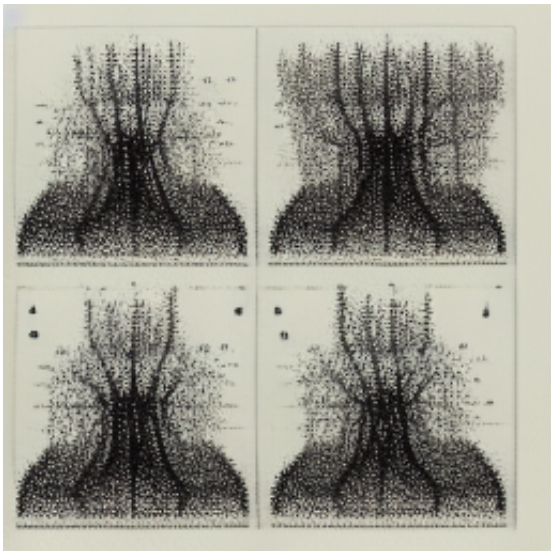
ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Συστηματική σύγκριση μεθόδων διαχωρισμού ομιλίας»



Του φοιτητή
Μίαρη Διονυσίου
Αρ. Μητρώου: 164842

Επιβλέπων
Ονοματεπώνυμο Κωνσταντίνος
Διαμαντάρας
Βαθμίδα καθηγητής

29 Ιανουαρίου 2023

Τίτλος Δ.Ε. Συστηματική σύγκριση μεθόδων διαχωρισμού ομιλίας
Κωδικός Δ.Ε. 22220
Ονοματεπώνυμο φοιτητή/ών: Διονύσιος Μίαρης
Ονοματεπώνυμο εισηγητή: Κωνσταντίνος Διαμαντάρας
Ημερομηνία ανάληψης Δ.Ε: 12-04-2022
Ημερομηνία περάτωσης Δ.Ε: 17-01-2023

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

«Αφιερώνω αυτή την εργασία στους γονείς μου που με στήριξαν με όλη τους την δύναμη»

Πρόλογος

Ένα σημαντικό κομμάτι της μηχανικής μαθήσεως είναι η προεπεξεργασία των δεδομένων. Οπότε έχοντας ασχοληθεί παλαιοτέρα με αναγνώριση εικόνων θέλησα να καλύψω κάποια κενά στις γνώσεις μου με άλλων ειδών δεδομένα. Κάπου εκεί ένας φίλος πρότεινε να ασχοληθώ με τον τομέα του speech seperation, πράγμα το οποίο μου κίνησε απευθείας το ενδιαφέρον καθώς δεν είχα ασχοληθεί με δεδομένα μορφής σημάτων. Βλέπω αυτήν την εργασία σαν μια «γέφυρα» για την αγορά εργασίας και πιστεύω ότι έχω αποκτήσει μια ευρύτερη γνώση στον τομέα της μηχανικής μαθήσεως.

Περίληψη

Σε αυτήν την εργασία γίνεται ανάλυση του τομέα διαχωρισμού ομιλίας κυρίως δυο ομιλητών. Αναλύεται το θεωρητικό υπόβαθρο που χρειάζεται για την κατανόηση του και επιπλέον έχει γίνει η εκπαίδευση και επικύρωση δυο βαθιών νευρωνικών μοντέλων μηχανικής μαθήσεως, το Convtasnet και το SudoImprovedNet. Τα αποτελέσματα του μοντέλου Convtasnet είναι ευθέως συγκρίσιμα με αποτελέσματα των ερευνητών στην βιβλιογραφία. Ενώ το μοντέλο SUdoImprovedNet δοκιμάστηκε σε διαφορετικό σετ δεδομένων από ότι στην βιβλιογραφία δείχνοντας αρκετά υψηλές αποδόσεις χωρίς κάποια περαιτέρω βελτίωση. Τέλος, γίνεται διαθέσιμος κώδικας που εμπεριέχει τα βάρη εκπαίδευσης των πειραμάτων και μοντέλα της βιβλιογραφίας όπως τα Sepformer και DualPathRnn καθώς και την δυνατότητα να εισάγουμε δική μας ηχογράφηση για διαχωρισμό.

«Συστηματική σύγκριση μεθόδων διαχωρισμού ομιλίας»

(Systematic comparison of speech separation methods)

«Διονύσιος Μίαρης»

(Dionysios Miaris)

Abstract

In this work, the speech separation domain of mainly two speakers is analyzed. The theoretical background needed for its understanding is analyzed. In addition, the training and validation of two deep neural machine learning models, such as Convtasnet and SydoImprovedNet, have been done. The results of the Convtasnet model are directly comparable to results of researchers in the literature. On the contrary for the SUdoImprovedNet model tested on a different data set than in the literature showing quite high performances without any further improvement. Finally, a code, that includes the training weights of the experiments ,other models in the literature such as Sepformer, DualPathRnn as well as the possibility to import our own recording for separation, becomes available.

Ευχαριστίες

«Θα ήθελα να ευχαριστήσω τους γονείς μου που πάρα τις συνεχείς αντιξοότητες και δυσκολίες κατάφεραν να με σπουδάσουν.»

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

Δ.Ε.	Διπλωματική Εργασία
ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
Π.Ε.	Πτυχιακή Εργασία
SAN	self-attentive network
SI-SDR	Scale-invariant signal-to-distortion ratio
Lstm	long-short term memory
BiLstm	Biderictional long-short term memory
Upit	utterance level permutation invariant training
PIT	permutation invariant training
STFT	Short-term-fourier-transform

Κεφάλαιο 1ο: Εισαγωγή

1.1 Ορισμός προβλήματος

Η διαδικασία διαχωρισμού μεμονωμένων πηγών ήχου από ένα μείγμα ηχητικών σημάτων είναι γνωστή ως διαχωρισμός πηγής (source seperation). Αυτό είναι ένα θεμελιώδες ζήτημα στην επεξεργασία σήματος ήχου, επειδή μας επιτρέπει να αναλύουμε και να χειριζόμαστε μεμονωμένους ήχους μέσα σε μια εγγραφή. Ο διαχωρισμός πηγής στην αναγνώριση ομιλίας μπορεί να χρησιμοποιηθεί για τον διαχωρισμό επικαλυπτόμενων σημάτων ομιλίας προκειμένου να βελτιωθεί η ακρίβεια της μεταγραφής. Ο διαχωρισμός πηγής μπορεί να χρησιμοποιηθεί στην αποκατάσταση ήχου για την αφαίρεση του θορύβου. Συνολικά, ο διαχωρισμός πηγής είναι ένα βασικό εργαλείο στον τομέα της επεξεργασίας σήματος ήχου, με ένα ευρύ φάσμα εφαρμογών στην παραγωγή μουσικής, την αναγνώριση ομιλίας και την αποκατάσταση ήχου. Επιτρέποντας τον διαχωρισμό μεμονωμένων πηγών ήχου από ένα μείγμα, μας επιτρέπει να αναλύουμε και να χειριζόμαστε τα ηχητικά σήματα με πιο ακριβή και αποτελεσματικό τρόπο.

Ο διαχωρισμός τυφλών πηγών (Blind Source Seperation) είναι ένα υποπεδίο της επεξεργασίας σήματος ήχου που ασχολείται με το διαχωρισμό μεμονωμένων πηγών ήχου από ένα μείγμα σημάτων ήχου χωρίς προηγούμενη γνώση των πηγών ή των ιδιοτήτων τους. Αυτό έρχεται σε αντίθεση με άλλες προσεγγίσεις για το διαχωρισμό πηγών, οι οποίες συνήθως βασίζονται σε προηγούμενη γνώση σχετικά με τις πηγές για να τις διαχωρίσουν. Το BSS είναι ένα ιδιαίτερα δύσκολο πρόβλημα επειδή απαιτεί τον διαχωρισμό των πηγών αποκλειστικά με βάση τις στατιστικές τους ιδιότητες.

Χρησιμοποιώντας μόνο ένα μικρόφωνο ή κανάλι ήχου, ο διαχωρισμός τυφλής πηγής ενός καναλιού (1C-BSS) είναι ένας κλάδος της επεξεργασίας σήματος ήχου που ασχολείται με τον διαχωρισμό διακριτών πηγών ήχου από ένα μείγμα σημάτων ήχου χωρίς προηγούμενη γνώση των πηγών.

