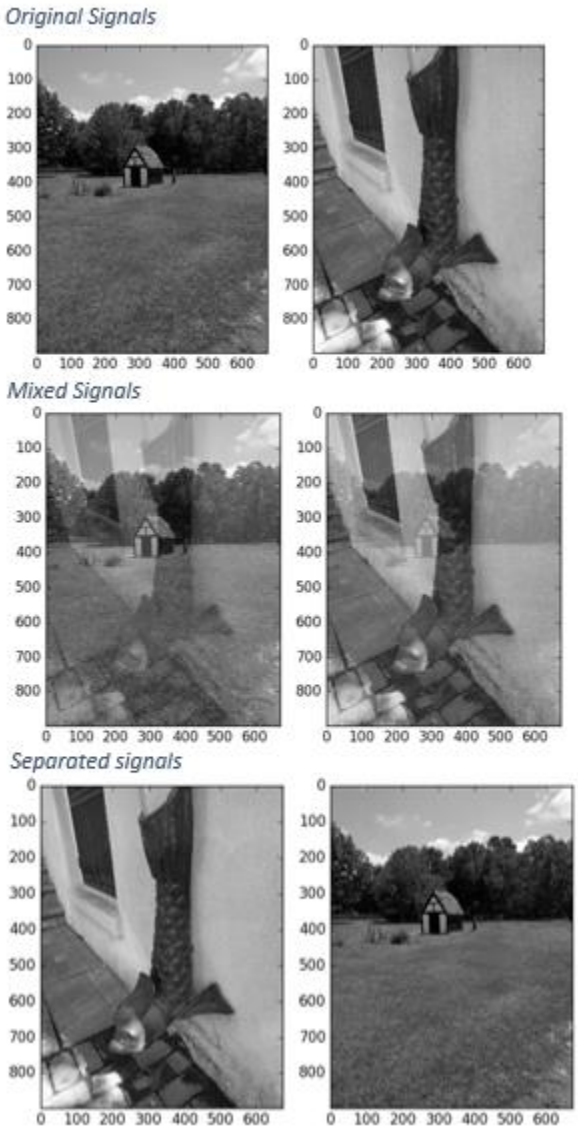
Στο κεφάλαιο αυτό θα αναφερθούμε στο θεωρητικό κομμάτι της διπλωματικής μας εργασίας. Πιο συγκεκριμένα, θα μιλήσουμε αρχικά για το γενικότερο πρόβλημα του τυφλού διαχωρισμού σήματος. Θα αναφερθούμε στο τι είναι το πρόβλημα αυτό, πως μπορούμε να το λύσουμε και που μπορούμε να το συναντήσουμε. Μία εφαρμογή του τυφλού διαχωρισμού σήματος είναι και το cocktail party problem το οποίο θα δούμε αναλυτικότερα. Στη συνέχεια θα μιλήσουμε για το Independent component analysis το οποίο πρόκειται για μια μέθοδο κλειδί στην εκτέλεση του τυφλού διαχωρισμού σήματος. Τέλος, θα περάσουμε στο πιο εξειδικευμένο θέμα μας το διαχωρισμό μουσικού σήματος που είναι και το κύριο θέμα της εργασίας μας.

2.1 Blind signal separation problem / source separation – cocktail party problem

Ο τυφλός διαχωρισμός σήματος είναι ο διαχωρισμός ενός συνόλου σημάτων πηγής από ένα σύνολο μικτών σημάτων. Ο διαχωρισμός αυτός γίνεται με ελάχιστη ή και καθόλου βοήθεια πληροφοριών σχετικών με τα σήματα πηγής ή τη διαδικασία ανάμειξης. Στόχος είναι η ανάκτηση των αρχικών συστατικών σημάτων από ένα σήμα μείγματος. Πρόκειται για ένα πρόβλημα που βρίσκει έδαφος σε ένα μεγάλο εύρος εφαρμογών, όπως η επεξεργασία σήματος και εικόνας, ιατρικές εφαρμογές, προβλήματα μηχανικής και στη μουσική. Το πρόβλημα μας είναι σε γενικές γραμμές εξαιρετικά υποκαθορισμένο αλλά μπορούν να προκύψουν πολλές χρήσιμες λύσεις. Τα περισσότερα βιβλία επικεντρώνονται στον διαχωρισμό χρονικών σημάτων όπως ο ήχος. Όμως σήμερα ο τυφλός διαχωρισμός σήματος συναντάται σε πολυδιάστατα δεδομένα όπως εικόνες όπου ενδέχεται να μην υπάρχει καμία χρονική διάσταση. Τώρα ας μιλήσουμε για τη λύση του προβλήματος μας. Έχουν προταθεί πολλές προσεγγίσεις για τη λύση αλλά η ανάπτυξη αυτών βρίσκεται ακόμα σε εξέλιξη. Μερικές από τις πιο επιτυχημένες προσεγγίσεις είναι οι principal components analysis και independent component analysis για τις οποίες θα μιλήσουμε αναλυτικότερα στην επόμενη υποενότητα. Άλλη μία υποσχόμενη προσέγγιση έχει γίνει από το πεδίο της computational auditory scene analysis η οποία επιχειρεί να επιτύχει τον διαχωρισμό της ακουστικής πηγής χρησιμοποιώντας την ανθρώπινη ακοή. Τέλος, το πρόβλημα αυτό μπορεί και πρέπει να το λύσει ο ανθρώπινος εγκέφαλος σε πραγματικό χρόνο. Στην ανθρώπινη αντίληψη η ικανότητα αυτή ονομάζεται auditory scene analysis ή cocktail party effect.

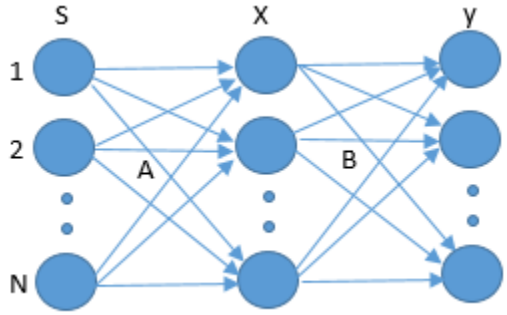
Ο τυφλός διαχωρισμός σήματος όπως αναφέραμε και παραπάνω συναντάται σε διάφορους τομείς και εφαρμογές. Στην παρακάτω εικόνα έχουμε την βασική έννοια του τυφλός διαχωρισμός σήματος στην επεξεργασία εικόνας. Εμφανίζονται τα

μεμονωμένα σήματα πηγής καθώς και τα μεικτά σήματα. Ο τυφλός διαχωρισμός σήματος χρησιμοποιείται για το διαχωρισμό των μεικτών σημάτων μόνο με γνώση των μικτών σημάτων χωρίς να ξέρει τίποτα για το αρχικό σήμα ή τον τρόπο με τον οποίο αναμείχθηκαν. Τα διαχωρισμένα σήματα είναι μόνο προσεγγίσεις των σημάτων προέλευσης.



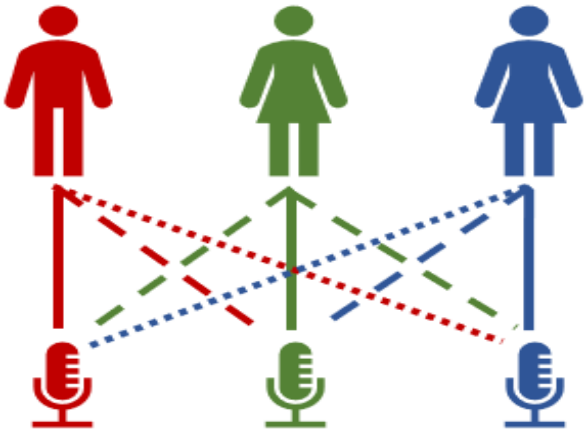
Επίσης, ο τυφλός διαχωρισμός σήματος συναντάται σε ιατρικές εφαρμογές. Πιο συγκεκριμένα, στο διαχωρισμό σημάτων που προέρχονται από καταγραφή της εγκεφαλικής δραστηριότητας. Επιπρόσθετα, συναντάται σε προβλήματα μηχανικής όπως στη καταγραφή σημάτων που προέρχονται από φθορές μηχανικών στοιχείων αλλά και τυφλός διαχωρισμός των σημάτων αυτών ώστε να εντοπιστεί το σήμα που σχετίζεται με τη φθορά και η πηγή της φθοράς που προκαλεί τη δημιουργία αυτού του σήματος. Τέλος έχουμε και τη μουσική όπου σε επόμενη υποενότητα θα μιλήσουμε αναλυτικά για το διαχωρισμό των μουσικών σημάτων.

Σε αυτό το σημείο θα μιλήσουμε για την μαθηματική προσέγγιση του τυφλού διαχωρισμού σήματος. Ξεκινώντας έχουμε ένα σύνολο μεμονωμένων σημάτων πηγής $s(t) = (s_1(t),\dots,s_n(t))^T$. Θα χρησιμοποιήσουμε έναν πίνακα $A= [a_{ij}] \in R^{m \cdot n}$ για να αναμείξουμε το προηγούμενο σύνολο και να δημιουργήσουμε ένα καινούργιο σύνολο μεικτών σημάτων $x(t) = (x_1(t),\dots,x_m(t))^T$. Συνήθως τα n και m είναι ίσα μεταξύ τους αλλά υπάρχουν και περιπτώσεις που αυτό δεν συμβαίνει. Αν έχουμε $m > n$ τότε το σύστημα των εξισώσεων είναι υπερπροσδιορισμένο και έτσι μπορεί να μην αναμειχθεί χρησιμοποιώντας μια συμβατική γραμμική μέθοδο. Αν έχουμε $n > m$ το σύστημα μας είναι υποπροσδιορισμένο και πρέπει να χρησιμοποιηθεί μη γραμμική μέθοδος για την ανάκτηση των μη αναμειγμένων σημάτων. Οπότε ο γενικός τύπος ο οποίος μας προκύπτει είναι $x(t) = A \cdot s(t)$. Ο τυφλός διαχωρισμός σήματος διαχωρίζει το σύνολο των μεικτών σημάτων x(t) μέσω ενός μη αναμειγμένου πίνακα $B= [B_{ij}] \in R^{n \cdot m}$. Αυτό το κάνει ώστε να ανακτήσει μία προσέγγιση των αρχικών σημάτων. Οπότε μας προκύπτουν οι εξής δύο εξισώσεις $y(t) = (y_1(t),\dots,y_n(t))^T$ και $y(t) = B \cdot x(t)$. Στην παρακάτω εικόνα παρουσιάζεται μέσω ενός flowchart το πώς προκύπτουν οι μαθηματικές σχέσεις τις οποίες βρήκαμε.



Στο τέλος αυτής της υποενότητας του 1ου κεφαλαίου της διπλωματικής μας εργασίας θα ασχοληθούμε με το φαινόμενο που ονομάζεται cocktail party effect. Το cocktail party effect είναι το φαινόμενο της ικανότητας του εγκεφάλου να εστιάσει την ακουστική του προσοχή σε ένα συγκεκριμένο ερέθισμα. Όπως για παράδειγμα ένα άτομο μπορεί να επικεντρωθεί σε μία συνομιλία σε ένα θορυβώδες δωμάτιο όπως είναι ένα cocktail party που αναφέρεται και η ονομασία του φαινομένου. Ένας ακροατής μπορεί να διαχωρίσει διαφορετικά ερεθίσματα σε διαφορετικά ρεύματα και στη συνέχεια να αποφασίσει ποιο από αυτά τα ρεύματα είναι πιο συναφή για αυτόν. Με αυτόν τον τρόπο έχει προταθεί ότι η αισθητήρια μνήμη κάποιου υποσυνείδητα αναλύει όλα τα ερεθίσματα και προσδιορίζει διακριτά κομμάτια πληροφοριών ταξινομώντας τα με βάση την εμβέλεια. Αυτό το αποτέλεσμα είναι που επιτρέπει στους περισσότερους ανθρώπους να δίνουν την προσοχή τους σε μια μόνο φωνή και να αγνοήσουν κατά κάποιο τρόπο τους άλλους. Αυτό το φαινόμενο

περιγράφεται συχνά με τον όρο επιλεκτικής ακοής ή επιλεκτικής προσοχής. Επίσης η κατάσταση αυτή μπορεί να περιγράψει άλλο ένα φαινόμενο κατά το οποίο κάποιος μπορεί να εντοπίσει αμέσως μια σημαντική λέξη για αυτόν όπως το όνομα του ανάμεσα σε ένα ευρύ φάσμα ακουστικών στοιχείων. Όταν λοιπόν κάποιο άτομο έχει αδυναμία διαχωρισμού των ερεθισμάτων με αυτόν τον τρόπο τότε αναφερόμαστε στο cocktail party problem ή αλλιώς cocktail party deafness.



Το cocktail party problem καθορίστηκε πρώτη φορά από τον Κόλιν Τσέρι το 1953. Το πρόβλημα ξεκίνησε από τους ελεγκτές εναέριας κυκλοφορίας. Πιο συγκεκριμένα, οι ελεγκτές λάμβαναν μηνύματα από πιλότους με μεγάφωνα από ένα μόνο ηχείο κάτι το οποίο έκανε το έργο του ελεγκτή πολύ δύσκολο. Ο Κόλιν Τσέρι ασχολήθηκε με το πρόβλημα αυτό κάνοντας διάφορα πειράματα. Ένα από τα συμπεράσματα που έβγαλε ήταν ότι η ικανότητα διαχωρισμού ήχων επηρεάζεται από μεταβλητές όπως το είδος του ηχείου, την κατεύθυνση του ήχου και το ρυθμό ομιλίας.

Στη συνέχεια διάφοροι ερευνητές εξακολούθησαν να ασχολούνται με την επιλεκτική προσοχή. Κάποιοι από τους ερευνητές αυτούς είναι ο Donald Broadbent, η Anne Treisman, η Diana Deutsch, ο Daniel Kahneman και πολλοί άλλοι. Γενικά αυτό που γνωρίζουμε είναι ότι η επιλεκτική προσοχή εμφανίζεται σε όλες τις ηλικίες ακόμα και στη βρεφική ηλικία όταν ένα μωρό ακούει τη φωνή των γονιών του. Επίσης η Anne Treisman δήλωσε ότι οι άνθρωποι είναι μόνιμα έτοιμοι να εντοπίσουν προσωπικές λέξεις και λέξεις ταμπού. Μία προσωπική λέξη είναι το όνομα μας και μια λέξη ταμπού μπορεί να είναι μια λέξη με σεξουαλικό περιεχόμενο. Βέβαια όσο μεγαλώνουμε η επιλεκτική προσοχή αρχίζει να αμφιταλαντεύεται καθώς φθείρεται η γνωστική ικανότητα, η μνήμη και η οπτική μας αντίληψη.