1.2 Πως έχει διαμορφωθεί ο τομέας BSS μέχρι σήμερα

Με την έρευνα των J.L. Lacour και P. Ruiz το 1988 πάνω στις συσσωρεύτριες βλέπουμε ότι τα σήματα που εκπέμπονται από τις πηγές δεν είναι γκαουσιανές. Η δεύτερη τάξη στατιστικής δεν περιέχει την πλήρη περιγραφή των στατιστικών ιδιοτήτων των σημάτων . Στην δημοσίευση αυτή φαίνεται ότι σε μη γκαουσιανές περιπτώσεις είναι πιθανή η ταυτοποίηση των πηγών χρησιμοποιώντας συσσωρεύτριες τρίτης και τέταρτης τάξης [1].

Ο Jean-Francois Cardoso το 1989 [2] για να δείξει ότι τα σήματα των πηγών ταυτοποιούνται αμέσως ως ιδιοδιανύσματα συνδιακύμανσης χρησιμοποίησε υψηλής τάξης ροπές, για την ταυτοποίηση της ”υπογραφής” στον πίνακα δεδομένων χωρίς κάποιο a-priori μοντέλο για διάδοση ή λήψη ηχητικού σήματος δηλαδή χωρίς κατευθυντήριο διάνυσμα παραμετροποίησης , δεδομένου ότι τα εκπεμπόμενα σήματα είναι ανεξάρτητα, με διαφορετικές κατανομές πιθανότητας και καμία γνώση για την τοποθέτηση των πηγών.

Οι Adel Belouchrani , Jean-Francois Cardoso το 1993 έδειξαν μια νέα τεχνική διαχωρισμού πηγών. Εκμεταλλεύτηκαν την χρονική ακολουθία των σημάτων σε αντίθεση με προηγούμενες τεχνικές που βασίζονταν μόνο σε σταθερές στατιστικές δεύτερης τάξης [3] .

Ο Pierre Comon το 1994 προτείνει έναν αποδοτικό αλγόριθμο, που επιτρέπει τον υπολογισμό της ανάλυσης ανεξαρτήτων συνιστωσών από ένα πίνακα δεδομένων εντός πολυωνυμικού χρόνου . Η ιδέα της ανάλυσης ανεξαρτήτων συνιστωσών μπορεί να οριστεί ως μια επέκταση της ανάλυσης κυρίων συνιστωσών, που μπορεί να ανακαλύψει ανεξαρτησία μέχρι 2ης τάξης κατά συνέπεια, ορίζει κατευθύνσεις που είναι ορθογώνιες [4].

Ο Anthony J Bell το 1995 έδειξε ότι οι μη γραμμικότητες στην μέθοδο μεταφοράς είναι ικανές να συλλάβουν μεγαλύτερης τάξης ροπές από την κατανομή των σημάτων και να πραγματοποιήσουν κάτι σαν true redundancy reduction (εξάλειψη πλεονάζουσας πληροφορίας) μεταξύ μονάδων στην έξοδο αναπαράστασης. Αυτό επιτρέπει στο δίκτυο τον διαχωρισμό στατιστικά ανεξαρτήτων τμημάτων στα σήματα, πράγμα που κάνει το δίκτυο ικανό διαχωρίζει έως και 10 ομιλητές [5].

Οι Andrzej Cichocki, Rolf Unbehauen το 1996 παρήγαγαν έναν αλγόριθμο που είναι για σειριακή εκπαίδευση ενός στρώματος εμπρόσθιας τροφοδοσίας νευρωνικού δικτύου. Επιπλέον παρήγαγαν και ένα δεύτερο αναδρομικό νευρωνικό δίκτυο για επιβεβαίωσή [6].

Ο S.Amari το 1996 έδειξε ότι η εξάρτηση μετριέται από την μέση κοινή πληροφορία των εξόδων. Τα σήματα και οι μίξεις είναι άγνωστα ενώ ο αριθμός των πηγών είναι γνωστός. Η Gram-Charlier αντί της Edgeworth επέκτασης χρησιμοποιήθηκε για την αξιολόγηση της κοινής πληροφορίας. Η προσέγγιση natural gradient (φυσικής κλίσης) χρησιμοποιείται για την ελαχιστοποίηση της κοινής πληροφορίας. Επιπλέον προτείνεται μια καινοτόμα συνάρτηση ενεργοποίησης [7].

Ο Aapo Hyvärinen το 1997 έδειξε ότι ένας κανόνας νευρωνικού δικτύου μπορεί να μεταλλαχθεί σε μια επανάληψη σταθερού σημείου (fixedpoint iteration). Αυτό μας προσφέρει ένα πολύ εύκολο αλγόριθμο, που δεν βασίζεται σε κάποια παράμετρο ορισμένη από τον χρήστη. Είναι γρήγορος στην σύγκλιση της βέλτιστης λύσης που επιτρέπουν τα δεδομένα. Οι υπολογισμοί μπορούν είτε να γίνουν τμηματικά είτε με ολικό τρόπο. Η σύγκλιση του αλγορίθμου έχει αποδειχθεί ότι έχει κυβική ταχύτητα. Τέλος γίνονται

κάποιες συγκρίσεις με αλγόριθμους βασισμένους σε gradient (κλίσεις) όπου η ταχύτητα του αλγορίθμου φαίνεται να είναι μεταξύ 10-100 φορές γρηγορότερη [8].

Η έρευνα του Jean-Francois Cardoso το 1997 ορίζει ως βασική αρχή για των διαχωρισμό πηγών την βελτιστοποίηση μιας συνάρτησης που ονομάζετε contrast function (συνάρτηση αντίθεσης). Βασισμένος στην αρχή infomax έφτιαξε μια καινούργια contrast function (συνάρτηση αντίθεσης). Σε αυτήν την έρευνα εξετάζεται η contrast function συσχετισμένη με την καλά εδραιωμένη maximum likelihood principle (αρχή της μέγιστης πιθανοφάνειας) [9].

Ο Dinh Tuan Pham το 1997 παρουσίασε 2 μεθόδους όπου οι πληροφορίες σχετικά με την κατανομή πιθανοτήτων αποκτούνται μέσω μιας λύσης maximum likelihood (αρχή της μέγιστης πιθανοφάνειας). Η πρώτη μέθοδος είναι ειδικά σχεδιασμένη για προσωρινά ανεξάρτητες μη γκαουσιανές πηγές και βασίζεται στη χρήση μη γραμμικών μεθόδων διαχωρισμού. Η δεύτερη μέθοδός είναι ειδικά σχεδιασμένη για αντίστοιχα σήματα με διακριτά φάσματα και είναι βασισμένη στη χρήση των γραμμικών διαχωριστικών φίλτρων [10].

Ο Jean-Francois Cardoso το 1998 [11] έθεσε ως στόχο να εξετάσει τις μεθόδους επίλυσης του προβλήματος που προκύπτει με την χρήση ανάλυση ανεξαρτήτων συνιστωσών εκμεταλλευόμενος μόνο την υπόθεση της αμοιβαίας ανεξαρτησίας μεταξύ των σημάτων. Επιπλέον στο άρθρο [12] ξεκινά από ανάλυση ανεξαρτήτων συνιστωσών για τον τυφλό διαχωρισμό πηγών και δείχνει ότι μπορεί να ταυτοποιήσει πηγές, δεδομένου ότι το μοντέλο ανάλυσης ανεξαρτήτων συνιστωσών, είναι σωστά παραμετροποιημένο με όρους από μονοδιάστατο υποχώρο. Στη συνέχεια η έρευνα κατευθύνεται στο πώς η κανονική ανάλυση ανεξαρτήτων συνιστωσών μπορεί να υιοθετηθεί σε MICA decomposition (Η αποσύνθεση MICA (Μη παραμετρική εξερεύνηση με βάση τη μέγιστη πληροφορία) είναι μια τεχνική για την ανάλυση συνόλων δεδομένων υψηλών διαστάσεων).

Ο Shiro Ikeda's το 1998 εστιάζει στην χρονική δομή των σημάτων. Η ιδέα είναι η εφαρμογή του decorrelation method που προτάθηκε από τους Molgedey και Schuster [13] στον τομέα της χρονοσυχνότητας [14].

Οι Adel Belouchrani, Moeness G. Amin το 1998 παρουσιάζουν μια μέθοδο για τυφλό διαχωρισμό πηγών εκμεταλλευόμενοι την διαφορά μεταξύ των υπογραφών χρονοσυχνότητας. Η μέθοδος βασίζεται σε μέθοδο διαγωνιοποίησης ενός συνδυασμένου σετ από χωρικές κατανομές χρόνου-συχνότητας [15].

Οι Daniel Schobben, Kari Torkkola, Paris Smaragdis το 1999 προτείνουν μια πλατφόρμα από συνθετικό περιβάλλον διαχωρισμού πηγών και πραγματικές καθαρές πηγές παρόλο που οι πραγματικές πηγές έχουν το μειονέκτημα της μη ακριβούς αξιολόγησης σε θέμα ποιότητας [16].

Οι Anisse Taleb, Christian Jutten το 1999 αρχικά προτείνουν θεωρητικά αποτελέσματα που στην πορεία αποδεικνύονται. Αποδεικνύεται οτι δεν είναι πιθανός ο διαχωρισμός των πηγών χωρίς nonlinear distortion (Η μη γραμμική παραμόρφωση είναι ένας όρος που χρησιμοποιείται για να εξηγήσει το φαινόμενο μιας μη γραμμικής σχέσης μεταξύ των σημάτων "εισόδου" και "εξόδου" μιας συσκευής). Ο πρώτος αλγόριθμος βασίζεται σε Gram–Charlier επέκταση. Δυστυχώς τα αποτελέσματα για μη γραμμικά μείγματα είναι μη επιθυμητά. Ένας δεύτερος αλγόριθμος βασιζόμενος σε δυναμική εκτίμηση των λογαριθμικών-παραγώγων πυκνοτήτων έχει πολύ καλύτερα αποτελέσματα [17].

Οι Lucas Parra και Clay Spence εξετάζουν το πρόβλημα εκμεταλλευόμενοι ευθέως την μη στασιμότητα θέσης για ακουστικές πηγές αλλάζοντας διασυσχετίσεις (Η Cross-correlation είναι μια ποσοτικοποίηση που παρακολουθεί τις σχετικές κινήσεις δύο ή περισσότερων συνόλων δεδομένων χρονοσειρών). Πολλές φορές δίνεται ένα ικανοποιητικό σετ από περιορισμούς για άγνωστα κανάλια. Μια βελτίωση με χρήση ελαχίστων τετραγώνων επιτρέπει την εκτίμηση του εμπρόσθιου (forward) μοντέλου. Με τον ίδιο τρόπο βρίσκουν μία μέθοδο φιλτραρίσματος σημάτων χρησιμοποιώντας ένα φίλτρο πεπερασμένης απόκρισης παλμών. Σε ένα φίλτρο προς τα πίσω, η απόκριση παλμού αναστρέφεται και εφαρμόζεται με την αντίστροφη σειρά σε σύγκριση με ένα κανονικό φίλτρο προς τα εμπρός [18].

Ο Sam T. Roweis το 2000 μας αναφέρει πως αλγόριθμοι από ανάλυση ανεξαρτήτων συνιστωσών και οι επεκτάσεις τους ανακτούν τις πηγές με επαναβάθμιση σειράς πολλαπλών παρατηρήσεων. Έτσι δεν μπορεί να λειτουργήσει όταν μόνο μία παρατήρηση από ένα σήμα είναι διαθέσιμη. Παρουσιάζεται μια τεχνική που λέγεται επαναφιλτράριζμα που επανακτά πηγές από μια μη-στατική επαναβάθμιση ("masking") από συχνότητες που αποτελούνται από τμήματα από μια μόνο έγγραφη. Έπειτα παρουσίασε τα αποτελέσματα από απλό παραγοντικό hidden markov μοντέλο το οποίο μαθαίνει από τις εγγραφές από μονούς ομιλητές. Το μοντέλο μπορεί να διαχωρίσει μίξεις χρησιμοποιώντας μόνο μια παρατήρηση με τον υπολογισμό της συνάρτησης και μετά επαναφιλτράρισμα [19].

Οι P. Bofill , M. Zibulevsky το 2001 παρουσίασαν τον στόχο αυτής της έρευνας ως τον διαχωρισμό N πηγών από M γραμμικές μίξεις όταν το υπόβαθρο σύστημα έχει $\mathrm{M} < \mathrm{N}$. Εάν η κατανομή των σημάτων είναι αραιή ο πίνακάς μείξης μπορεί να εκτιμηθεί είτε από εξωτερική βελτιστοποίηση ή λύνοντας ένα χαμηλών διαστάσεων γραμμικού προγραμματισμού πρόβλημά. Ωστόσο όταν τα σήματα δεν τηρούν αυτή την υπόθεση, η αραιότητα μπορεί να αποκτηθεί μεταφέροντας τον διαχωρισμό σε ακόμη αραιότερη περιοχή. Σε αυτή την περίπτωση υπολογίζουμε και τον αριθμό των πηγών και τον πίνακα μείξης από τις κορυφές μιας πιθανής συνάρτησης κατά τον κύκλο του μεγέθους της μονάδας. Έτσι λαμβάνεται η ελάχιστη αναπαράσταση $l1$ νόρμας για κάθε σημείο στα δεδομένα από ένα γραμμικό συνδυασμό των ζευγαριών της βάσης διανυσμάτων που το περιβάλλουν [20].

Οι Dinh-Tuan Pham,Jean-Francois Cardoso το 2001 ανέπτυξαν μια καινοτόμα προσέγγιση βασισμένη στην maximum likelihood (αρχή της μέγιστης πιθανοφάνειας) και ελάχιστης κοινής πληροφορίας. Αυτές οι αρχές δίνουν πλεονέκτημα στους αποτελεσματικούς αλγόριθμους και στις εκτός σύνδεσης περιπτώσεις (δηλαδή χρήση ομαδοποίησης δεδομένων και όχι σειριακά μέσω μιας νέας διαδικασίας από κοινού διαγωνιοποίησης). Επιπλέον είναι αποδοτική και στην περίπτωση σειριακής επεξεργασίας δεδομένων (μέσω μιας παρόμοιας τεχνικής του Νεύτωνα). Στο τέλος αυτό που κάνει του αλγορίθμους να δουλεύουν είναι ότι κάθε αναγνώριση των σημάτων έχει ένα χρονικά ποικιλόμορφο επικάλυμμα (time-varying envelope) [21].

Ο M.D. Plumbley το 2003 ανακάλυψε πως για ανεξάρτητες πηγές με μη μηδενική συνάρτηση πυκνότητας πιθανοτήτων κοντά στο 0 είναι εύλογο να βρεθεί η ορθοκανονική περιστροφή $y = Wz$ των πηγών που ήδη έχουν υποστεί λεύκανση $z = Vx$. Αυτό που ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα της ανακατασκευής των Z από την διορθωμένη έκδοση $y+$ του y. Έπειτα προτείνονται κάποιοι αλγόριθμοι που εκτελούν την σκεπτική αυτή. [22].

Οι Ozgur Yılmaz, Scott Rickard το 2004 πρότειναν ότι άψογη απόμειξη μέσω δυαδικών μασκών χρονοσυχνότητων είναι πιθανή άμα υπάρχουν αναπαραστάσεις χρονοσυχνότητα. Εάν οι αναπαραστά-

σεις των πηγών δεν αλληλοκαλύπτονται τότε η περίπτωση καλείται W-αποσυνδεόμενης ορθογωνικότητας (W-disjoint orthogonality). Μια μέθοδος προσέγγισης W-αποσυνδεόμενης ορθογωνικότητας (W-disjoint orthogonality) παρουσιάζεται την ώρα που ο καθορισμός των μασκών από μια μείξη είναι ανοικτό πρόβλημά. Δείχνεται ότι μπορεί να εκτιμηθούν οι ιδανικές μάσκες στην περίπτωση όπου δυο μείξης χωρίς αντανάκλαση παρουσιάζονται. Υποκινείται από την παράμετρο ανάμειξης της μέγιστης πιθανότητας. Εκτιμητές, ορίζουν ένα δισδιάστατο ιστόγραμμα σταθμισμένης ισχύος αποτελούμενο από την αναλογία των αναπαραστάσεων χρόνο-συχνότητας των μείξεων. Στο ιστόγραμμα έχει δειχθεί ότι οι αναπαραστάσεις έχουν μια κορυφή για κάθε πηγή με κάθε σημείο κορυφής να εξισώνεται με την σχετική ”αραίωση” (attenuation) και τις παραμέτρους καθυστέρησης μείξης.Το ιστόγραμμα χρησιμοποιείται για την δημιουργία των μασκών χρονοσυχνότητας που κομματιάζει την μείξη σε αρχικές πηγές [23].