2.2 Independent Component Analysis

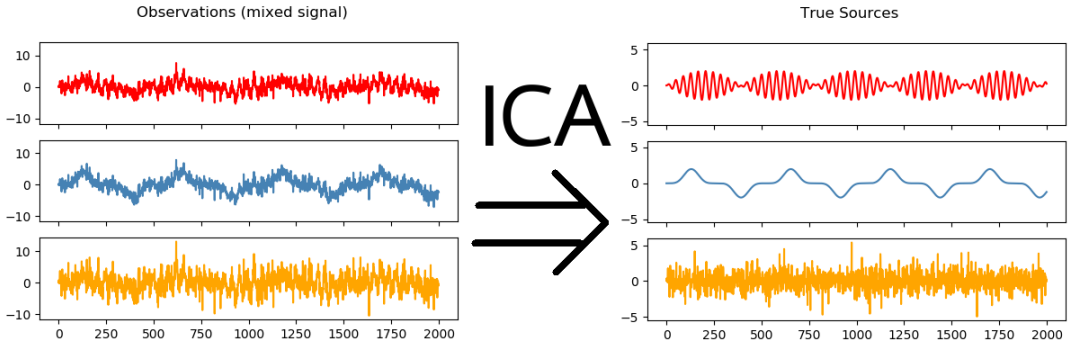
Το Independent component analysis (ICA) είναι μία υπολογιστική μέθοδος για το διαχωρισμό ενός σήματος πολλαπλών μεταβλητών σε πρόσθετα υποσυστατικά. Αυτό για να επιτευχθεί υποθέτουμε ότι τα υποσυστατικά είναι ενδεχομένως μη Gauss σήματα και ανεξάρτητα το ένα από το άλλο. Το ICA είναι όπως έχουμε αναφέρει και προηγουμένως στην εργασία μας μία ειδική περίπτωση τυφλού διαχωρισμού σήματος με χαρακτηριστικότερο παράδειγμα το cocktail party problem. Επίσης άλλο ένα παράδειγμα το οποίο είναι και το αντικείμενο της εργασίας μας είναι ο ήχος. Ο ήχος είναι συνήθως ένα σήμα που αποτελείται από την αριθμητική πρόσθεση σημάτων από διάφορες πηγές. Στόχος λοιπόν του τυφλού διαχωρισμού ICA είναι να διαχωρίσει αυτές τις πηγές από το συνολικό σήμα κάτι το οποίο έχει παρατηρηθεί ότι γίνεται με μεγάλη αποτελεσματικότητα.

Όσων αναφορά την κατασκευή των σημάτων M από τα N συστατικά μπορούμε να τοποθετήσουμε τα βάρη ανάμειξης τους σε έναν πίνακα M x N. Σημαντική προϋπόθεση είναι ότι εάν υπάρχουν N πηγές τότε απαιτούνται τουλάχιστον N παρατηρήσεις (π.χ. μικρόφωνα για τον ήχο) για την ανάκτηση των αρχικών σημάτων. Αν έχουμε ίσο αριθμό σημάτων και παρατηρήσεων τότε ο πίνακας ανάμιξης είναι τετράγωνος με M=N.

Τα πολύ καλά αποτελέσματα που δίνει ο διαχωρισμός ICA μικτών σημάτων βασίζεται σε δύο υποθέσεις και τρία αποτελέσματα ανάμιξης σημάτων πηγής. Οι δύο υποθέσεις είναι πρώτον ότι τα σήματα πηγής είναι ανεξάρτητα μεταξύ τους και δεύτερον ότι οι τιμές κάθε σήματος πηγής έχουν μη Gauss κατανομές. Τα τρία αποτελέσματα της ανάμιξης σημάτων πηγής αφορούν την ανεξαρτησία, την κανονικότητα και την πολυπλοκότητα. Για την ανεξαρτησία έχουμε ότι ενώ τα σήματα είναι ανεξάρτητα τα μείγματα σημάτων τους δεν είναι κάτι το οποίο συμβαίνει επειδή τα μείγματα σημάτων μοιράζονται τα ίδια σήματα πηγής. Για την κανονικότητα ξεκινάμε από το θεώρημα κεντρικών ορίων το οποίο μας λέει ότι η κατανομή ενός αθροίσματος ανεξάρτητων τυχαίων μεταβλητών με πεπερασμένη διακύμανση τείνει προς την κατανομή Gauss. Επίσης ένα άθροισμα δύο ανεξάρτητων τυχαίων μεταβλητών έχει συνήθως μια κατανομή που είναι πιο κοντά στο Gauss από οποιαδήποτε από τις δύο αρχικές μεταβλητές. Εδώ θεωρούμε την τιμή κάθε σήματος ως τυχαία μεταβλητή. Τέλος ως προς την πολυπλοκότητα έχουμε ότι οποιοδήποτε μείγμα σήματος έχει μεγαλύτερη χρονική πολυπλοκότητα από εκείνη του απλούστερου συστατικού σήματος πηγής. Κατά αυτόν τον τρόπο όλες αυτές οι αρχές συμβάλλουν στην βασική εγκατάσταση του ICA.

Ας αναφερθούμε τώρα στην ανεξαρτησία των συστατικών ή αλλιώς παραγόντων των σημάτων. Η ICA βρίσκει τα ανεξάρτητα συστατικά μεγιστοποιώντας τη στατιστική ανεξαρτησία των εκτιμώμενων συστατικών. Οι δύο ευρύτεροι ορισμοί της

ανεξαρτησίας για το ICA είναι η ελαχιστοποίηση τη αμοιβαίας ενημέρωσης και η μεγιστοποίηση χωρίς Gauss. Η οικογένεια των αλγορίθμων ICA της ελαχιστοποίησης των αμοιβαίων πληροφοριών χρησιμοποιεί μέτρα όπως η απόκλιση Kullback-Leibler και η μέγιστη εντροπία. Η οικογένεια των αλγορίθμων ICA της μεγιστοποίησης χωρίς Gauss με κίνητρο το θεώρημα του κεντρικού ορίου χρησιμοποιεί την κύρτωση και την αρνητική εντροπία. Το ICA είναι σημαντικό για τον διαχωρισμό τυφλού σήματος και έχει πολλές πρακτικές εφαρμογές όπως είναι η αναζήτηση ενός παραγοντικού κώδικα δεδομένων.



Στην παραπάνω εικόνα βλέπουμε στην αριστερή πλευρά τρία σήματα με μείξη και στη δεξιά πλευρά το σήμα που προκύπτει από την εφαρμογή του ICA.

Στη συνέχεια θα μιλήσουμε για τους μαθηματικούς ορισμούς του ICA. Η γραμμική ανεξάρτητη ανάλυση εξαρτημάτων μπορεί να χωριστεί σε αθόρυβη και θορυβώδεις περιπτώσεις όπου η αθόρυβη είναι μια ειδική περίπτωση της θορυβώδους ICA. Η μη γραμμική ICA είναι μια ξεχωριστή περίπτωση. Σύμφωνα με τον γενικό ορισμό τα δεδομένα αντιπροσωπεύονται από το παρατηρούμενο τυχαίο διάνυσμα $x = (x_1,\ldots,x_m)^T$ και τα κρυμμένα συστατικά ως το τυχαίο διάνυσμα $s = (s_1,\ldots,s_n)^T$. Ο στόχος είναι να μετατρέψουμε τα παρατηρούμενα δεδομένα x χρησιμοποιώντας έναν γραμμικό στατικό μετασχηματισμό W ως $s = W_x$ σε ένα διάνυσμα μεγίστων ανεξαρτήτως συνιστωσών s μετρούμενο με κάποια συνάρτηση ανεξαρτησίας $F(s_1,\ldots,s_n)$.

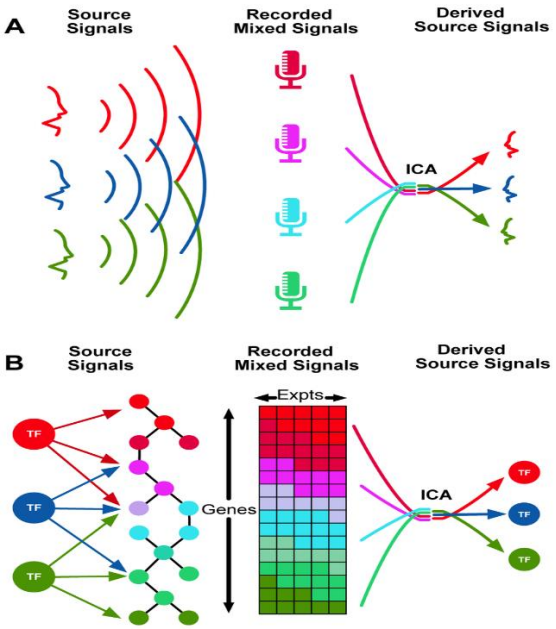
Μετά θα μιλήσουμε για κάθε περίπτωση ξεχωριστά ξεκινώντας από την γραμμική αθόρυβη ICA. Στην περίπτωση αυτή τα συστατικά x_i του παρατηρούμενου τυχαίου διανύσματος $x = (x_1,\ldots,x_m)^T$ δημιουργούνται ως άθροισμα των ανεξάρτητων συνιστωσών s_k, $k = 1,\ldots,n$ και έτσι έχουμε $x_i = a_{i,1}s_1 + \ldots + a_{i,k}s_k + \ldots + a_{i,n}s_n$ σταθμισμένο με τα βάρη ανάμιξης $a_{i,k}$. Μπορούμε να γράψουμε το ίδιο γενετικό μοντέλο σε μορφή διανύσματος ως $x = \sum_{k=1}^{n} s_k \; a_k$ όπου το παρατηρούμενο τυχαίο διάνυσμα x αντιπροσωπεύεται από τα βασικά διανύσματα $a_k = (a_{1,k},\ldots,a_{m,k})^T$. Τα βασικά διανύσματα a_k σχηματίζουν τις στήλες του πίνακα ανάμειξης $A = (a_1,\ldots,a_n)$ και ο γενικός τύπος μπορεί να γραφεί ως $x = A_s$ όπου $s = (s_1,\ldots,s_n)^T$. Βασικός στόχος είναι να εκτιμηθεί τόσο ο πίνακας ανάμειξης A όσο και οι πηγές s λαμβάνοντας υπόψη το

μοντέλο και τα δείγματα $x_1,\ldots,x_N$ του τυχαίου διανύσματος x. Ο στόχος αυτός επιτυγχάνεται με τον υπολογισμό των διανυσμάτων w και τη ρύθμιση μιας συνάρτησης κόστους που είτε μεγιστοποιεί τη non-gaussianity του υπολογιζόμενου $s_k = w^T x$ είτε ελαχιστοποιεί την αμοιβαία πληροφορία.

Για την γραμμική θορυβώδη ICA το μοντέλο ICA παίρνει τη μορφή $x = A_s + n$. Αυτό προκύπτει με την προστιθέμενη υπόθεση μηδενικού μέσου και ασυσχέτου Gaussian θορύβου $n \sim N(0, diag(\sum))$.

Τέλος έχουμε τη μη γραμμική ICA. Εδώ χρησιμοποιώντας μια μη γραμμική συνάρτηση ανάμιξης $f(\cdot|\theta)$ με παραμέτρους θ το μη γραμμικό μοντέλο ICA είναι $x = f(s|\theta) + n$.

Όσων αναφορά το ιστορικό κομμάτι και το πως ξεκίνησε η μέθοδος ICA το πρώιμο γενικό πλαίσιο εισήχθη από τους Jeanny Herault και Bernard Ans το 1984. Το 1985 το ICA αναπτύχθηκε περισσότερο από τον Christian Jutten και το 1994 ο Pierre Comon δημοσιοποίησε τη μέθοδο αυτή στην εργασία του. Το 1995 ο Tony Bell και ο Terry Sejnowski εισήγαγαν έναν γρήγορο και αποτελεσματικό αλγόριθμο ICA βασισμένο σε μία αρχή με το όνομα informax. Το 1999 οι Sepp Hochreiter και Jurgen Schmidhuber έδειξαν πως να αποκτήσουν μη γραμμικό ICA χωρίς να απαιτείται εκ των προτέρων γνώση σχετικά με τον αριθμό των ανεξάρτητων πηγών. Υπάρχουν πολλοί αλγόριθμοι που διατίθενται στη βιβλιογραφία και κάνουν ICA. Ένας ευρέως χρησιμοποιημένος είναι ο αλγόριθμος FastICA που αναπτύχθηκε από τους Hyvarinen και Oja. Άλλα παραδείγματα σχετίζονται με τον διαχωρισμό τυφλών πηγών όπου χρησιμοποιείται μια γενικότερη προσέγγιση.



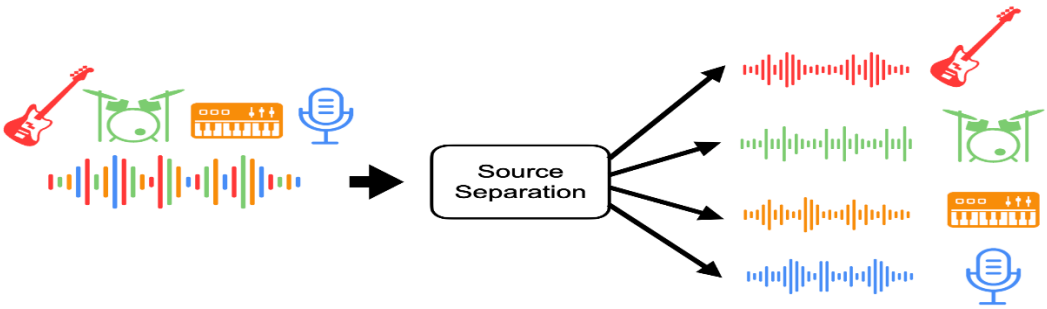
Καταλήγοντας υπάρχουν πολλές εφαρμογές του ICA ακόμα και στην καθημερινή μας ζωή. Ορισμένες εφαρμογές είναι: οπτική απεικόνιση νευρώνων, αναγνώριση

προσώπου, πρόβλεψη τιμών χρηματιστηρίου, επικοινωνίες κινητών τηλεφώνων, ανίχνευση της ωριμότητας μιας ντομάτας με βάση το χρώμα, πειράματα αλληλουχίας RNA με ένα κύτταρο, μελέτες του δικτύου ηρεμίας του εγκεφάλου, αστρονομία και κοσμολογία. Στην παραπάνω εικόνα βλέπουμε πως η εφαρμογή της μεθόδου ICA μας βοηθάει στο A κομμάτι να λύσουμε το cocktail party problem δημιουργώντας ξεχωριστά σήματα πηγής. Στο B κομμάτι το ICA εντοπίζει τα σήματα πηγής ρυθμιστών μεταγραφής από πολύπλοκες μετρήσεις γονιδιακής έκφρασης.

2.3 Music Source Separation

Στην τελευταία υποενότητα αυτού του κεφαλαίου θα μιλήσουμε για το music source separation που είναι και η βασικό αντικείμενο της εργασίας καθώς αυτό καλούμαστε να πραγματοποιήσουμε στο πειραματικό μέρος. Όπως αναφέραμε και προηγουμένως τα προβλήματα διαχωρισμού πηγής στην επεξεργασία ψηφιακού σήματος είναι εκείνα στα οποία έχουν αναμειχθεί πολλά σήματα σε ένα συνδυασμένο σήμα. Στόχος είναι να ανακτηθούν τα αρχικά συστατικά σήματα από το συνδυασμένο σήμα. Στην περίπτωση της μουσικής στην οποία συνδυάζονται πολλές μουσικές πηγές πρέπει να τη διαχωρίσουμε για την καλύτερη υποτίμηση της.