Ο Hiroshi Sawada το 2004 με αυτήν η έρευνα παρουσιάζει μια ”στιβαρή” και ακριβή μέθοδο για την επίλυση του permutation problem (Είναι ένα πρόβλημα εύρεσης όλων των δυνατών τρόπων διάταξης ενός συνόλου στοιχείων με μία συγκεκριμένη σειρά. Με άλλα λόγια, είναι το πρόβλημα της εύρεσης όλων των πιθανών μεταθέσεων ενός δεδομένου συνόλου στοιχείων). Βασίζεται σε δυο προηγούμενες μεθόδους: την εκτίμηση κατεύθυνσης αφίξεως και την συσχέτιση τοπικής συχνότητας. Με την εύρεση των πλεονεκτημάτων και μειονεκτημάτων των δυο αυτών μεθόδων επιτρέπεται να χρησιμοποιηθούν για την εκμετάλλευση των επιμέρους πλεονεκτημάτων. Επιπλέον προτείνεται μια κλειστή φόρμουλα για την εκτίμηση της κατεύθυνσης των αρχικών σημάτων από τον πίνακα διαχωρισμού χρησιμοποιώντας ανάλυση ανεξαρτήτων συνιστωσών [24].

Οι Mikkel N.Schmidt, Rasmus K.Olsson το 2006 με την χρήση αραιής μη αρνητικής παραγοντοποίησης μήτρας (sparse non-negative matrix factorization) μπορεί να μάθει τις αναπαραστάσεις αραιότητας των δεδομένων χωρίς επίβλεψη. Αυτό εφαρμόζεται στην εκμάθηση του προσωπικού λεξιλογίου για έναν ομιλητή. Και μετά χρησιμοποιείται για διαχωρισμό του ακουστικού κύματος στα κομμάτια του. Φαίνεται ότι μπορεί να υπάρχει υπολογιστική οικονομία εφόσον τεμαχίσουν τα δεδομένα εκπαίδευσης σε επίπεδα ”φωνήματος” (phoneme) [25].

Ο Paris Smaragdis το 2007 παρουσιάζει μια μέθοδο συνελικτικής αποσύνθεσης βάσης (convolutive basis decomposition) και την εφαρμογή της σε ταυτόχρονους ομιλητές από έγγραφες μικροφώνων .Το μοντέλο που προτείνεται είναι μια συνελικτική έκδοση του αλγορίθμου παραγοντοποίησης μη αρνητικού πίνακα (non-negative matrix factorization algorithm) [26].

Ο Taesu Kim το 2007 ανακάλυψε έναν νέο αλγόριθμο που υποθέτει ότι οι εξαρτήσεις υπάρχουν μεταξύ ”καλαθιών” συχνοτήτων αντί για τον ορισμό της ανεξαρτησίας για καθένα καλάθι. Με αυτόν τον τρόπο μπορεί να αποφύγει το πρόβλημα μετάθεσης συχνότητας (frequency permutation problem). Για την εξαγωγή αλγορίθμου μάθησης, ορίζεται μια συνάρτηση κόστους που είναι μια επέκταση της αμοιβαίας πληροφορίας μεταξύ τυχαίων μεταβλητών με μεγάλη διακύμανση [27].

Ο Po-Sen Huang το 2007 σκέφτηκε έναν συνδυασμό από βαθιά νευρωνικά δίκτυα και αναδρομικά νευρωνικά δίκτυα σε συνδυασμό με ένα επιπλέον masking layer (στρώμα μάσκας), το οποίο εφαρμόζει τον περιορισμό της ανακατασκευής. Επιπλέον, εξερευνάται ένα διακριτικό κριτήριο εκπαίδευσης για τον εμπλουτισμό της ακρίβειας διαχωρισμού στα νευρωνικά δίκτυα. Με την εισαγωγή μιας πηγής πριν την μοντελοποίηση των έμφυτων εξαρτήσεων συχνότητας, λαμβάνουμε μια απλή μορφή μιας συνάρτησης πολλαπλών μεταβλητών [28].

Ο Felix Weninger το 2014 χρησιμοποίησε ένα γενικό διακριτικό κριτήριο εκπαίδευσης που αντιστοιχίζει την ιδανική ανακατασκευή από μάσκες χρονοσυχνοτήτων. Το κριτήριο αυτό εισάγεται στον τομέα του διαχωρισμού λόγου για να μειώσει τον χώρο χαρακτηριστικών (Meldomain) [29].

Ο Yusuf Isik το 2016 ξεκινά με την βελτίωσή της απόδοσης του βασικού συστήματος με την υιοθέτηση καλύτερης ομαλοποίηση (regularization). Επίσης λαμβάνεται μεγαλύτερο χρονικό πλαίσιο καθώς και βαθύτερη αρχιτεκτονική. Επιπλέον επεκτείνεται το μοντέλο για την εισαγωγή ενός στρώματος εμπλουτισμού για την βελτίωσή των εκτιμήσεων των σημάτων. Ο τρόπος που γίνεται η εκπαίδευση είναι ότι όλες οι απαραίτητες ενέργειες γίνονται από το δίκτυο (end-to-end). Ο νέος στόχος προσέγγισης σήματος, σε συνδυασμό με end-to-end τρόπο, παράγει ανεπανάληπτα αποτελέσματα, μειώνοντάς το ποσοστό σφάλματος ανά λέξη από 89.1%σε 30.8% [30].

Ο Morten Kolbæk το 2017 εφαρμόζει μια εκπαίδευση αναλλοίωτης μετάθεσης σε επίπεδο εκφοράς (utterance-level Permutation Invariant Training) εφαρμόσιμη με τρόπο end-to-end (δηλαδή το εγχείρημα δεν διασπάτε σε περισσότερα κομμάτια) για πολλαπλό διαχωρισμό ανεξαρτήτου ομιλητή. Συγκεκριμένα, επεκτείνει την πρόσφατα προτεινομένη Εκπαίδευση αναλλοίωτης μετάθεσης (Permutation Invariant Training [31]), εξαλείφοντας έτσι το πρόβλημα του permutation problem. Με αυτό τον τρόπο εξαναγκάζονται τα διαχωρισμένα καρέ που ανήκουν στον ίδιο ομιλητή να είναι ευθυγραμμισμένα με την ίδια ροή εξόδου. Στη πραγματικότητα αυτό επιτρέπει στα εκπαιδευμένα από uPIT μοντέλα τον διαχωρισμό πολλών μείξεων ομιλίας χωρίς κάποια προηγουμένη γνώση για την διάρκεια, αριθμό ομιλητών, ταυτότητα η φύλο του ομιλητή [32].

Οι Yi Luo, Nima Mesgarani προτείνουν το Time-domain Audio Separation Network (TasNet) για την απαλοιφή περιορισμών όπως η αποσύνδεση φάσης/μεγέθους και μεγάλο χρονικό παράθυρο που χρειάζεται για να πετύχει ικανοποιητική ανάλυση συχνότητας. Στο μοντέλο αυτό γίνεται απευθείας μοντελοποίηση του σήματος σε χρονικό τομέα χρησιμοποιώντας μια πλατφόρμα κωδικοποιητή-αποκωδικοποιητή και παριστάνει των διαχωρισμό πηγών ως μη-αρνητικά κωδικοποιημένες εξόδους. Αυτή η μέθοδος αφαιρεί το βήμα αποσύνθεσης συχνότητας και μειώνει το πρόβλημα διαχωρισμού με εκτίμηση των μασκών των πηγών σε εξόδους κωδικοποιητή που συνθυιθενται στην πορεία από τον αποκωδικοποιητή [33].

Οι Zhuo Chen, Yi Luo, Nima Mesgarani το 2017 προτείνουν μια νέα μέθοδος για διαχωρισμό πηγών με την χρήση ενός καναλιού. Δημιουργούνται σημεία προσέλκυσης σε χώρο ενσωμάτωσης υψηλών διαστάσεων για τα ακουστικά σήματα τα οποία συσσωρεύονται σε καλάθια χρόνο-συχνοτήτων που αντιπροσωπεύουν μια πηγή. Τα σημεία προσέλκυσης δημιουργούνται βρίσκοντας τα σημεία που απέχουν λιγότερο από το κέντρο των πηγών στον χώρο ενσωμάτωσης. Τα οποία σημεία ακολούθως χρησιμοποιούνται για καθορισμό της ομοιότητας κάθε καλαθιού στην μειξη με άλλες πηγές. Το δίκτυο εκπαιδεύεται για ελαχιστοποίηση του σφάλματος επανακατασκευής για κάθε πηγή με την βελτίωση των ενσωματώσεων. Η προτεινόμενη μέθοδος είναι διαφορετική από προηγούμενες μεθόδους καθώς εφαρμόζει εκπαιδευση end-to-end (δηλαδή το εγχείρημα δεν διασπάται σε περισσότερα κομμάτια) και δεν περιορίζεται από τον αριθμό των πηγών στο μείγμα. Δυο στρατηγικές ερευνώνται περαιτέρω K-means και fixed attractor points όπου η δεύτερή δεν χρειάζεται επιπλέον επεξεργασία και μπορεί να χρησιμοποιηθεί σε πραγματικό χρόνο [34] .

Ο Yi Luo το 2018 έθεσε γνωστό το πρόβλημα για μεθόδους διαχωρισμού ομιλίας με μονό κανάλι εισαγωγής. Ανεξάρτητα από την ταυτότητα των ομιλητών, η απόδοσή, ο χρόνος και το υπολογιστικό

κόστος από αυτές τις μεθόδους παραμένουν μη ικανοποιητικά. Στην μέθοδο που προτείνεται ο διαχωρισμός των ομιλητών γίνεται μέσω εφαρμογής ενός σετ από συναρτήσεις διαβάθμισης μασκών (weighting functions) στην έξοδο του κωδικοποιητή. Οι τροποποιημένες αναπαραστάσεις κωδικοποιητή αντιστρέφονται πίσω σε μορφή κύματος χρησιμοποιώντας ένα γραμμικό αποκωδικοποιητή. Οι μάσκες βρίσκονται χρησιμοποιώντας χρονικός συνελικτικό δίκτυο (temporal convolutional network) που αποτελείται από ένα μονοδιάστατο διευρυμένο συνελικτικό τμήμα το οποίο επιτρέπει στο δίκτυο μακροχρόνιες ανεξαρτησίες από το σήμα ομιλίας ενώ παράλληλα διατηρεί μικρό μέγεθος μοντέλου [35].

Ο Yi Luo το 2020 έκανε γνωστά τα προβλήματα με χρονικές μεθόδους διαχωρισμού που συχνά λαμβάνουν σειρές αποτελούμενες από μεγάλο αριθμό βημάτων. Συνήθως αυτό εισάγει προβλήματα για μοντελοποίηση πολύ μεγάλων σειρών. Τυπικά αναδρομικά νευρωνικά δίκτυα δεν είναι πολύ αποτελεσματικά για μοντελοποίηση τόσο μεγάλων σειρών εξαιτίας προβλημάτων στην βελτιστοποίηση. Παράλληλα τα Μονό-διάστατα συνελικτικά δίκτυα δεν μπορούν να κάνουν μοντελοποίηση ακολουθίας σε επίπεδο ομιλίας όταν ο αντίστοιχος τομέας λήψης είναι μικρότερος από την ακολουθία. Σε αυτή την έρευνα προτείνεται το dual-path recurrent neural network (DPRNN), μια απλή άλλα ικανή μοντελοποίηση για οργάνωση αναδρομικών νευρωνικών δικτύων σε μια βαθιά δομή για την μοντελοποίηση πολύ μεγάλων σειρών [36].

Ο Jingjing Chen το 2020 Σε αυτή την έρευνά προτείνεται ένα dual-path transformer network (DPTNet) για end-to-end διαχωρισμό ομιλίας. Η καινοτομία του αλγορίθμου είναι η εισαγωγή άμεσης επίγνωσης του πλαισίου στην μοντελοποίηση της ακολουθίας. Εισάγοντας ένα βελτιωμένο transformer τα στοιχεία σε ακολουθίες ομιλίας μπορούν να αλληλοεπιδρούν άμεσα. Ο βελτιωμένος transformer μαθαίνει την σειρά της πληροφορίας χωρίς κάποια σειριακή κωδικοποίηση με την χρήση αναδρομικών νευρωνικών δικτύων στον αρχικό transformer. Επιπλέον η δομή κάνει καλύτερη την μοντελοποίηση μεγάλων σειρών [37].

Ο Eliya Nachmani το 2020 παρουσίασε μια μέθοδο για διαχωρισμό μείξης, όπου πολλαπλές ομιλίες συμβαίνουν ταυτόχρονα. Η νέα αυτή μέθοδος χρησιμοποιεί νευρωνικά δίκτυα πυλών τα οποία είναι εκπαιδευμένα για διαχωρισμό ομιλιών σε πολλαπλά βήματα, κρατώντας τον ομιλητή σε κάθε κανάλι εξόδου σταθερό. ένα απλό μοντέλο εκπαιδεύεται για κάθε πιθανό αριθμό ομιλητών στο δωμάτιο και το μοντέλο με τους περισσοτέρους ομιλητές διαλέγεται [38].

Ο Ziqiang Shi το 2020 προτείνει αρκετές βελτιώσεις στο διπλής κατεύθυνσης BiLSTM δίκτυο για end-to-end προσέγγιση. Πρώτα ένα διπλής κατεύθυνσης δίκτυο τοπικής επεξεργασίας BiLSTM και ένα δίκτυο μακροχρόνιας επεξεργασίας BiLSTM ορίζονται για την μείωση της αστάθειας της απόδοσης σε διαφορετικά τμήματα. Έπειτα προτείνεται η εισαγωγή της μακροχρόνιας επίγνωσης πλαισίου για καλύτερη αντίληψη του γενικότερου νοήματος της πληροφορίας. Τελικά ένα σπειροειδής LSTM διπλής κατεύθυνσης πολλαπλών σταδίων ορίζεται για την σταδιακή βελτίωση της διαχωριστικής ικανότητας. Ως στόχος εκπαίδευσης ορίζεται η απευθείας βελτίωση του επίπεδου εκφοράς αναλογίας σήματος προς παραμόρφωση αμετάβλητης κλίμακας(utterance level scale-invariant signal-to-distortion ratio) (SI-SDR) με τρόπο αμετάβλητης εκπαίδευσης(ermutation invariant training) (PIT) [39].

Ο Manuel Pariente το 2020 έδειξε ότι ο τομέας ενός καναλιού έχει σημειώσει πρόσφατα σημαντική πρόοδο.Χάρη σε γνωστές τράπεζες φίλτρων όπως αυτές που χρησιμοποιούνται στο ConvTasNet εμπλουτίζει τις μαθημένες και παραμετροποιημένες τράπεζες φίλτρων με πραγματική αξία σε αναλυτικές

τράπεζες φίλτρων σύνθετης αξίας.

Οι Scott Wisdom, Efthymios Tzinis το 2020 έδειξαν ότι η εξάρτηση από συνθετικά δεδομένα για εκπαίδευση είναι προβληματική. Εφόσον καλά αποτελέσματα εξαρτιούνται από την καλή ταύτιση του πραγματικού κόσμου με την προσομοίωση (Ειδικότερα για ακουστικές συνθήκες και κατανομή πηγών οι ακουστικές ιδιότητες μπορούν να αποτελέσουν δύσκολο έργο προς προσομοίωση και η κατανομή των τύπων των ήχων δύσκολη στην επανάληψη). Προτείνεται μια μέθοδος χωρίς επίβλεψη mixture invariant training (MixIT) που χρειάζεται μόνο ακουστικά μείγματα ενός καναλιού. [40].