Υπάρχουν πολλές εφαρμογές που χρειάζονται το music source separation. Για παράδειγμα στη σκηνή ανάκτησης μουσικής πληροφοριών (MIR) εάν θέλουμε να εκτιμήσουμε τον ρυθμό της μουσικής πρέπει πρώτα να αποκτήσουμε τον κρουστό ήχο της ενώ εάν θέλουμε να εκτιμήσουμε τη χορδή πρέπει να τον αφαιρέσουμε. Πρέπει να πάρουμε τα φωνητικά για να εκτιμήσουμε τους στίχους και πρέπει να διαχωρίσουμε όλα τα όργανα για τη μουσική μεταγραφή. Κάτι παρόμοιο θα κάνουμε και εμείς στη συνέχεια ξεχωρίζοντας σε ένα ολόκληρο τραγούδι τα φωνητικά, τα ντραμς, το μπάσο και τα υπόλοιπα μουσικά όργανα.



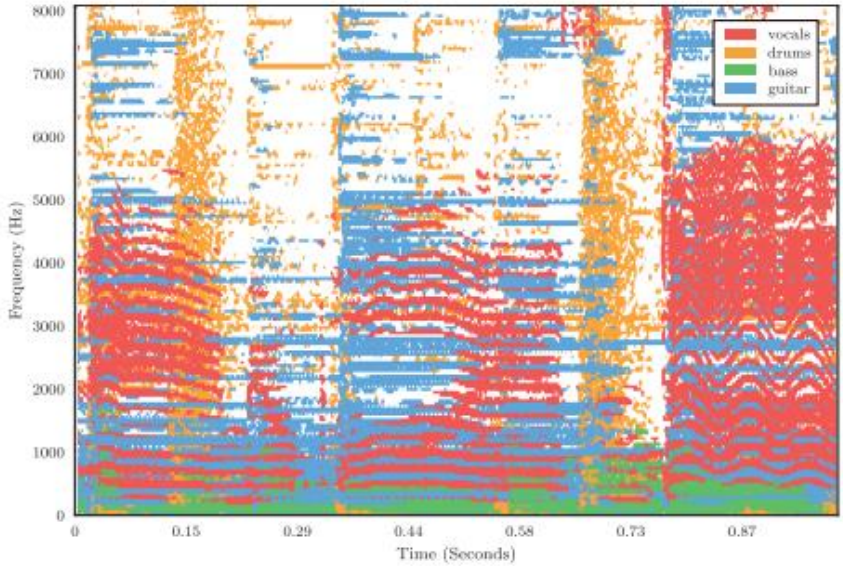
Το music source separation δεν είναι μόνο χρήσιμο για το MIR αλλά και από μόνο του. Ας πούμε ότι θέλουμε να εκπαιδεύσουμε ένα τραγούδι. Ο φωνητικός διαχωρισμός μπορεί να μας βοηθήσει παρέχοντας το φωνητικό σήμα που αφαιρείται από το αρχικό αρχείο μουσικής. Επίσης αν μπορούμε να ορίσουμε κάθε

διαχωρισμένη πηγή στη διαφορετική εικονική τοποθεσία είναι δυνατό να κάνουμε την αναμειγμένη έκδοση από μονοφωνικό σε στερεοφωνικό ήχο.

Όπως και σε πολλές άλλες εργασίες μια έρευνα για το διαχωρισμό της πηγής μουσικής ξεκινά από τον προσδιορισμό του προβλήματος. Πρέπει να προσδιορίσουμε διάφορα ερωτήματα όπως τι θέλουμε να χωρίσουμε, πόσα όργανα υπάρχουν, είναι μονοφωνικό ή στερεοφωνικό, έχουμε κάποια εκ των προτέρων εκπαιδευμένη βάση δεδομένων υπάρχει κάποια άλλη παράπλευρη πληροφορία. Ανάλογα με την εφαρμογή υπάρχουν πολλά προβλήματα που θέλουμε να λύσουμε.

Το music source separation έχει από καιρό γοητεύσει τους επιστήμονες. Οι μηχανικοί προσπάθησαν πρώτα να απομονώσουν τα φωνητικά ή τις κιθάρες ενός τραγουδιού προσαρμόζοντας το αριστερό και το δεξί κανάλι σε μια στερεοφωνική εγγραφή ή παίζοντας με τις ρυθμίσεις ισοσταθμιστή για να αυξήσουν ή να κόψουν ορισμένες συχνότητες. Άρχισαν να πειραματίζονται με την τεχνητή νοημοσύνη για να διαχωρίσουν τους ήχους συμπεριλαμβανομένων των μουσικών ηχογραφήσεων στις αρχές της δεκαετίας του 2000.

Σήμερα οι πιο συχνά χρησιμοποιημένες τεχνικές music source separation με τεχνητή νοημοσύνη λειτουργούν με την ανάλυση φασματογραμμάτων τα οποία είναι οπτικοποιήσεις που μοιάζουν με θερμικούς χάρτες των διαφορετικών συχνοτήτων ήχου ενός τραγουδιού. Τα φασματογράμματα μπορεί να είναι ωραία αλλά τα μοντέλα τεχνητής νοημοσύνης που τα χρησιμοποιούν έχουν αρκετούς σημαντικούς περιορισμούς. Προσπαθούν ιδιαίτερα για να διαχωρίσουν κομμάτια από drums και μπάσο τείνοντας να παραλείπουν σημαντικές πληροφορίες σχετικά με την αρχική ηχογράφηση πολλαπλών κομματιών. Αυτό συμβαίνει κυρίως επειδή προσπαθούν να συσπειρώσουν τους ήχους σε ένα προκαθορισμένο πίνακα συχνότητας και χρόνου αντί να τους αντιμετωπίσουν όπως είναι στην πραγματικότητα.



Στην παραπάνω εικόνα βλέπουμε την αναπαράσταση ενός μουσικού μίγματος με άξονες τη συχνότητα και το χρόνο. Η κυρίαρχη μουσική πηγή σε κάθε περιοχή συχνότητας εμφανίζεται με το χαρακτηριστικό χρώμα του μουσικού οργάνου.

Τα συστήματα τεχνητής νοημοσύνης που βασίζονται σε φασματογράφημα είναι σχετικά αποτελεσματικά στο διαχωρισμό των νοτών οργάνων που ηχούν σε μία μόνο συχνότητα σε οποιοδήποτε δεδομένο χρονικό σημείο. Αυτές εμφανίζονται σε ένα φασματογράφημα ως διακριτές αδιάσπαστες οριζόντιες γραμμές που τρέχουν από δεξιά προς τα αριστερά. Αλλά η απομόνωση κρουστών ήχων που παράγουν υπολειπόμενο θόρυβο όπως ντραμς, μπάσο και πιάνο είναι πολύ πιο σκληρή δουλειά. Ένα τύμπανο ακούγεται σαν ένα μοναδικό ολόκληρο γεγονός σε πραγματικό χρόνο αλλά στην πραγματικότητα περιέχει διάφορα μέρη. Για ένα τύμπανο αυτό περιλαμβάνει ένα αρχικό ήχο που καλύπτει ένα ευρύ φάσμα υψηλότερων συχνοτήτων ακολουθούμενη από έναν άλλο ήχο σε μικρότερο εύρος χαμηλών συχνοτήτων.

Τα φασματογράμματα τα οποία μπορούν να αναπαραστήσουν μόνο τα ηχητικά κύματα ως χρόνο και συχνότητα δεν μπορούν να αποτυπώσουν τέτοιες αποχρώσεις. Κατά συνέπεια επεξεργάζονται ένα drumbeat ως πολλές μη συνεχόμενες κάθετες γραμμές και όχι ως ένας καθαρός ήχος. Αυτός είναι και ο λόγος για τον οποίο τα κομμάτια των ντραμς και των μπάσων που έχουν διαχωριστεί μέσω φασματογράμματος δεν ακούγονται καθαρά.

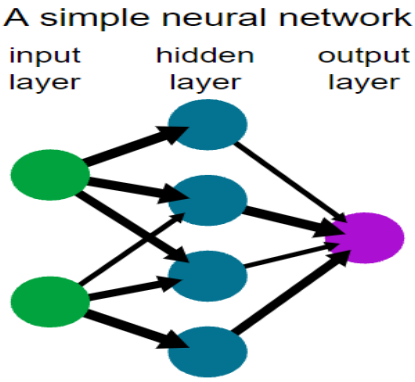
3. Τεχνικές Μηχανικής Μάθησης για διαχωρισμό μουσικού σήματος

Στο κεφάλαιο αυτό θα αναφερθούμε σε τεχνικές μηχανικής μάθησης που έχουν εφαρμοστεί για το διαχωρισμό ενός μουσικού σήματος. Αρχικά θα μιλήσουμε για γενικές τεχνικές που διάφοροι ερευνητές έχουν προτείνει μέσα από εργασίες και papers. Στην συνέχεια θα αναφερθούμε στα τέσσερα μοντέλα τα οποία θα αξιολογήσουμε και εμείς στην δική μας εργασία. Να υπενθυμίσουμε ότι τα μοντέλα που θα χρησιμοποιήσουμε είναι τα *demucs*, *OpenUnmix*, *D3Net* και *Spleeter.* Θα δημιουργήσουμε μία υποενότητα για κάθε μοντέλο αντίστοιχα και θα αναφερθούμε στην αρχιτεκτονική και τις τεχνικές που χρησιμοποιούν για το διαχωρισμό των μουσικών σημάτων.

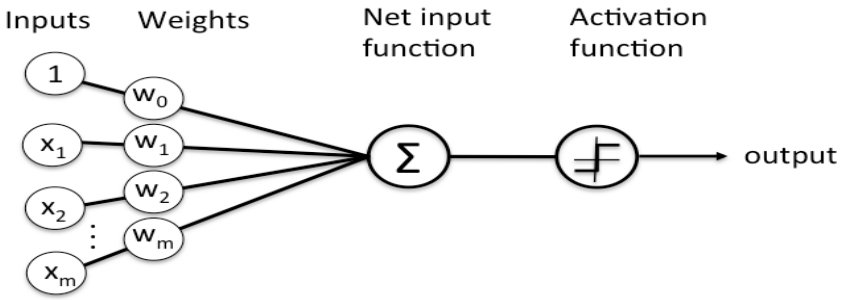
3.1 Γενικές Τεχνικές

Διάφοροι αλγόριθμοι έχουν εφαρμοστεί για την επίτευξη του source separation και ειδικότερα του music source separation όπως το Independent component analysis που αναλύσαμε στο προηγούμενο κεφάλαιο και το principal component analysis. Αυτοί οι αλγόριθμοι επισκιάστηκαν ελαφρά από την επιτυχία της τεχνικής της μηχανικής μάθησης (machine learning) και της βαθιάς μάθησης (deep learning). Οι αλγόριθμοι βαθιάς μάθησης είναι σε θέση να μοντελοποιήσουν μη γραμμικότητες και να παρέχουν γρηγορότερες εφαρμογές από τους προηγούμενους αλγόριθμους. Αυτά συνήθως διατυπώνονται ως εποπτευόμενο μαθησιακό πρόβλημα χρησιμοποιώντας συγκεκριμένους στόχους και διαφορετικές λειτουργίες κόστους. Οι περισσότεροι αλγόριθμοι χρησιμοποιούν επαναλαμβανόμενες δομές νευρωνικού δικτύου όπως μονάδες LSTM για να διερευνήσουν τις εξαρτήσεις χρόνου. Οπότε σε αυτή την υποενότητα θα εξηγήσουμε τις βασικές έννοιες για το πως λειτουργούν οι αλγόριθμοι βαθιάς μάθησης του music source separation. Θα ξεκινήσουμε με τον τρόπο λειτουργίας ενός νευρωνικού δικτύου και μετά θα εξηγήσουμε τρεις από τις πιο συνηθισμένες τεχνικές που χρησιμοποιούνται στο music source separation.

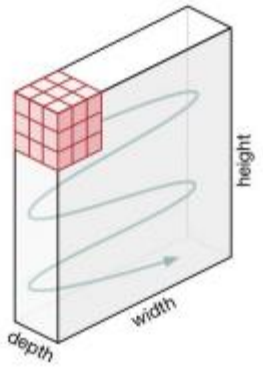
Η βαθιά μάθηση είναι ένα υποπεδίο της μηχανικής μάθησης που αφορά αλγόριθμους εμπνευσμένους από τη δομή και τη λειτουργία του εγκεφάλου που ονομάζονται τεχνητά νευρωνικά δίκτυα. Ένα νευρωνικό δίκτυο είναι ένας συνδυασμός πολλών στρωμάτων νευρώνων. Τα νερωνικά δίκτυα αποτελούνται από στρώματα εισόδου και εξόδου και στις περισσότερες περιπτώσεις έχουν και ένα κρυφό στρώμα αποτελούμενο από μονάδες που μετατρέπουν την είσοδο σε κάτι που μπορεί να χρησιμοποιήσει το επίπεδο εξόδου. Σε μία μονάδα νευρώνα τα δεδομένα εισόδου μετασχηματίζονται με γραμμικό ή μη γραμμικό τρόπο για να παράγουν μία μόνο έξοδο.



Τώρα ας μιλήσουμε για τα βάρη στα νευρωνικά δίκτυα. Τα βάρη μετατρέπονται σε κάθε επανάληψη έως ότου οι έξοδοι δικτύου πλησιάσουν τις επιθυμητές εξόδους. Αυτή η πραγμάτωση βάρους ονομάζεται αναπαραγωγή και χρησιμοποιεί τη διαφορά μεταξύ της πραγματικής εξόδου και της ιδανικής εξόδου κάθε στρώματος από το επίπεδο εξόδου στο επίπεδο εισόδου. Ορισμένες παράμετροι χρησιμοποιούνται για την επίτευξη καλής απόδοσης δικτύου όπως το ρυθμό εκμάθησης που ελέγχει πόσο γρήγορα γίνεται η προσαρμογή βαρών ή η κανονικοποίηση. Οι παράμετροι αυτοί ελέγχουν το πόσο εμείς θέλουμε να είναι ευέλικτο το μοντέλο μας.



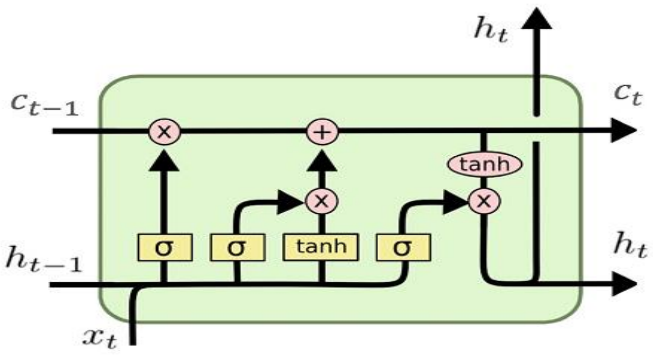
Στη συνέχεια θα μιλήσουμε για ένα από τα πιο γνωστά και εξαιρετικά χρήσιμα είδη νευρωνικών δικτύων που ονομάζεται Convolutional Neural Networks (CNN). Τα CNN σχεδιάστηκαν αρχικά για επεξεργασία εικόνας. Ένα συνελικτικό στρώμα σχεδιάστηκε για να μειώσει τις εικόνες σε μια μορφή που είναι ευκολότερη συην επεξεργασία χωρίς να χάνονται χαρακτηριστικά που είναι απαραίτητα για μια καλή πρόβλεψη. Το CNN είναι σε θέση να καταγράψει με επιτυχία τις χωρικές και χρονικές εξαρτήσεις μέσω της εφαρμογής σχετικών φίλτρων.



Στην παραπάνω εικόνα έχουμε ένα παράδειγμα φίλτρου πυρήνα. Μπορούμε να παρατηρήσουμε ότι ο πυρήνας δηλαδή το κόκκινο κουτάκι κινείται σε ολόκληρη την εικόνα με μια συγκεκριμένη τιμή. Η έξοδος σε κάθε θέση είναι το άθροισμα των στοιχειωδών προϊόντων μεταξύ του φίλτρου και της πολυδιάστατης εισόδου. Κατά τη διάρκεια της εκπαίδευσης στο δίκτυο οι συντελεστές φίλτρου αλλάζουν για τη βελτιστοποίηση των αποτελεσμάτων.

Οπότε μπορούμε να πούμε ότι η ικανότητα μείωσης των διαστάσεων και η καταγραφή χωρικής εξάρτησης καθιστά το CNN μια καλή επιλογή στην επεξεργασία ήχου και στη δική μας περίπτωση στο music source separation. Ανάλογα με τη διάσταση της εισόδου μπορούμε να χρησιμοποιήσουμε μονοδιάστατα ή δυσδιάστατα στρώματα.

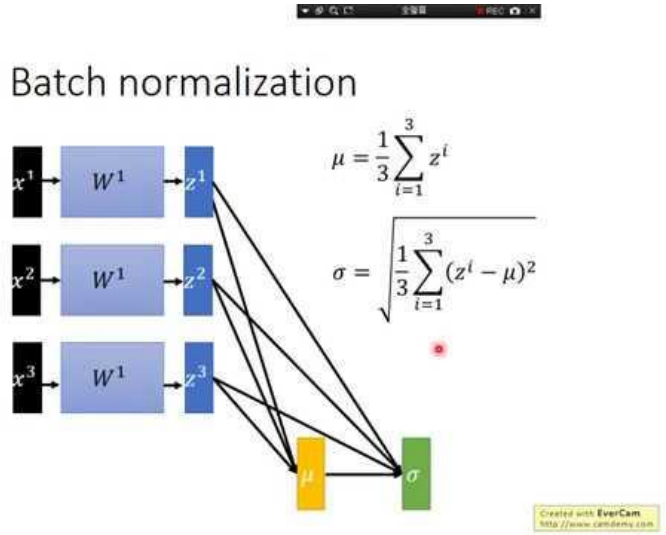
Ένα ακόμα είδος νευρωνικών δικτύων που χρησιμοποιείτε εξίσου αρκετά στην επιστημονική κοινότητα όσο και τα CNN είναι τα επαναλαμβανόμενα νευρωνικά δίκτυα και συγκεκριμένα τα Long-Short Time Memory (LSTM). Πρόκειται για δίκτυα με βρόχους μέσα τους επιτρέποντας τη διατήρηση των πληροφοριών. Τα LSTM είναι ικανά να μάθουν μακροπρόθεσμες εξαρτήσεις μεταξύ των δειγμάτων εισόδου. Επίσης θυμούνται πολύ καλά πληροφορίες για μεγάλα χρονικά διαστήματα.



LSTM
(Long−Short Term Memory)

Τα LSTM έχουν δομή σαν αλυσίδα με επαναλαμβανόμενη μονάδα. Για παράδειγμα σε μια εφαρμογή με LSTM κάθε μονάδα αποτελείται από ένα κελί, μια πύλη εισόδου, μια πύλη εξόδου και μια forget πύλη. Το κελί θυμάται τιμές σε αυθαίρετα χρονικά διαστήματα και οι τρείς πύλες ρυθμίζουν τη ροή πληροφοριών μέσα και έξω από το κελί.

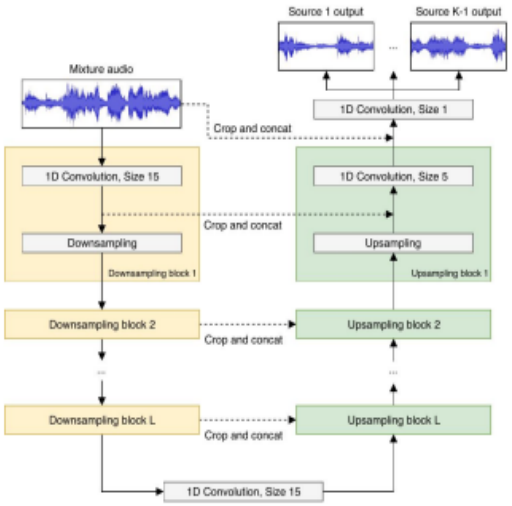
Ακόμα μια τεχνική την οποία χρησιμοποιούμε συνέχεια στην μηχανική μάθηση είναι το batch normalization το οποίο χρησιμοποιείται για την αύξηση της σταθερότητας ενός νευρωνικού δικτύου. Η διαδικασία συνιστάται στην ομαλοποίηση της εξόδου ενός προηγούμενου στρώματος ενεργοποίησης αφαιρώντας το μέσο όρο παρτίδας και διαιρώντας με την τυπική απόκλιση της παρτίδας. Με τον τρόπο αυτό βεβαιώνεται ότι δεν υπάρχει καμία ενεργοποίηση που είναι πολύ υψηλή ή πολύ χαμηλή και προσαρμόζει την παραμετροποίηση ενός μοντέλου ώστε η απώλεια επιφάνειας να είναι πιο ομαλή.



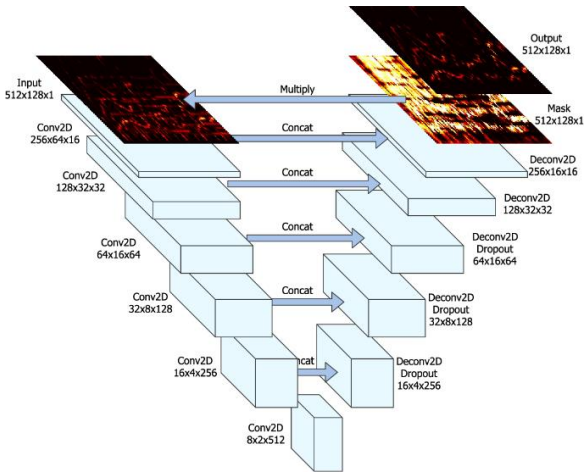
Στο τέλος αυτής της υποενότητας θα αναφερθούμε στους αλγόριθμους του music source separation που μπορεί να έχει στην αρχιτεκτονική του ένα μοντέλο διαχωρισμού μουσικής. Οι αλγόριθμοι χωρίζονται σε δύο οικογένειες. Ανάλογα με την είσοδο μπορούμε να διακρίνουμε δύο τύπους αλγορίθμων. Έχουμε τους αλγόριθμους κυματομορφής (waveform) και αλγόριθμους προ-επεξεργασμένων κυματομορφών (pre-processed waveform).

Οι αλγόριθμοι κυματομορφής έχουν σχεδιαστεί ως συστήματα από άκρο σε άκρο. Η εργασία σε αυτόν τον τομέα επιτρέπει τη μοντελοποίηση πληροφοριών φάσης αποφεύγοντας σταθερούς φασματικούς μετασχηματισμούς και υπολογισμούς χαμηλής καθυστέρησης. Οι περισσότεροι από αυτούς τους αλγορίθμους χρησιμοποιούν πολλαπλά μονοδιάστατα στρώματα συνέλιξης και δομή κωδικοποίησης-αποκωδικοποίησης χρησιμοποιώντας μείωση και αύξηση δειγματοληψίας. Αυτό το είδος δομής ονομάζεται δομή U-Net. Εισήχθη στη βιοϊατρική απεικόνιση για να βελτιώσει την ακρίβεια και τον εντοπισμό

μικροσκοπικών εικόνων. Οι πιο διάσημοι αλγόριθμοι είναι οι Demucs, Wave-U-Net, Conv-Tasnet. Τον Demucs όπως έχουμε αναφέρει θα τον χρησιμοποιήσουμε και εμείς στην συνέχεια της εργασίας μας. Τον Wave-U-Net τον βλέπουμε στην παρακάτω εικόνα. Παρόλο που δεν είναι οι πιο εκτεταμένοι αλγόριθμοι τα τελευταία αποτελέσματα δείχνουν τις δυνατότητες των συστημάτων από άκρο σε άκρο στο music source separation και είναι από τα μοντέλα που χρησιμοποιούνται περισσότερο για το πρόβλημα αυτό.



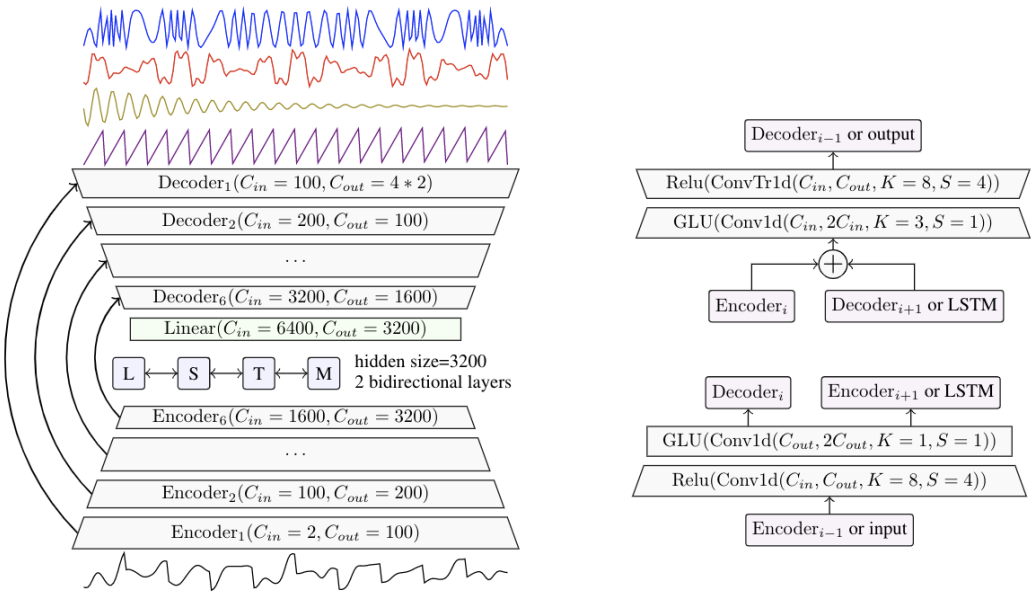
Οι αλγόριθμοι προ-επεξεργασμένων κυματομορφών λειτουργούν με τα φασματογράμματα που δημιουργούνται από το Short-Time Fourier Transform (STFT). Παράγουν μια μάσκα στα φάσματα μεγέθους για κάθε πλαίσιο και κάθε πηγή. Ο ήχος εξόδου δημιουργείται εκτελώντας ένα αντίστροφο STFT στα καλυμμένα φασματογράμματα. Ωστόσο η έξοδος STFT εξαρτάται από πολλές παραμέτρους όπως το μέγεθος και η επικάλυψη των καρέ ήχου και μπορεί να επηρεάσει την ανάλυση χρόνου και συχνότητας. Όπως και στον τομέα κυματομορφής αυτοί οι αλγόριθμοι συνήθως διερευνούν δίκτυα δομής U-Net. Αλλά σε αυτή την περίπτωση χρησιμοποιώντας δυσδιάστατες συνελικτικές στρώσεις. Ένα παράδειγμα θα μπορούσε να είναι το Spleeter το οποίο θα χρησιμοποιήσουμε και εμείς.



Σε αυτό το σημείο ολοκληρώνουμε τις γενικές τεχνικές της μηχανικής μάθησης και θα περάσουμε στα μοντέλα τα οποία θα χρησιμοποιήσουμε, θα αξιολογήσουμε και θα βγάλουμε συμπεράσματα μέσα από αυτά.

3.2 Μοντέλο Demucs

Ένα από τα μοντέλα τα οποία θα αξιολογήσουμε στην συνέχεια της εργασίας μας είναι το Demucs. Στην υποενότητα αυτή θα αναλύσουμε την αρχιτεκτονική του μοντέλου αυτού οπότε θα καλύψουμε το θεωρητικό κομμάτι του μοντέλου. Ξεκινώντας η λειτουργεία του demucs είναι να παίρνει ένα στερεοφωνικό μείγμα ως είσοδο και να εξάγει μια στερεοφωνική εκτίμηση για κάθε πηγή. Το μοντέλο έχει μια αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή που αποτελείται από έναν συνελικτικό κωδικοποιητή, ένα αμφίδρομο LSTM και έναν συνελικτικό αποκωδικοποιητή με τον κωδικοποιητή και αποκωδικοποιητή να συνδέονται με παραλείψεις U-Net συνδέσεων. Παρόμοια με άλλες εργασίες στην εικόνα και τον ήχο δεν χρησιμοποιούμε κανονικοποίηση κατά παρτίδες καθώς κάποια πειράματα έδειξαν ότι ήταν επιζήμια για την απόδοση του μοντέλου. Στην παρακάτω εικόνα απεικονίζεται η αρχιτεκτονική του μοντέλου.

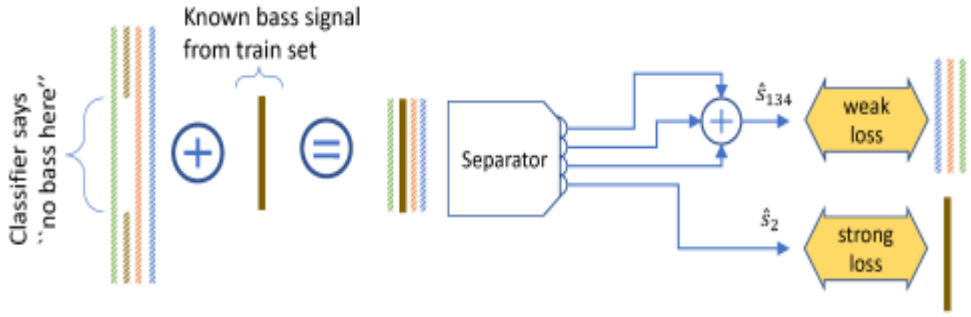


Το Demucs πρόκειται για ένα νέο μοντέλο κυματομορφής με αρχιτεκτονική πιο κοντά σε μοντέλα για παραγωγή ήχου με μεγαλύτερη χωρητικότητα στον κωδικοποιητή. Στα αποτελέσματα ξεπερνά τα μοντέλα του παρελθόντος ως προς την μετρική SDR αλλά έχει χαμηλότερη τιμή από το μοντέλο Conv-Tasnet που είναι παρόμοιο με αυτό. Έχει καλύτερη ποιότητα από το Conv-Tasnet αλλά περισσότερο contamination από άλλες πηγές κάτι το οποίο εξηγεί τη διαφορά στο SDR. Για μεγαλύτερα σύνολα

δεδομένων η διαφορά στις τιμές του SDR για Demucs και Conv-Tasnet μικραίνει άρα το μοντέλο αυτό μπορεί να χαρακτηριστεί ως πολύ υποσχόμενο.