Οι Thilo von Neumann, Keisuke Kinoshita το 2020 δείχνουν πώς να συνδυαστεί ένα διαχωριστικό τμήμα βασισμένο σε Convolutional Time domain Audio Separation Network (Conv-TasNet) με end to end αναγνώριση ομιλίας. Επιπλέον παρουσιάζουν το πως να εκπαιδευτεί ένα τέτοιο μοντέλο ολικά μοιράζοντας το σε πολλαπλές GPUs (κάρτες γραφικών) [41].

Οι Wangyou Zhang , Xuankai Chang σχεδίασαν πολλαπλές μεθόδους για τον εμπλουτισμό της ανάλυσης που περιλαμβάνουν: παράλληλο μηχανισμό προσοχής βασιζόμενο στους ομιλητές, προγραμματισμένη δειγματοληψία, εκπαίδευση με διαβάθμιση δυσκολίας και απόσταξη γνώσης (knowledge distillation). Συγκεκριμένα ο παράλληλος μηχανισμός προσοχής επεκτείνει τον βασικό σχεδιασμό για κοινόχρηστη προσοχή σε πολλά τμήματα προσοχής για κάθε ομιλητή, πράγμα που μπορεί να βελτιώσει την εύρεση και απομόνωση κάθε ομιλητή. Τότε η προγραμματισμένη δειγματοληψία και εκπαίδευση με διαβάθμιση δυσκολίας ακολουθούνται για καλύτερη βελτίωση του μοντέλου. Τέλος η απόσταξη γνώσης μεταφέρει την γνώση από το αρχικό μονοφωνικό μοντέλο στο πρόσφατο πολυφωνικό μοντέλο [42].

Οι Yuzhou Liu , Masood Delfarah 2020 ερευνούν τον διαχωρισμό της βάσης της ομιλίας αλλά και διαχωρισμό ομιλίας από θόρυβο .Η μοντελοποίηση αυτή είναι ανεξάρτητη από την ταυτότητα του ομιλητή. Επεκτείνεται το πρόσφατα προτεινόμενο deep CASA για την διαχείριση μείξεων με θόρυβο. Για την εγκαθίδρυση του εμπλουτισμού του λόγου, ένα τμήμα εξάλειψης θορύβου προστίθεται στο deep CASA έως ένα πρωταρχικό στρώμα επεξεργασίας [43].

Ο Cem Subakan το 2021 έδειξε ότι τα αναδρομίκα νευρωνικά δίκτυα (RNNs) ήταν για καιρό η κυρίαρχη αρχιτεκτονική για εκμάθηση ακολουθίας προς ακολουθία .Ωστόσο τα RNNs είναι σειριακά μοντέλα που δεν επιτρέπουν παραλληλισμό των υπολογισμών τους. Transformers έχουν αναπτυχθεί ως εναλλακτική λύση εναλλάσσοντας επαναλαμβανόμενες πράξεις με μηχανισμό προσοχής Σε αυτή την έρευνα προτείνεται το SepFormer ένα καινούργιο δίκτυο χωρίς RNN βασισμένο σε Transformers για διαχωρισμό ομιλίας. Το SepFormer μαθαίνει βραχυπρόθεσμες και μακροχρόνιες εξαρτήσεις με μια μέθοδο πολλαπλής κλιμάκωσης που χρησιμοποιεί transformers [44].

Ο Shaked Dovrat το 2021 παρουσιάσε μια μέθοδο που χρησιμοποιεί των Hungarian algorithm [45] για εκπαίδευσή σε (C^3) χρόνο όπου C είναι οι ομιλητές. Επίσης παρουσιάσε μια αρχιτεκτονική που μπορεί να διαχειριστεί μεγαλύτερο αριθμό ομιλητών $C <= 20$ και βελτιώνει τα προηγούμενα αποτελέσματα κατά ένα μεγάλο βαθμό [46].

Οι Max W. Y. Lam, Jun Wang σχεδίασαν το globally attentive locally recurrent (GALR) δίκτυο μια αρχιτεκτονική υψηλής ποιότητας χαμηλού κόστους . Παρόμοιά με το διπλής κατεύθυνσης αναδρομικό νευρωνικό δίκτυο (DPRNN) πρώτα χωρίζουν την ακολουθία χαρακτηριστικών σε δισδιάστατα κομμά-

τια, και έπειτα επεξεργάζονται την ακολουθία μαζί με τις διαστάσεις των τοπικών και μακροχρόνιων κομματιών. Η κύρια πρωτοτυπία βρίσκεται στο ότι τα χαρακτηριστικά επεξεργάζονται επαναληπτικά μαζί με τις διαστάσεις στα μακροχρόνια κομμάτια. Σε αυτο το σημείο ένας μηχανισμός αυτό-προσοχής (self-attention mechanism) εγκαθιδρύεται πάνω στην ακολουθία μαζί με τις διαστάσεις στα τοπικά κομμάτια, που συγκεντρώνει πληροφορίες που εμπεριέχουν γνώση συμφραζομένων και επίσης επιτρέπει την παραλληλοποίηση [47].

ΟΙ Max W. Y. Lam,Jun Wang το 2021 παρουσίασαν ένα μοντέλο με κύρια πρωτοτυπία τα πολυκοκκοποίημενα χαρακτηριστικά, που είναι αναγκαία για εμπλουτισμό της μοντελοποίησης της γνώσης των συμφραζομένων . Ένα δίκτυο με μηχανισμό αυτό-προσοχής και σχήμα κλεψύδρας παρουσιάζεται με όνομά Sandglasset που έχει την καλύτερη αποδοτικότητα με πολύ μικρότερο μέγεθος μοντέλου και υπολογιστικό κόστος. Η σμίκρυνσή των χαρακτηριστικών γίνεται σταδιακά πιο ισχυρή μέχρι να φτάσει τα μισά μπλοκ δικτύου και έπειτα η μεγέθυνση ακολουθεί αντίστοιχη φορά προς την έξοδο μέχρι την σύνθεση των διαχωρισμένων σημάτων. Επίσης βρέθηκε ότι η διαρροή πληροφορίας μεταξύ χαρακτηριστικών με ίδια επίπεδα σμίκρυνσης μέσω υπολειπόμενες συνδέσεις που είναι αναγκαίες για την διατήρηση της πληροφορίας αφού περάσει από το στρώμα σμίκρυνσης [48].

Οι Chenda Li, Jing Shi το 2021 παρουσιάσαν το ESPnet-SE ένα εργαλείο που σχεδιάστηκε για γρήγορη ανάπτυξη συστημάτων εμπλουτισμού και διαχωρισμού ομιλίας. Το ESPnet-SE είναι ένα νέο εργαλείο που εμπεριέχει πολλά μοντέλα πλούσια σε αυτόματη αναγνώριση ομιλίας. Προσφέρει πόρους και συστήματα για την υποστήριξη και επικύρωση των μοντέλων. Μπορεί να επεξεργαστεί δεδομένα από ένα ή και πολλά κανάλια, με πολλές λειτουργείς που περιλαμβάνουν εξάλειψη της αντήχησης,του θορύβου, εξαγωγή χαρακτηριστικών , εκπαίδευση, επικύρωση καθώς και ένα μεγάλο πλήθος από μετρικές [49].

TIMELINE 1: *History*

1988 - **J.L. Lacour's and P. Ruiz's** Sources lndentification : A Solution Based on the Cumulants

1989 - **Jean-Francois Cardoso's** Source separation using higher order momemnts

1993 - **Adel Belouchrani's** A Blind Source Separation Technique Using Second-Order Statistics

1994 - **Pierre Comon's** Independent component analysis, A new concept?

1995 - **J Bell's Terrence J Sejnow's** An information-maximisation approach to blind separation and blind deconvolution

1996 - **Andrzej Cichocki's and Rolf Unbehauen's** Robust Neural Networks with On-Line Learning for Blind Identification and Blind Separation of Sources

1996 - **s. Amari*'s** A New Learning Algorithm for Blind Signal Separation

1997 - **Aapo Hyvärinen's** A fast fixed-point algorithm for Independent Component Analysis

1998 - **Jean-Fran'sÿCois Cardoso's** Blind signal separation: statistical principles

1998 - **Jean-Fran'sÿCois Cardoso's** Multidimensional independent component analysis

1998 - **Shiro Ikeda's** An approach to blind source separation of speech signals

1998 - **Adel Belouchrani's** Blind source separation based on time-frequency signal representations

1999 - **Daniel Schobben's** Evaluation of blind signal separation methods

1999 - **Anisse Taleb's** Source separation in post-nonlinear mixtures

2000 - **Lucas Parra's** Convolutive Blind Separation of Non-Stationary Sources

2000 - **Sam T. Roweis's** One Microphone Source Separation

2001 - **P. Bofill's and M. Zibulevsky's.** Underdetermined blind source separation using sparse representations

TIMELINE 2: *History*

2003 - **Dinh-Tuan Pham's,ean-Francois Cardoso's** Blind separation of instantaneous mixtures of nonstationary source

2003 - **M.D. Plumbley's** Algorithms for nonnegative independent component analysis

2004 - **Ozgur Ylmaz's** Blind separation of speech mixtures via time-frequency masking

2004 - **Hiroshi Sawad's** A robust and precise method for solving the permutation problem of frequency-domain blind source separation

2006 - **Mikkel N. Schmidt's** Single-channel speech separation using sparse Non-Negative Matrix Factorization

2007 - **Paris Smaragdis's** Convolutive speech bases and their application to supervised speech separation

2007 - **Taesu Kim's** Blind source separation exploiting higher-order frequency dependencies

2007 - **Po-Sen Huan's** Deep learning for monaural speech separation

2014 - **Felix Weninger's** Discriminatively trained recurrent neural networks for single-channel speech separation

2016 - **Yusuf Isik's**Single-channel multi-speaker separation using deep clustering

2017 - **Morten Kolbæk's**Multi-talker speech separation with Utterance-level Permutation Invariant Training of deep recurrent neural networks

2017 - **Yi Luo's**Tasnet time-domain audio separation network for real-time,single-channel speech separation

2017 - **Zhuo Chen's** Deep attractor network for single-microphone speaker separation

2018 - **Yi Luo's** Conv-tasNet: surpassing ideal time-frequency magnitude masking for speech separation

2020 - **Yi Luo's** Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation

2020 - **Jingjing Chen's** Dual-path transformer network: direct context-aware modeling for end-to- end monaural speech separation

TIMELINE 3: *History*

2020 – **Eliya Nachmani's** Voice separation with an unknown number of multiple speakers

2020 – **Ziqiang Shi's** LaFurca: iterative refined speech separation based on context-aware dual-path parallel bi-lstm

2020 – **Manuel Pariente's** Filterbank design for end-to-end speech separation

2020 – **Efthymios Tzinis's** Unsupervised sound separation using mixture invariant training

2020 – **Thilo von Neumann's** End-to-end training of time domain audio separation and recognition

2020 – **Wangyou Zhang's** End-to-end training of time domain audio separation and recognition

2020 – **Yuzhou Liu's** Deep casa for talker-independent monaural speech separation

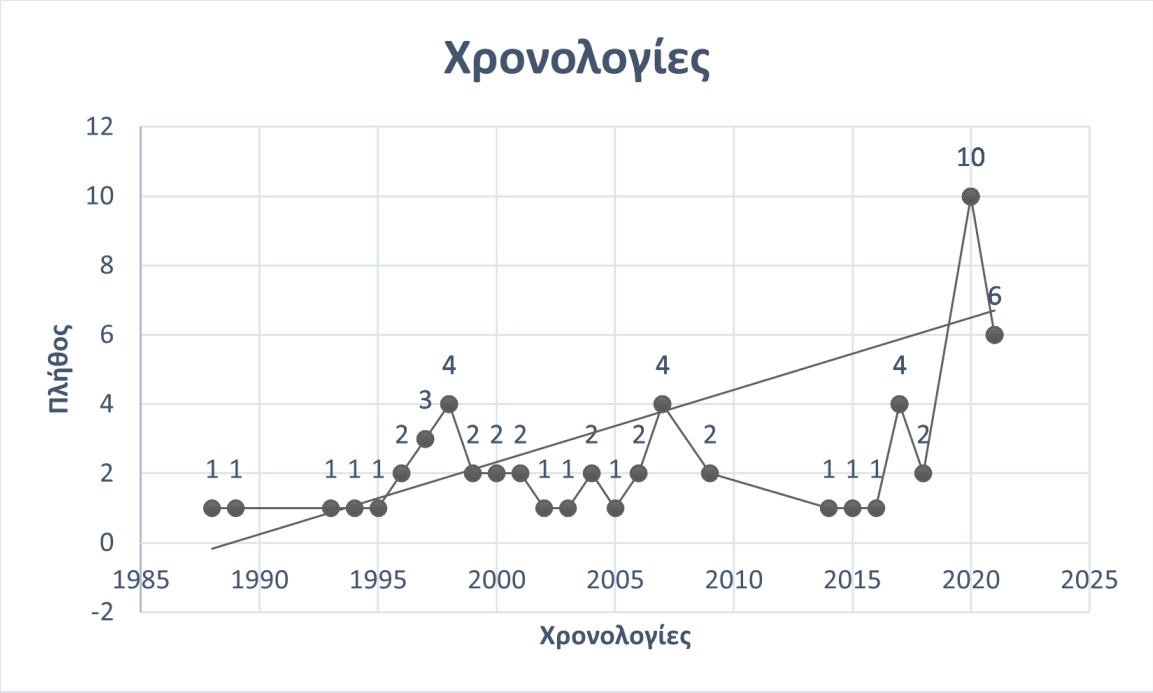
2021 – **Cem Subaka's** Attention is all you need in speech separation

2021 – **Shaked Dovrat's** Many-speakers single channel speech separation with optimal permutation training

2021 – **Max W. Y. Lam's** Effective low-cost time-domain audio separation using globally attentive locally recurrent networks

2021 – **Max W. Y. Lam's** Sandglasset a light multi-granularity self-attentive network for time-domain speech separation

2021 – **Chenda Li's** ESPnet-SE end-to-end speech enhancement and separation toolkit designed for asr integration



Σχήμα 1.1: Η τάση των δημοσιεύσεων

1.2.1 Δομή πτυχιακής

Σε αυτή την πτυχιακή η δομή τον κεφαλαίων θα είναι ως εξής:

- Κεφάλαιο 1 Εισαγωγή στο πρόβλημα με θεωρητικά θεμέλια και εξέλιξη του τομέα στην ιστορία.
- Κεφάλαιο 2 Εισαγωγή στις μεθοδολογίες ροπές, συσσωρεύτριες, παραγοντοποίηση πινάκων, ανάλυση ανεξάρτητων και κυρίων συνιστωσών.
- Κεφάλαιο 3 Ανάλυση θεωρίας βαθιών αναδρομικών νευρωνικών δικτύων bi-lstm και transformers.
- Κεφάλαιο 4 Ανάλυση αλγορίθμων.
- Κεφάλαιο 5 Σετ δεδομένων, μετρικές, αποδόσεις , συγκρίσεις και συμπεράσματα.