Όπως είπαμε το demucs έχει κωδικοποιητή και αποκωδικοποιητή. Ας μιλήσουμε αρχικά για τον κωδικοποιητή. Ο κωδικοποιητής αποτελείται από 6 στοιβαγμένα συνελικτικά μπλοκ L αριθμημένα από 1 έως L. Κάθε μπλοκ αποτελείται από μία συνέλιξη με μέγεθος πυρήνα 8, βήμα 4, είσοδο C_i -1 κανάλια, κανάλια εξόδου C_i, ενεργοποίηση ReLU ακολουθούμενη από συνέλιξη με μέγεθος πυρήνα 1, κανάλια εξόδου $2C_i$ και κλειστές γραμμικές μονάδες ως λειτουργία ενεργοποίησης. Δεδομένου ότι οι μονάδες GLU μειώνουν κατά το ήμισυ τον αριθμό των καναλιών η τελική έξοδος του μπλοκ i έχει κανάλια εξόδου C_i. Η χρήση των ενεργοποιήσεων GLU αυξάνει σημαντικά την απόδοση. Ο αριθμός των καναλιών στο μείγμα εισόδου είναι C0 = C = 2 ενώ χρησιμοποιούμε C1: = 64 ως τον αριθμό των καναλιών εξόδου για το πρώτο μπλοκ κωδικοποιητή. Ο αριθμός των καναλιών στη συνέχεια διπλασιάζεται σε κάθε επόμενο μπλοκ δηλαδή $C_i = 2C_i – 1$ για i = 2…L οπότε ο τελικός αριθμός καναλιών είναι CL = 2048. Στη συνέχεια χρησιμοποιούμε ένα αμφίδρομο LSTM με 2 επίπεδα και κρυφό μέγεθος CL. Το LSTM εξάγει κανάλια 2CL ανά χρονική θέση. Χρησιμοποιούμε ένα γραμμικό επίπεδο για να μειώσουμε αυτόν τον αριθμό σε CL.

Για τον αποκωδικοποιητή μπορούμε να πούμε ότι είναι το αντίστροφο του κωδικοποιητή. Αποτελείται από μπλοκ L αριθμημένα με αντίστροφη σειρά από L έως 1. Τα μπλοκ ξεκινούν με μια συνέλιξη με βήμα 1 και πλάτος πυρήνα 3 για να παρέχουν στοιχεία σχετικά με βήματα χρόνου, κανάλια εισόδου-εξόδου C_i και ενεργοποίηση ReLU. Τέλος χρησιμοποιούμε μια συνέλιξη με πλάτος πυρήνα 8, βήμα 4 και εξόδους $C_i – 1$ και ενεργοποίηση ReLU. Οι πηγές S συντίθενται μόνο στο τελικό στρώμα μετά από όλα τα μπλοκ αποκωδικοποιητών. Το τελικό στρώμα είναι γραμμικό με κανάλια εξόδου $S \cdot C0$ ένα για κάθε πηγή χωρίς καμία πρόσθετη λειτουργία ενεργοποίησης. Κάθε ένα από αυτά τα κανάλια δημιουργεί απευθείας την αντίστοιχη κυματομορφή.



Στη συνέχεια θα μιλήσουμε για τη δομή του U-Net που χρησιμοποιούμε. Παρόμοια με το Wave-U-Net υπάρχουν συνδέσεις παράκαμψης μεταξύ των μπλοκ κωδικοποιητή και αποκωδικοποιητή με τον ίδιο δέκτη. Ενώ το κύριο κίνητρο προέρχεται από εμπειρικές επιδόσεις ένα πλεονέκτημα των συνδέσεων παράλειψης είναι να δοθεί άμεση πρόσβαση στο αρχικό σήμα και συγκεκριμένα επιτρέπει την

απευθείας μεταφορά της φάσης του σήματος εισόδου στην έξοδο. Σε αντίθεση με το Wave-u-Net χρησιμοποιούμε μετατοπισμένες περιστροφές αντί για γραμμική παρεμβολή ακολουθούμενη από συνέλιξη με βήμα 1. Για την ίδια αύξηση του δεκτικού πεδίου οι μετατροπές μετατόπισης απαιτούν 4 φορές λιγότερες λειτουργείες και μνήμη. Αυτό περιορίζει τον συνολικό αριθμό καναλιών που μπορούν να χρησιμοποιηθούν πριν εξαντληθεί η μνήμη τους.

Στο Demucs μοντέλο οι συνδέσεις παράκαμψης του U-Net και η πύλη που εκτελούνται από τις GLU υποδηλώνουν ότι αυτή η αρχιτεκτονική είναι αρκετά εκφραστική για να αντιπροσωπεύει μάσκες σε μία έμπειρη αναπαράσταση του σήματος εισόδου με παρόμοιο τρόπο με ένα γνωστό μοντέλο το Conv-Tasnet. Η προσέγγιση του Demucs είναι πιο εκφραστική από το Conv-Tasnet και τα κύρια πλεονεκτήματα της είναι οι αναπαραστάσεις πολλαπλής κλίμακας της εισόδου και οι μη γραμμικοί μετασχηματισμοί προς και από τον τομές της κυματομορφής.

Ένα από τα κύρια μειονεκτήματα του μοντέλου Demucs σε σύγκριση με άλλη αρχιτεκτονική είναι το μεγάλο μέγεθος του μοντέλου του το οποίο είναι περισσότερο από 1014MB έναντι άλλον όπως το Conv-Tasnet που είναι 42MB. Για να μειώσουμε το ζήτημα αυτό μπορούμε είτε να μειώσουμε το αρχικό αριθμό καναλιών που θα βελτιώσουν τόσο το μέγεθος του μοντέλου όσο και θα μειώσουν την υπολογιστική πολυπλοκότητα του μοντέλου είτε να χρησιμοποιήσουμε την τεχνική κβαντοποίησης. Στο σημείο αυτό έχουμε αναφέρει ότι χρειάζεται για το μοντέλο Demucs και μπορούμε να προχωρήσουμε στο επόμενο μοντέλο.

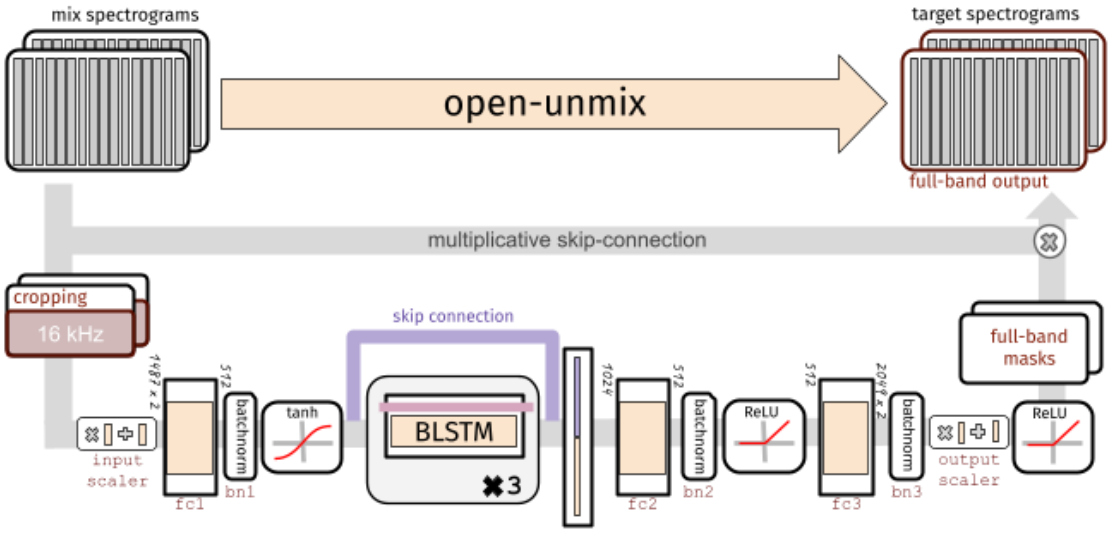
3.3 Μοντέλο Open-Unmix

Στην Τρίτη υποενότητα του κεφαλαίου αυτού θα ασχοληθούμε και θα γνωρίσουμε το δεύτερο μοντέλο της εργασίας μας που είναι το Open-Unmix. Το Open-Unmix αναπτύχθηκε από τους Fabian-Robert Stoter και Antoine Liutkus στο Inria Montpellier. Η έρευνα σχετικά με την αρχιτεκτονική του βαθύ νευρωνικού δικτύου καθώς και η διαδικασία κατάρτισης έγινε σε στενή συνεργασία με τους Stefan Uhlich και Yuki Mitsufuji από τη Sony Corporation. Το Open-Unmix είναι μέρος ενός οικοσυστήματος λογισμικού σύνολο δεδομένων και διαδικτυακών πόρων που ονομάζεται sigsep και βασίζεται στο αμφίδρομο μοντέλο LSTM.

Οι δημιουργοί του μοντέλου προτείνουν το Open-Unmix το οποίο εφαρμόζει τη μηχανική μάθηση για συγκεκριμένες εργασίες του διαχωρισμού μουσικής. Πιστεύουν πως η έλλειψη γνώσης στον τομές των μουσικών σημάτων οδηγεί συχνά σε κακή απόδοση όπου τα θέματα είναι δύσκολο να εντοπιστούν χρησιμοποιώντας αλγόριθμους που βασίζονται στη μάθηση. Για το λόγο αυτό σχεδιάσανε το Open-

Unmix το οποίο βασίζεται σε διαδικασίες που έχουν επαληθευτεί από την κοινότητα της μηχανικής μάθησης.

Οι σχεδιαστικές επιλογές που έγιναν για το Open-Unmix επιδίωξαν να επιτύχουν δύο κάπως αντιφατικούς στόχους. Ο πρώτος στόχος του είναι να έχει υπερσύγχρονες επιδόσεις και ο δεύτερος στόχος είναι να είναι εύκολα κατανοητός ώστε να μπορεί να χρησιμεύει ως βάση για την έρευνα που θα επιτρέπει βελτιωμένες επιδόσεις στο μέλλον. Στο παρελθόν πολλοί ερευνητές αντιμετώπισαν δυσκολίες πριν και μετά την επεξεργασία που θα μπορούσαν να αποφευχθούν με την ανταλλαγή γνώσεων τομέα. Για το λόγο αυτό στόχος των δημιουργών ήταν να σχεδιάσουν ένα σύστημα που θα επιτρέπει στους ερευνητές να επικεντρωθούν σε νέες αναπαραστάσεις και νέες αρχιτεκτονικές.



Το μοντέλο έχει αναπτυχθεί στη γλώσσα Python και συγκεκριμένα με τη βιβλιοθήκη PyTorch λόγω της ισορροπίας της μεταξύ απλότητας και αρθρωτότητας. Επίσης οι δημιουργοί έχουν ανεβάσει το μοντέλο στο διαδίκτυο και συνεχώς προσπαθούν να το ενημερώνουν έτσι ώστε να μπορούν όλο και περισσότεροι ερευνητές να ασχοληθούν. Για παράδειγμα σχεδιάζουν να κυκλοφορήσουν μια θύρα για το Tensorflow 2.0 μόλις κυκλοφορήσει το framework. Όμως δεν περιλαμβάνουν προ-εκπαιδευμένα μοντέλα έτσι ώστε οι ερευνητές να μπορούν να κάνουν συγκρίσεις.

Στη συνέχεια θα αναφερθούμε στα τρία χαρακτηριστικά που οι δημιουργοί φρόντισαν να βασίζεται το Open-Unmix ώστε όλο και περισσότεροι ερευνητές θα μπορούσαν να βασιστούν στο μοντέλο αυτό και να ασχοληθούν με το διαχωρισμό ήχου. Αρχικά φροντίσανε ώστε το Open-Unmix να είναι μια απλή επέκταση. Δηλαδή το μέρος του κώδικα πριν και μετά την επεξεργασία, η φόρτωση δεδομένων, η εκπαίδευση και τα μοντέλα είναι εύκολο να αντικατασταθούν. Συγκεκριμένα έγινε μια συγκεκριμένη προσπάθεια για να διευκολυνθεί η αντικατάσταση του μοντέλου. Το 2ο χαρακτηριστικό είναι ότι το Open-Unmix δεν είναι πακέτο. Το λογισμικό αποτελείται σε μεγάλο βαθμό από ανεξάρτητα και αυτόνομα μέρη διατηρώντας το

εύκολο στη χρήση και στην αλλαγή. Το τρίτο χαρακτηριστικό είναι το ότι η φιλοσοφία του μοιάζει πολύ με το χαρακτηριστικό παράδειγμα του MNIST. Λόγω του στόχου των δημιουργών να διευκολύνουν τους εμπειρογνώμονες μηχανικής μάθησης να δοκιμάσουν τον διαχωρισμό μουσικής κάνανε το καλύτερο δυνατό για να μειώσουν τη φιλοσοφία των βασικών εφαρμογών για αυτήν την κοινότητα. Επίσης το μοντέλο έχει τη δυνατότητα άμεσης εκπαίδευσης σε ένα σύνολο δεδομένων που μπορεί να κατέβει αυτόματα.

Επιπρόσθετα η δημιουργία του Open-Unmix είναι μια προσπάθεια να παρέχει μια αξιόπιστη υλοποίηση που θα τη τηρεί τις καθιερωμένες πρακτικές προγραμματισμού. Πιο συγκεκριμένα παρέχει αναπαραγώγιμο κώδικα. Αυτό σημαίνει ότι παρέχονται τα πάντα για την ακριβή αναπαραγωγή των πειραμάτων και την εμφάνιση των αποτελεσμάτων. Επίσης παρέχει προ-εκπαιδευμένα μοντέλα που επιτρέπουν στον χρήστη να χρησιμοποιήσει το μοντέλο αμέσως ή να το προσαρμόσει σε δικά του δεδομένα. Τέλος, το Open-Unmix παρέχει δοκιμές. Η έκδοση περιλαμβάνει δοκιμές μονάδας και παλινδρόμησης χρήσιμες για την οργάνωση μελλοντικής συνεργασίας.

Το Open-Unmix είναι ένα έργο επικεντρωμένο στην κοινότητα. Συνεπώς, οι δημιουργοί το δημιούργησαν για να ενθαρρύνουν την κοινότητα να υποβάλει διορθώσεις σφαλμάτων και σχόλια για να βελτιώσει την υπολογιστική του απόδοση. Ωστόσο οι ίδιοι δεν αναζητούν αλλαγές που εστιάζουν μόνο στη βελτίωση της απόδοσης διαχωρισμού. Θέλουν πολλοί ερευνητές να χρησιμοποιούν το λογισμικό ως βάση για την έρευνα τους και να επεκτείνουν με τις επιλογές τους τον κώδικα. Καταλήγοντας το Open-Unmix είναι ένα μοντέλο το οποίο δημιουργήθηκε ώστε όλο και περισσότεροι ερευνητές να αρχίσουν να ασχολούνται με τον πρόβλημα του διαχωρισμού του ήχου ξεκινώντας έχοντας μια πολύ σημαντική βάση.

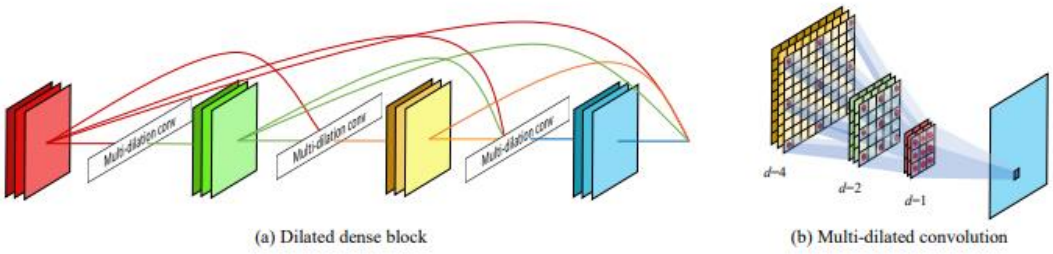
3.4 Μοντέλο DenseNet(D3Net)

Το τρίτο μοντέλο με το οποίο θα ασχοληθούμε στην εργασία μας ονομάζεται DenseNet (D3Net). Ένας ερευνητής που ονομαζόταν Takahashi εφάρμοσε μια αρχιτεκτονική CNN με ένα πυκνό μοτίβο συνδεσιμότητας που ονομάζεται DenseNet στο διαχωρισμό μουσικής λαμβάνοντας αποτελέσματα τελευταίας τεχνολογίας. Μια τέτοια πυκνή συνδεσιμότητα επιτρέπει τη μέγιστη ροή πληροφοριών και βαθύτερο CNN διατηρώντας παράλληλα μικρό το μέγεθος του μοντέλου με την αποτελεσματική επαναχρησιμοποίηση των ενδιάμεσων αναπαραστάσεων των προηγούμενων στρωμάτων.

Παρόλο που η δομή του down-up sampling και η διασταλμένη συνέλιξη επιτρέπουν ένα μεγάλο δεκτικό πεδίο κάθε στρώμα στο δίκτυο βλέπει μόνο μια ανάλυση κάθε

φορά. Το DenseNet αντιμετωπίζει εν μέρει αυτό το πρόβλημα μέσω της πυκνής σύνδεσης παράλειψης που επιτρέπει την άμεση συγκέντρωση χαρακτηριστικών από πρώιμα επίπεδα και λειτουργίες σε μεταγενέστερα στρώματα σε ένα μόνο στρώμα συνέλιξης. Ωστόσο ενδέχεται να είναι πολύ αργή η μετατροπή των τοπικών χαρακτηριστικών σε καθολικά χαρακτηριστικά και είναι αναποτελεσματική η ύπαρξη πολλών παραμέτρων ειδικά για δεδομένα υψηλής ανάλυσης.

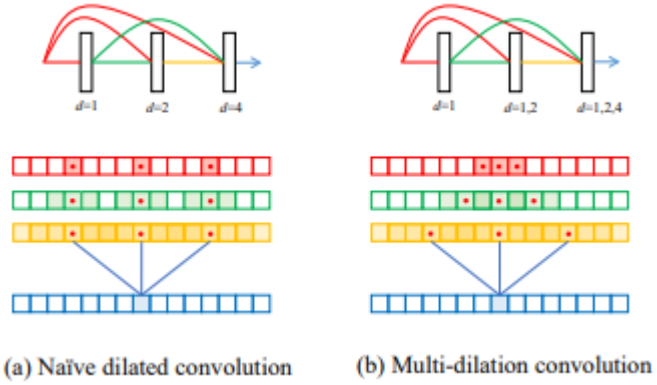
Συνδυάζοντας τα πλεονεκτήματα του DenseNet και της διασταλμένης συνέλιξης οι ερευνητές Takahashi και Mitsufuji δημιουργήσαν μια νέα αρχιτεκτονική δικτύου που ονόμασαν διασταλμένο DenseNet(D2Net). Για να το καταφέρουν αυτό πρότειναν μια πολυεπίπεδη στρώση συνέλιξης που έχει έναν συντελεστή πολλαπλής διαστολής μέσα σε ένα στρώμα. Ο συντελεστής διαστολής εξαρτάται από τη σύνδεση παράκαμψης από τη οποία προέρχονται τα κανάλια όπως βλέπουμε στην παρακάτω εικόνα. Η πολυεπίπεδη συνέλιξη μπορεί αν αποτρέψει την ψευδαίσθηση που συμβαίνει όταν εφαρμόζεται μια τυπική διασταλμένη συνέλιξη για να εμφανίσει χάρτες με δεκτικά πεδία μικρότερα από τον συντελεστή διαστολής. Παρόλο που έχει προταθεί ένας αφελής συνδυασμός του DenseNet με διαστολή χρησιμοποιούνται τυπικές διασταλμένες συστροφές και οι παράγοντες διαστολής καθορίζονται ανάλογα με το βάθος του στρώματος γεγονός που προκαλεί σημαντική αλλοίωση. Επίσης οι δημιουργοί πρότειναν μια αρχιτεκτονική διασταλμένων πυκνών μπλοκ για την αποτελεσματική επανάληψη των παραγόντων διαστολής πολλές φορές με πυκνές συνδέσεις που εξασφαλίζουν το επαρκές νάθος που απαιτείται για την μοντελοποίηση κάθε ανάλυσης. Η αρχιτεκτονική ονομάστηκε D3Net.



(a) Dilated dense block (b) Multi-dilated convolution

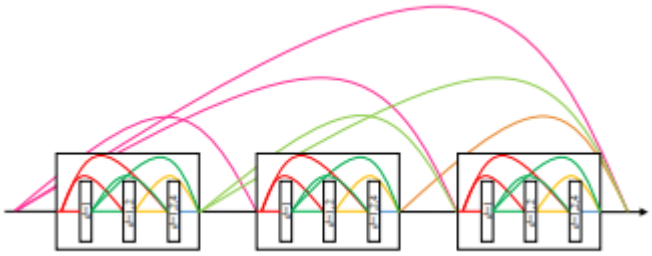
Στο σημείο αυτό θα μιλήσουμε για την πολυεπίπεδη συνέλιξη του DenseNet. Στο DenseNet οι έξοδοι του συνελικτικού στρώματος x_l υπολογίζονται χρησιμοποιώντας φίλτρα k_l και εξόδους όλων των προηγούμενων στρωμάτων ως $x_l = \psi([x_0,x_1,\ldots,x_{l-1}])\ (\cdot)\ k_l$. Όπου ψ() δηλώνει τη σύνθετη λειτουργία κανονικοποίησης παρτίδας και μη γραμμικότητας $[x_0,x_1,\ldots,x_{l-1}]$ τη συνένωση των χαρακτηριστικών χαρτών από 1,…, l-1 στρώματα και (·) τη συνέλιξη. Ένας αφελής τρόπος ενσωμάτωσης της διασταλμένης συνέλιξης είναι η αντικατάσταση της συνέλιξης (·) με τη διασταλμένη συνέλιξη $(\cdot)_d$ με τον παράγοντα διαστολής d = 2l – 1. Ωστόσο αυτό προκαλεί ένα σοβαρό πρόβλημα ψευδαισθήσεων. Για να ξεπεραστεί αυτό το πρόβλημα χρησιμοποιούμε την πολλαπλή συνέλιξη $(\cdot)_l^m$ που ορίζεται ως $Y_l\ (\cdot)_l^m\ k_l = \sum_{i=0}^{l-1} y_i\ (\cdot)\ d_i\ k_l^i$ όπου $Y_l = [y_0,y_1,\ldots,y_{l-1}] = \psi([x_0,x_1,\ldots,x_{l-1}])$ είναι η σύνθετη έξοδος στρώματος k_l^i το υποσύνολο φίλτρων που

αντιστοιχούν στην i-στη σύνδεση παράκαμψης και $d_i = 2^i$. Όπως απεικονίζεται στην παρακάτω εικόνα το DenseNet με την προτεινόμενη πολυδιάστατη συνέλιξη έχει διαφορετικούς παράγοντες διαστολής ανάλογα με το στρώμα από το οποίο προέρχεται το κανάλι. Αυτό επιτρέπει στο δεκτικό πεδίο να καλύπτει το πεδίο εισόδου χωρίς απώλεια κάλυψης μεταξύ των δειγμάτων που θα εφαρμοστούν τα φίλτρα και ως εκ τούτου να μάθουν κατάλληλα φίλτρα για να αποτρέψουν την αλλοίωση. Ένα πλεονέκτημα του διασταλμένου πυκνού μπλοκ είναι η ικανότητα του να ενσωματώνει πληροφορίες από πολύ τοπικό σε εκθετικά μεγάλο δεκτικό πεδίο μέσα σε ένα μόνο στρώμα. Αυτή η γρήγορη ροή πληροφοριών παρέχει μεγαλύτερη ευελιξία στη μοντελοποίηση πληροφοριών σε ένα ευρύ φάσμα αναλύσεων.



(a) Naïve dilated convolution (b) Multi-dilation convolution

Σε αυτό το σημείο θα αναφερθούμε στο D3Net. Παρόλο που το μπλοκ D2 παρέχει ένα εκθετικά μεγάλο δεκτικό πεδίο καθώς αυξάνεται ο αριθμός των επιπέδων αξίζει να παρέχεται επαρκής ευελιξία για τον μετασχηματισμό των χαρακτηριστικών χαρτών σε κάθε ανάλυση. Στο μοντέλο WaveNet οι συντελεστές διαστολής επαναφέρονται στο ένα μετά τη στοίβαξη και επανάληψη πολλών στρωμάτων. Δηλαδή οι συντελεστές διαστολής στο l-στο στρώμα δίνεται με $d_l = 2^{l-1 \bmod M}$ όπου mod είναι η λειτουργία modulo και M είναι ο αριθμός των στρωμάτων στα οποία ο συντελεστής διαστολής διπλασιάζεται. Εμπνευσμένοι από το μοντέλο αυτό οι δημιουργοί του D3Net προτείνουν μια σύνθετη αρχιτεκτονική των μπλοκ D2 όπως φαίνεται στην παρακάτω εικόνα.



Τα μπλοκ D2 θεωρούνται ως ενιαία σύνθετα στρώματα και συνδέονται πυκνά με τον ίδιο τρόπο όπως στο ίδιο το μπλοκ D2. Επίσης χρησιμοποιούν έναν μηχανισμό μείωσης καναλιών στο τέλος κάθε μπλοκ D2 για να μετριάσουν την ανάπτυξη ενός

υπερβολικού αριθμού καναλιών και έτσι να βελτιώσουν την υπολογιστική αποδοτικότητα. Η μείωση του καναλιού μπορεί να πραγματοποιηθεί είτε με συνέλιξη 1x1 είτε απλά περνώντας την έξοδο των τελευταίων N στρωμάτων στο επόμενο μπλοκ. Στο μοντέλο αυτό χρησιμοποιούμε την τελευταία προσέγγιση καθώς τα χαρακτηριστικά απόδοσης και των δύο μεθόδων είναι παρόμοια αλλά η 1η προσέγγιση απαιτεί ελαφρώς περισσότερους υπολογισμούς.

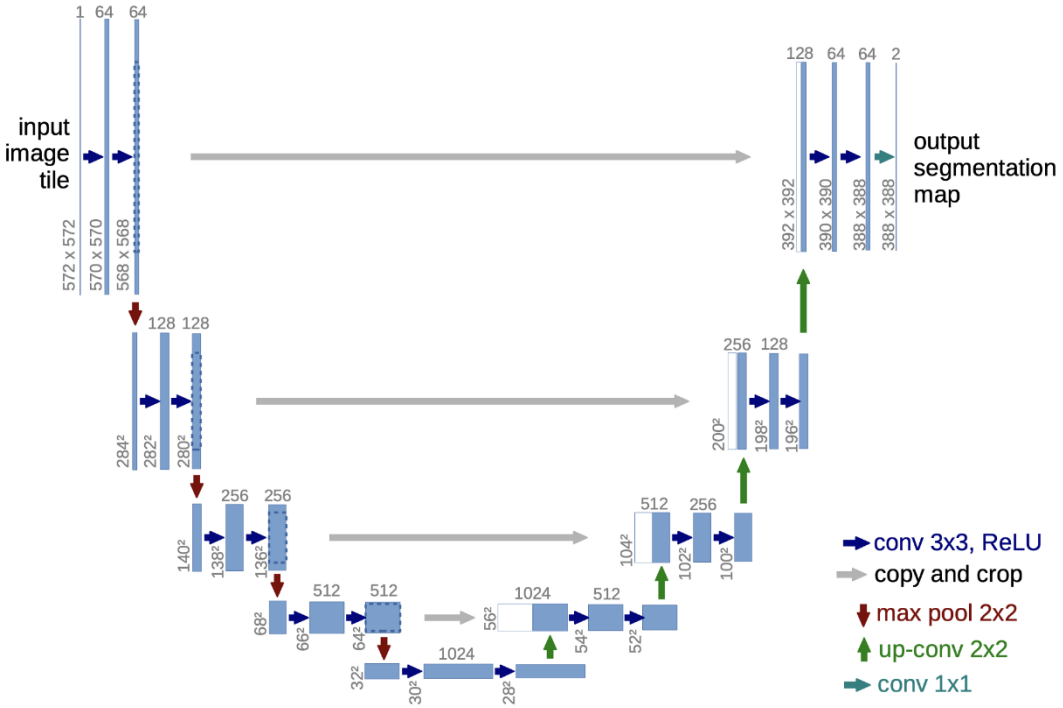
Όσων αναφορά την αρχιτεκτονική του μοντέλου το D3Net χρησιμοποιεί την αρχιτεκτονική πολλαπλών κλιμακίων στην οποία έχουμε ενότητες που προορίζονταν για ζώνες και μια ενότητα πλήρους ζώνης η καθεμία με αρχιτεκτονική κωδικοποιητή – αποκωδικοποιητή συμφόρησης με συνδέσεις παραλείψεων. Οι έξοδοι δικτύου χρησιμοποιούνται για τον υπολογισμό του πολυκάναλου φίλτρου Wiener για τη λήψη των τελικών διαχωρισμών όπως συνήθως γίνεται σε μεθόδους διαχωρισμού πηγών ήχου τομέα συχνοτήτων.

Καταλήγοντας το μοντέλο D3Net πρόκειται για μια νέα αρχιτεκτονική νευρωνικών δικτύων. Το D3Net χρησιμοποιεί την πολυεπίπεδη συνέλιξη με πυκνές συνδέσεις παράλειψης που επιτρέπει τις τοπικές και παγκόσμιες πληροφορίες χαρακτηριστικών να μοντελοποιούνται ταυτόχρονα σε ένα μόνο στρώμα.

3.5 Μοντέλο Spleeter

Στην τελευταία και 5η υποενότητα αυτού του κεφαλαίου θα παρουσιάσουμε το τελευταίο μοντέλο που θα χρησιμοποιήσουμε στην εργασία μας για το διαχωρισμό πηγών μουσικής που ονομάζεται Spleeter. Το Spleeter σχεδιάστηκε με γνώμονα την ευκολία χρήσης, την απόδοση διαχωρισμού και την ταχύτητα. Επίσης το μοντέλο αυτό βασίζεται στο Tensorflow και καθιστά δυνατό το διαχωρισμό αρχείων ήχου με μία μόνο γραμμή εντολών και την εκπαίδευση μοντέλων διαχωρισμού ή τη βελτίωση προ-εκπαιδευμένων μοντέλων. Το Spleeter είναι επίσης πολύ γρήγορο καθώς μπορεί να διαχωρίσει ένα μίγμα ήχου σε 4 μέρη 100 φορές πιο γρήγορα από τον πραγματικό χρόνο σε μία GPU.

Οι δημιουργοί του Spleeter το κυκλοφορούν με προ-εκπαιδευμένα μοντέλα για να βοηθήσουν την κοινότητα ανάκτησης μουσικής πληροφοριών να αξιοποιήσει τη δύναμη του διαχωρισμού πηγής σε διάφορες εργασίες. Οι εργασίες αυτές μπορεί να είναι ανάλυση φωνητικών στίχων, ηχογράφηση μουσικής, αναγνώριση τραγουδιστή και ταξινόμηση είδους ενός τραγουδιού. Το Spleeter καθιστά επίσης δυνατή τη λεπτομερή ρύθμιση των παρεχόμενων μοντέλων τελευταίας τεχνολογίας για να προσαρμόσουμε το σύστημα σε μια συγκεκριμένη περίπτωση χρήσης. Τέλος η ύπαρξη ενός διαθέσιμου εργαλείου διαχωρισμού πηγών όπως το Spleeter θα επιτρέψει στους ερευνητές να συγκρίνουν τις επιδόσεις των νέων μοντέλων τους με ένα σύγχρονο μοντέλο στα δικά τους σύνολα δεδομένων αντί για το Musdb που είναι και το πιο δημοφιλές από όλα.



Τώρα θα μιλήσουμε για κάποιες λεπτομέρειες της εφαρμογής του Spleeter. Αρχικά το Spleeter περιέχει εκπαιδευμένα μοντέλα για διαχωρισμό φωνητικών, διαχωρισμό σε 4 μέρη(φωνητικά, μπάσο, τύμπανο και άλλα) και διαχωρισμό σε 5 μέρη(φωνητικά, μπάσο, τύμπανο, πιάνο και άλλα) τον οποίο δεν τον έχει πραγματοποιήσει άλλο μοντέλο. Το Spleeter χρησιμοποιεί αρχιτεκτονική U-Net 12 επιπέδων και συγκεκριμένα 6 στρώματα για τον κωδικοποιητή και 6 στρώματα για τον αποκωδικοποιητή. Τα προ-εκπαιδευμένα μοντέλα εκπαιδεύτηκαν σε εσωτερικά σύνολα δεδομένων και ο χρόνος εκπαίδευσης διήρκησε περίπου μια βδομάδα σε μία GPU. Ο διαχωρισμός έγινε στη συνέχεια από τα φασματογράμματα πηγής χρησιμοποιώντας φιλτράρισμα Wiener πολλαπλών καναλιών. Η εκπαίδευση και το συμπέρασμα εφαρμόζονται στο Tensorflow το οποίο καθιστά δυνατή την εκτέλεση του κώδικα στην κεντρική μονάδα επεξεργασίας GPU.

Τέλος θα αναφερθούμε στην ταχύτητα του Spleeter καθώς είναι ένα από τα κύρια χαρακτηριστικά στο οποίο έδωσαν μεγάλη βάση οι δημιουργοί του μοντέλου. Καθώς ολόκληρος ο αγωγός διαχωρισμού μπορεί να τρέξει σε μια GPU και το μοντέλο βασίζεται σε ένα CNN το μοντέλο μπορεί να τρέξει πού γρήγορα. Για παράδειγμα ο Spleeter είναι σε θέση να διαχωρίσει ολόκληρο το σύνολο δεδομένων Musdb που είναι συνολικά 3 ώρες και 27 λεπτά μουσικών ήχων σε 4 μέρη σε λιγότερο από 2 λεπτά αν διαθέτουμε μια GPU τελευταίας τεχνολογίας. Επίσης το Spleeter μπορεί στη συνέχεια να χωρίσει σε 4 μέρη 100 δευτερόλεπτα στερεοφωνικού ήχου σε λιγότερο από 1 δευτερόλεπτο γεγονός που το καθιστά πολύ χρήσιμο για την αποτελεσματική επεξεργασία μεγάλων συνόλων δεδομένων.

Καταλήγοντας το Spleeter είναι ένα από τα πιο σημαντικά μοντέλα που γνωρίσαμε στην εργασία μας. Αυτό γιατί είναι πολύ γρήγορο διαθέτει ένα ξεχωριστό χαρακτηριστικό που είναι ο διαχωρισμός ήχου σε 5 μέρη κάτι που το κάνει μοναδικό στο είδος του και τέλος είναι ένα μοντέλο στο οποίο οι ερευνητές μπορούν να βασιστούν ώστε να ξεκινήσουν την έρευνα στον διαχωρισμό μουσικής. Σε αυτό το σημείο λοιπόν ολοκληρώσαμε όσα πρέπει να γνωρίζουμε για το θεωρητικό κομμάτι των μοντέλων που θα ασχοληθούμε και είμαστε σε θέση να περάσουμε στο πειραματικό και πιο ενδιαφέρον κομμάτι της εργασίας μας.

4. Πειράματα-Αξιολόγηση

4.1 Πειράματα

Στο σημείο αυτό και αφού έχουμε κατανοήσει το θεωρητικό υπόβαθρο της εργασίας μας μπορούμε να περάσουμε στο πειραματικό μέρος. Στο μέρος αυτό θα τρέξουμε τα τέσσερα μοντέλα στα οποία έχουμε αναφερθεί και θα αναλύσουμε τα αποτελέσματα που θα προκύψουν. Αυτό το οποίο θα κάνουμε αρχικά είναι να τρέξουμε τα μοντέλα με δεδομένα ένα dataset το οποίο είναι από τα πιο διαδεδομένα στον τομέα του διαχωρισμού ήχου. Το dataset αυτό είναι το MusDB. Στη συνέχεια για τα δεδομένα αυτά θα βγάλουμε διάφορες μετρικές τις οποίες θα συγκρίνουμε μεταξύ τους για να δούμε σε τι υπερτερεί το κάθε μοντέλο και αν στο τέλος υπάρχει το μοντέλο το οποίο ξεχωρίζει σε σχέση με τα άλλα.

Αρχικά θα μιλήσουμε για το dataset μας και πως θα το διαχειριστούμε. Το MusDB είναι ένα σύνολο δεδομένων το οποίο αποτελείται από 150 τραγούδια με πλήρη επίβλεψη σε στερεοφωνικό ήχο και δειγματοληψία στα 44100Hz. Εμείς στα μοντέλα μας θα χρησιμοποιήσουμε 84 τραγούδια για το train set, τα επόμενα 16 τραγούδια θα είναι το validation set και τα υπόλοιπα 50 τραγούδια θα είναι το test set.

Τώρα θα μιλήσουμε για τις μετρικές που θα βγάλουμε από τα μοντέλα και με τις οποίες θα κάνουμε την αξιολόγηση των μοντέλων. Οι μετρικές αυτές θα είναι οι εξής τέσσερις: SDR(Signal to Distortion Ratio), SAR(Signal to Artifacts Ratio), SIR(Signal to Interference Ratio) και ISR(Image to Spatial Distortion Ratio). Οι μετρικές αυτές είναι μέχρι σήμερα οι πιο ευρέως χρησιμοποιούμενες μέθοδοι για την αξιολόγηση της εξόδου ενός συστήματος διαχωρισμού πηγής. Όσο μεγαλύτερη η τιμή της μετρικής τόσο καλύτερο το αποτέλεσμα. Το SDR θεωρείται συνήθως ως ένα συνολικό μέτρο για το πόσο καλά ακούγεται μια πηγή. Το SAR ερμηνεύεται ως το ποσό των ανεπιθύμητων τεχνουργημάτων που έχει μια εκτίμηση πηγής σε σχέση με την πραγματική πηγή. Το SIR ερμηνεύεται ως το ποσό των άλλων πηγών που μπορούν να ακουστούν σε μια εκτίμηση πηγής ή αλλιώς η διαρροή του ήχου. Το ISR ερμηνεύεται παρόμοια με την μετρική SDR.

Στην συνέχεια περνάμε στον τρόπο με τον οποίο θα κάνουμε τα πειράματα μας. Για τις ανάγκες της εργασίας μας θα χρησιμοποιήσουμε το Colab του Google. Πρόκειται για ένα εργαλείο που προσφέρει η Google και μπορούμε να τρέξουμε τον κώδικα μας χωρίς να χρησιμοποιήσουμε τη μνήμη και τη GPU μας καθώς προσφέρεται από το εργαλείο αυτό. Με τον τρόπο αυτό άμα δεν έχουμε τον κατάλληλο εξοπλισμό μπορούμε με το εργαλείο αυτό να έχουμε ένα γρήγορο και επιθυμητό αποτέλεσμα. Οπότε θα τρέξουμε και τα τέσσερα μοντέλα μας με τη βοήθεια του Colab και θα πάρουμε τα αποτελέσματα μας.

Στην αρχή το αποτέλεσμα που θέλουμε είναι να βγάλουμε τις μετρικές. Κάθε τραγούδι χωρίζεται με την βοήθεια των μοντέλων σε 4 μέρη. Τα μέρη αυτά είναι το μπάσο, τα ντράμς, τα φωνητικά και οι υπόλοιποι ήχοι του κομματιού. Εμείς για κάθε ένα από τα μέρη αυτά θα υπολογίσουμε τις μετρικές μας ώστε να μπορούμε να κάνουμε μια ολοκληρωτική ανάλυση των αποτελεσμάτων. Οπότε παρακάτω παρουσιάζεται ένας πίνακας στον οποίο υπάρχουν όλες οι τιμές των μετρικών σε κάθε μοντέλο και κάθε μέρος αντίστοιχα.

	Demucs	**Open-Unmix**	**D3Net**	**Spleeter**
Vocals SDR	7.05	6.32	7.24	6.86
Vocals SIR	13.94	13.33	14.51	15.86
Vocals SAR	7.00	6.52	7.44	6.99
Vocals ISR	12.04	11.93	11.97	11.95
Bass SDR	6.70	5.23	5.25	5.51
Bass SIR	13.03	10.93	10.83	10.30
Bass SAR	6.68	6.34	6.17	5.96
Bass ISR	9.99	9.23	9.40	9.61
Drums SDR	7.08	5.73	7.01	6.71
Drums SIR	13.74	11.12	13.22	13.67
Drums SAR	7.04	6.02	6.29	6.54
Drums ISR	11.96	10.51	10.60	10.69
Other SDR	4.47	4.02	4.53	4.55
Other SIR	7.11	6.59	7.38	8.16
Other SAR	5.26	4.74	4.56	4.88
Other ISR	10.86	9.31	9.76	9.87

Τα παραπάνω αποτελέσματα θα τα σχολιάσουμε στην επόμενη υποενότητα που θα γίνει η γενική αξιολόγηση των μοντέλων με βάση αυτές τις τιμές.

Το επόμενο βήμα με το οποίο θα προχωρήσουμε την εργασία μας είναι να κάνουμε τον διαχωρισμό σε κάποια μουσικά κομμάτια. Πιο συγκεκριμένα, ανεξάρτητα από το

σύνολο δεδομένων που χρησιμοποιήσαμε για να βγάλουμε τις μετρικές τώρα θα χρησιμοποιήσουμε τέσσερα τραγούδια της δικιάς μας επιλογής. Τα τραγούδια αυτά φροντίσαμε να είναι γνωστά στο ευρύ κοινό και να ανήκουν στην κατηγορία της ροκ μουσικής καθώς στο είδος αυτό υπάρχουν περισσότερα μουσικά όργανα τα οποία συμπεριλαμβάνουν το μπάσο και τα ντραμς τα οποία θέλουμε. Τα τραγούδια που επιλέξαμε είναι τα Back In Black(AC_DC), TNT(AC_DC), Californication(Red Hot Chili Peppers) και Wake me up when September ends(Green Day). Οπότε παρακάτω για κάθε τραγούδι θα υπάρχει και η αντίστοιχη ανάλυση σε τέσσερα μέρη από κάθε μοντέλο. Τα τέσσερα μέρη αυτά θα είναι τα ντραμς, μπάσο, φωνητικά και τα υπόλοιπα. Έχουμε δύο εξαιρέσεις. Η πρώτη εξαίρεση που υπάρχει είναι στο μοντέλο Spleeter. Στο μοντέλο αυτό γίνεται διαχωρισμός σε δύο μέρη και όχι σε τέσσερα. Τα δύο αυτά μέρη είναι τα φωνητικά και η συνολική συνοδεία του τραγουδιού. Η δεύτερη εξαίρεση είναι στο μοντέλο OpenUnmix. Εδώ το μοντέλο δεν μπορεί να χωρίσει όλο το τραγούδι με ένα τρέξιμο οπότε θα έχουμε δύο αρχεία για κάθε μέρος όπου στο 1ο αρχείο θα είναι το πρώτο μισό του τραγουδιού και στο δεύτερο αρχείο θα είναι το υπόλοιπο τραγούδι. Όλα αυτά που περιγράφουμε τώρα παρουσιάζονται παρακάτω και μπορούμε να τα ακούσουμε για να δούμε και το πρακτικό αποτέλεσμα.

	Back In Black(AC_DC)	TNT(AC_DC)	Californication (Red Hot Chili Peppers)	Wake me up when September ends (Green Day)
Demucs bass	 bass.wav	 bass.wav	 bass.wav	 bass.wav
Demucs drums	 drums.wav	 drums.wav	 drums.wav	 drums.wav
Demucs other	 other.wav	 other.wav	 other.wav	 other.wav
Demucs vocals	 vocals.wav	 vocals.wav	 vocals.wav	 vocals.wav

OpenUnmix bass	bass.wav bass2.wav	bass.wav bass2.wav	bass.wav bass2.wav bass3.wav	bass.wav bass2.wav
OpenUnmix drums	drums.wav drums2.wav	drums.wav drums2.wav	drums.wav drums2.wav drums3.wav	drums.wav drums2.wav
OpenUnmix other	other.wav other2.wav	other.wav other2.wav	other.wav other2.wav other3.wav	other.wav others2.wav
OpenUnmix vocals	vocals.wav vocals2.wav	vocals.wav vocals2.wav	vocals.wav vocals2.wav vocals3.wav	vocals.wav vocals2.wav
D3Net bass	bass.wav	bass.wav	bass.wav	bass.wav
D3Net drums	drums.wav	drums.wav	drums.wav	drums.wav

D3Net other	other.wav	other.wav	other.wav	other.wav
D3Net vocals	vocals.wav	vocals.wav	vocals.wav	vocals.wav
Spleeter accompaniment	accompaniment.wav	accompaniment.wav	accompaniment.wav	accompaniment.wav
Spleeter vocals	vocals.wav	vocals.wav	vocals.wav	vocals.wav

4.2 Αξιολόγηση

Η υποενότητα αυτή είναι ίσως η πιο σημαντική για την εργασία μας καθώς και η πιο ενδιαφέρουσα. Έχουμε πάρει τα αποτελέσματα μας στην προηγούμενη υποενότητα και τώρα είμαστε σε θέση να τα αξιολογήσουμε. Αρχικά θα ξεκινήσουμε αξιολογώντας τις μετρικές που βγάλαμε. Να αναφέρουμε ξανά ότι όσο μεγαλύτερη είναι η τιμή σε μία μετρική τόσο καλύτερο το αποτέλεσμα. Βλέποντας λοιπόν τις τιμές των μετρικών μπορούμε να ξεχωρίσουμε το μοντέλο το οποίο έχει τις καλύτερες τιμές στις περισσότερες μετρικές αλλά και το μοντέλο με τα χειρότερα αποτελέσματα σε σύγκριση με τα υπόλοιπα. Πιο συγκεκριμένα, μπορούμε πολύ εύκολα να δούμε ότι το μοντέλο το οποίο κερδίζει τα υπόλοιπα είναι το Demucs. Αυτό συμβαίνει γιατί το Demucs έχει την καλύτερη τιμή σε μετρική στις 11 από τις 16 περιπτώσεις που τρέξαμε. Στις υπόλοιπες περιπτώσεις είχαμε καλύτερη τιμή στο μοντέλο Spleeter για τα Vocals SIR, Other SDR, Other SIR και στο μοντέλο D3Net για τα Vocals SDR και Vocals SAR. Σε καμία από τις περιπτώσεις μας δεν είχε καλύτερη τιμή το μοντέλο OpenUnmix.

Οπότε μπορούμε να πούμε ότι έχουμε ως καλύτερο μοντέλο το Demucs και ως χειρότερο μοντέλο το OpenUnmix σε σύγκριση πάντα με τα τέσσερα μοντέλα που εμείς τρέξαμε. Επίσης να αναφέρουμε ότι το ιδανικό μοντέλο θα ήταν αυτό το οποίο θα υπερτερούσε σε όλες τις περιπτώσεις σε σχέση με τα άλλα μοντέλα αλλά αυτό θα ήταν κάτι πολύ δύσκολο. Αυτό συμβαίνει γιατί τις περισσότερες φορές για να αυξήσουμε την τιμή μια μετρικής αλλάζουμε την αρχιτεκτονική του μοντέλου με τρόπο που μπορεί να μειωθεί μία άλλη.

Στη συνέχεια θα αναφερθούμε στο διαχωρισμό που κάναμε στα τέσσερα μουσικά κομμάτια που επιλέξαμε. Η σύγκριση εδώ δεν μπορεί να γίνει εύκολα καθώς δεν υπάρχει κάποια τιμή για να συγκρίνουμε αλλά μόνο ήχος. Πρέπει λοιπόν να ακούσουμε όλα τα αρχεία διαχωρισμού και να τα συγκρίνουμε με άλλα κριτήρια αυτή τη φορά όπως είναι το πόσο καθαρός είναι ο ήχος ή αν έγινε σωστά ο διαχωρισμός που θέλαμε.

Αφού ακούσαμε όλα τα αρχεία ήχου μπορούμε να πούμε ότι ο διαχωρισμός υπάρχει και είναι σωστός σε όλα τα τραγούδια πράγμα που ήταν και ο κύριος στόχος της εργασίας μας. Μετά μπορούμε να συγκρίνουμε μια συγκεκριμένη κατηγορία για παράδειγμα τα vocals στο τραγούδι Wake me up When September Ends. Τα vocals στο demucs δεν ακούγονται πολύ καθαρά σε σχέση με το D3Net στο οποίο έχουμε πιο καθαρό ήχο. Στο OpenUnmix έχουμε τον πιο καθαρό ήχο από όλα τα μοντέλα καθώς και πιο δυνατό σε ένταση ήχο. Τέλος στο Spleeter ο ήχος ακούγεται κάπως πιο βαθύς και όχι τόσο καθαρός. Αυτά τα συμπεράσματα ισχύουν και στις υπόλοιπες κατηγορίες όπως στα Drums, το Bass και το Other. Τα συμπεράσματα αυτά μπορεί να διαφέρουν από άνθρωπο σε άνθρωπο καθώς είναι με βάση την ακοή και το πώς ο κάθε άνθρωπος αντιλαμβάνεται τον ήχο.

Με αυτά που αναφέραμε εμείς στην εργασία αυτή έχουμε έρθει σε ένα πολύ ενδιαφέρον αποτέλεσμα. Όταν αξιολογήσαμε με βάση τις μετρικές καταλήξαμε στο ότι το demucs είναι το καλύτερο μοντέλο και το OpenUnmix το χειρότερο. Σε αντίθεση με το συμπέρασμα αυτό όταν ακούσαμε τα αρχεία ήχου όπου είχαμε κάνει το διαχωρισμό των δικών μας μουσικών κομματιών καταλήξαμε στο ότι στο OpenUnmix είχαμε τον καλύτερο ήχο και στο demucs το χειρότερο. Αυτό το αποτέλεσμα είναι αρκετά ενδιαφέρον καθώς μπορούμε να πούμε ότι μπορεί στη θεωρία ένα μοντέλο να φαίνεται βάση τιμών το καλύτερο άλλα στην πράξη αυτό μπορεί να αλλάξει ανάλογα με τις συνθήκες που χρειάζεται κάποιος.

Στο σημείο αυτό τελειώσαμε και την 4^{η} και πιο σημαντική ενότητα της εργασίας μας που είχε το πειραματικό μέρος. Στη συνέχεια θα περάσουμε στην τελευταία ενότητα στην οποία θα εκφράσουμε τα συμπεράσματα μας και θα ολοκληρώσουμε την διπλωματική μας εργασία.

5. Συμπεράσματα-Επίλογος

Στην τελευταία αυτή ενότητα θα εκφράσουμε τα συμπεράσματα που βγάλαμε από την εργασία μας. Τα συμπεράσματα βγαίνουν κατά κύριο λόγο από το πειραματικό μέρος και την αξιολόγηση που κάναμε καθώς το θεωρητικό μέρος δεν μπορεί να μας βοηθήσει στην ενότητα αυτή.

Το πρώτο συμπέρασμα βγαίνει με βάση τις τιμές των μετρικών των μοντέλων που τρέξαμε. Όπως είδαμε στις 11 από τις 16 τιμές που βγάλαμε το μοντέλο Demucs υπερτερούσε σε σχέση με τα άλλα μοντέλα. Οπότε καταλήγουμε στο ότι το Demucs είναι το καλύτερο μοντέλο και αυτό το οποίο θα προτείναμε σε κάποιον για το διαχωρισμό ενός μουσικού σήματος ή σε έναν ερευνητή για να αναπτύξει ένα ακόμα καλύτερο μοντέλο. Από την άλλη μεριά το μοντέλο με την χειρότερη απόδοση είναι το OpenUnmix. Εννοείται βέβαια ότι και τα άλλα μοντέλα είναι αποτελεσματικά αλλά το Demucs είναι αυτό που ξεχωρίζει.

Το επόμενο συμπέρασμα που βγάλαμε είναι με βάση το διαχωρισμό τεσσάρων μουσικών κομματιών της επιλογής μας. Εδώ το καλύτερο μοντέλο με βάση αυτό που ακούσαμε είναι το μοντέλο OpenUnmix το οποίο είχε τον πιο καθαρό και δυνατό ήχο. Αντίθετα το μοντέλο με τον χειρότερο ήχο ήταν το μοντέλο Demucs. Οπότε εδώ θα προτείναμε το OpenUnmix για το διαχωρισμό ενός μουσικού κομματιού.

Στο σημείο αυτό βγάζουμε άλλο ένα συμπέρασμα στο οποίο αναφερθήκαμε και στην προηγούμενη ενότητα. Βλέπουμε ότι με βάση τις μετρικές έχουμε ως αποδοτικότερο μοντέλο το Demucs και ως χειρότερο το OpenUnmix ενώ με βάση τα αρχεία διαχωρισμού έχουμε το ακριβώς αντίθετο. Όπως αναφέραμε και παραπάνω το 2^{ο} συμπέρασμα μπορεί να διαφέρει από άνθρωπο σε άνθρωπο καθώς δεν υπάρχει κάποια μετρική να συγκρίνουμε αλλά μόνο ήχος. Οπότε εδώ βγάζουμε το συμπέρασμα ότι το Demucs είναι στη θεωρία το καλύτερο μοντέλο άλλα στην πράξη μπορεί κάποιο άλλο μοντέλο να είναι καλύτερο ανάλογα με αυτό που θέλουμε ως αποτέλεσμα. Για παράδειγμα εμείς θεωρούμε ότι το OpenUnmix έχει τον καλύτερο ήχο άλλα κάποιος ειδικός που ασχολείται με την μουσική μπορεί να έχει διαφορετική άποψη.

Εν κατακλείδι, εμείς στην εργασία μας θα προτείναμε το μοντέλο Demucs. Στο σημείο αυτό έχουμε ολοκληρώσει την διπλωματική εργασία μας συγκρίνοντας τις μετρικές σε τέσσερα διαφορετικά μοντέλα μηχανικής μάθησης και κάνοντας διαχωρισμό σε τέσσερα τραγούδια της επιλογής μας. Τα αποτελέσματα μας ήταν λογικά και αυτά που θέλαμε να πετύχουμε καθώς πετύχαμε το στόχο μας που ήταν να λύσουμε το πρόβλημα του διαχωρισμού μουσικών σημάτων με τη βοήθεια τεχνικών μηχανικής μάθησης.

6. Βιβλιογραφία

Alexandre Défossez, Nicolas Usunier, Léon Bottou, Francis Bach, Music Source Separation in the Waveform Domain. Facebook AI Research; 2021.

Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus and Yuki Mitsufuji, Open-Unmix - A Reference Implementation for Music Source Separation. Stöter et al., 2019.

Naoya Takahashi, Yuki Mitsufuji, D3NET: DENSELY CONNECTED MULTIDILATED DENSENET FOR MUSIC SOURCE SEPARATION. Sony Corporation, Japan, 2021.

Romain Hennequin, Anis Khlif, Felix Voituret, Manuel Moussallam, SPLEETER: A FAST AND STATE-OF-THE ART MUSIC SOURCE SEPARATION TOOL WITH PRE-TRAINED MODELS. Deezer R&D, Paris, 2019.

P. Comon and C. Jutten . "Handbook of Blind Source Separation, Independent Component Analysis and Applications" Academic Press, ISBN 978-2-296-12827-9

P. Comon, Contrasts, Independent Component Analysis, and Blind Deconvolution, "Int. Journal Adapt. Control Sig. Proc.", Wiley, Apr. 2004.

Kevin Hughes "Blind Source Separation on Images with Shogun" http://shoguntoolbox.org/static/notebook/current/bss_image.html

Aapo Hyvarinen, Juha Karhunen, and Erkki Oja. "Independent Component Analysis" https://www.cs.helsinki.fi/u/ahyvarin/papers/bookfinal_ICA.pdf pp. 147–148, pp. 410–411, pp. 441–442, p. 448

Congedo, Marco,Gouy-Pailler, Cedric, Jutten, Christian (December 2008). "On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics". Clinical Neurophysiology. 119 (12): 2677–2686. arXiv:0812.0494. doi:10.1016/j.clinph.2008.09.007. PMID 18993114.

Jean-Francois Cardoso "Blind Signal Separation: statistical Principles" http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.9738&rep=rep1&type=pdf

Rui Li, Hongwei Li, and Fasong Wang. "Dependent Component Analysis: Concepts and Main Algorithms" http://www.jcomputers.us/vol5/jcp0504-13.pdf

Shlens, Jonathon. "A tutorial on independent component analysis." arXiv:1404.2986

Bronkhorst, Adelbert W. (2000). "The Cocktail Party Phenomenon: A Review on Speech Intelligibility in Multiple-Talker Conditions". Acta Acustica United with Acustica. 86: 117–128. Retrieved 2020-11-16.

Hawley ML, Litovsky RY, Culling JF (February 2004). "The benefit of binaural hearing in a cocktail party: effect of location and type of interferer" . The Journal of the Acoustical Society of America. 115 (2): 833–43. doi:10.1121/1.1639908.

Wood N, Cowan N (January 1995). "The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel?". Journal of Experimental Psychology. Learning, Memory, and Cognition. 21 (1): 255–60. doi:10.1037/0278-7393.21.1.255. PMID 7876773.

Conway AR, Cowan N, Bunting MF (June 2001). "The cocktail party phenomenon revisited: the importance of working memory capacity". Psychonomic Bulletin & Review. 8(2): 331–5. doi:10.3758/BF03196169. PMID 11495122.

Marinato G, Baldauf D (February 2019). "Object-based attention in complex, naturalistic auditory streams". Scientific Reports. 9 (1): 2854. doi:10.1038/s41598-019-39166-6. PMC 6393668. PMID 30814547.

Hyvärinen, Aapo (2013). "Independent component analysis: recent advances". Philosophical Transactions: Mathematical, Physical and Engineering Sciences. 371 (1984): 20110534. Bibcode:2012RSPTA.37110534H. doi:10.1098/rsta.2011.0534. ISSN 1364- 503X. JSTOR 41739975. PMC 3538438. PMID 23277597.

Isomura, Takuya; Toyoizumi, Taro (2016). "A local learning rule for independent component analysis". Scientific Reports. 6: 28073. Bibcode:2016NatSR...628073I. doi:10.1038/srep28073. PMC 4914970. PMID 27323661.

Painsky, Amichai; Rosset, Saharon; Feder, Meir (2014). Generalized Binary Independent Component Analysis. IEEE International Symposium on Information Theory (ISIT), 2014. pp. 1326–1330. doi:10.1109/ISIT.2014.6875048. ISBN 978-1-4799-5186-4. S2CID 18579555.

Delorme, A; Sejnowski, T; Makeig, S (2007). "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis". NeuroImage. 34 (4): 1443– 1449. doi:10.1016/j.neuroimage.2006.11.004. PMC 2895624. PMID 17188898.

Trapnell, C; Cacchiarelli, D; Grimsby, J (2014). "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". Nature Biotechnology. 32 (4): 381– 386. doi:10.1038/nbt.2859. PMC 4122333. PMID 24658644.