Κεφάλαιο 2ο: Θεωρητικό υπόβαθρο

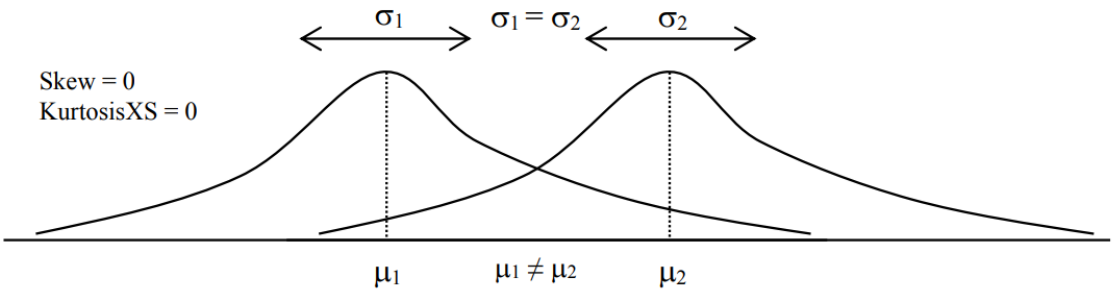
2.1 Εισαγωγή στις ροπές και συσσωρεύτριες

2.1.1 Ροπές

Ορισμός ροπών: *Οι περισσότερες κατανομές μπορούν να αναλυθούν με 4 ροπές. Η πρώτη ροπή περιγράφει την τοποθεσία ή το κέντρο (προσδοκώμενες τιμές) της κατανομής. Η δεύτερή ροπή περιγράφει το πλάτος η την διασπορά στο χώρο. Η τρίτη ροπή είναι η στρέβλωση της κατανομής. Ενώ η τέταρτη ροπή είναι η κυρτότητα της κατανομής. Και οι τέσσερις ροπές θα πρέπει να υπολογιστούν για την παραγωγή μιας περιεκτικής εικόνας της κατανομής [50].*

2.1.1.1 Καθορίζοντάς το κέντρο της κατανομής : Πρώτη ροπή

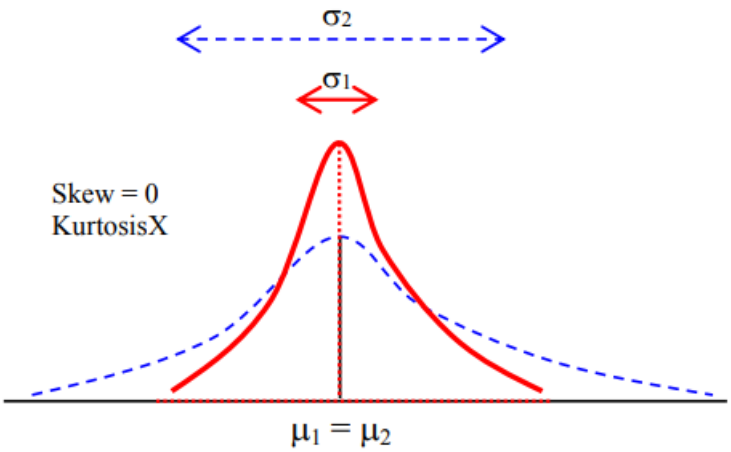
Η πρώτη ροπή μιας κατανομής είναι η προσδοκώμενη τιμή (Expected value). Μετρά την τοποθεσία της κατανομής των δεδομένων και τις πιθανές τιμές κατά μέσο όρο. Οι βασικοί όροι στατιστικής για την πρώτη ροπή είναι ο μέσος όρος, το κέντρο της κατανομής και η πολυπληθέστερη τιμή. Το γράφημα 2.1 δείχνει την πρώτη ροπή που σε αυτή την περίπτωση μετριέται από τον μέσο όρο [50].



Σχήμα 2.1: Το κέντρο της κατανομής [50]

2.1.1.2 Μετρώντας την διασπορά της κατανομής: Δεύτερη ροπή

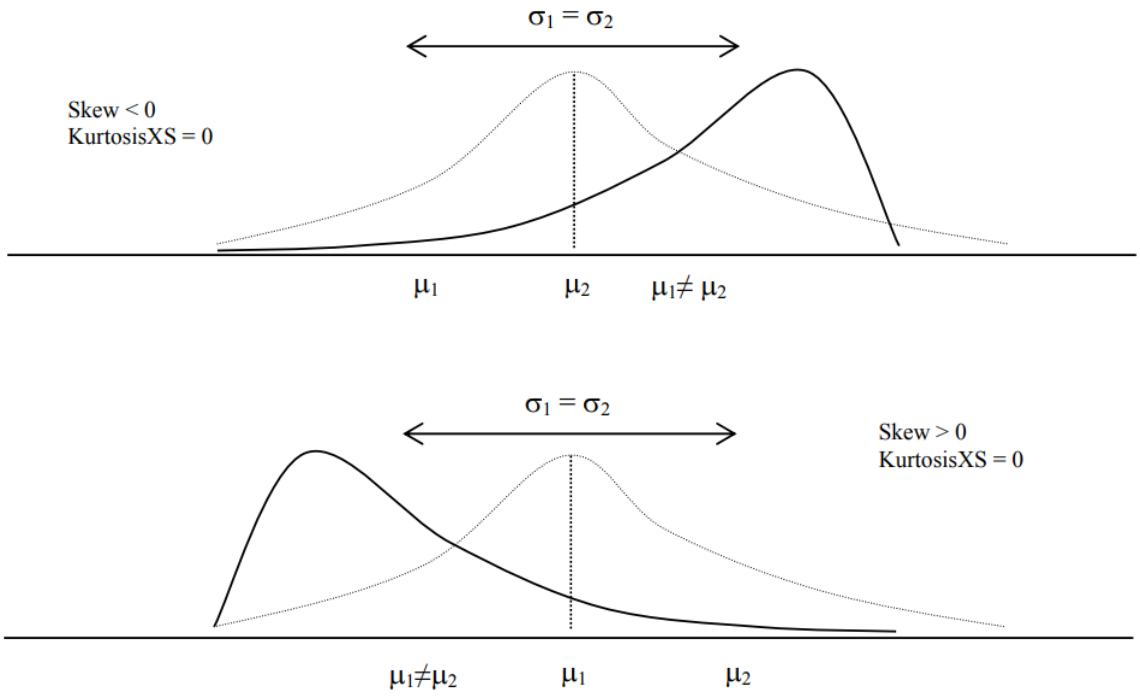
Η δεύτερή ροπή μετρά την διασπορά της κατανομής. Η διασπορά η το πλάτος μιας κατανομής αναφέρει την ποικιλία των τιμών μιας μεταβλητής. Δηλαδή την πιθανότητα η μεταβλητή να βρεθεί σε διαφορετικά σημεία στην κατανομή [50]. Παρατηρείστε ότι στο γράφημά 2.2 και οι δυο κατανομές έχουν ίδιες πρώτες ροπές άλλα υπάρχουν εμφανείς διάφορες. Η διάφορα στο πλάτος μπορεί να μετρηθεί. Μέσω των στατιστικών όρων όπως εύρος (range), απλή θεμελιακή εκτροπή (standard deviation), διακύμανση (varience), συντελεστής διακύμανσης (coefficient of variation), αστάθεια (volatility), διατεταρτημοριακό εύρος (interquartile range), εκατοστημορίων (percentiles) [50].



Σχήμα 2.2: Διασπορά της κατανομής [50]

2.1.1.3 Μετρώντας την στρέβλωση της κατανομής: Τρίτη ροπή

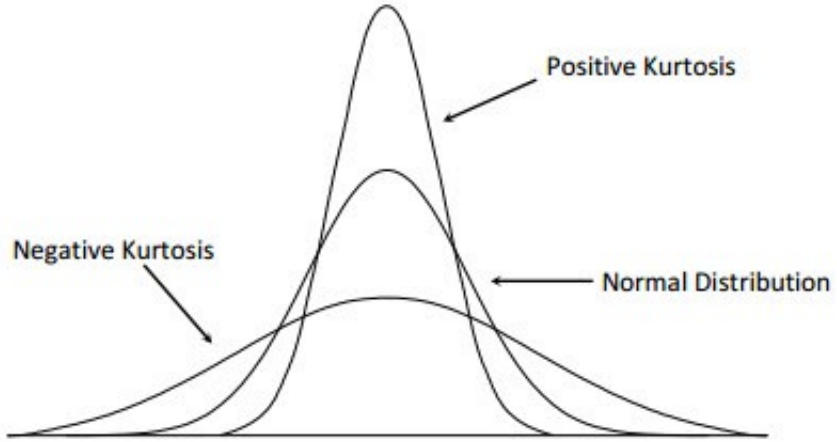
Η τρίτη ροπή μετρά την στρέβλωση της κατανομής. Δηλαδή κατά πόσο μια κατανομή είναι «τραβηγμένη» από την μια η την άλλη κατεύθυνση. Το πρώτο σχήμα δείχνει αρνητική στρέβλωση, ενώ το δεύτερο έχει θετική στρέβλωση. Ο μέσος όρος είναι πάντοτε πιο κοντά στην ουρά της κατανομής ενώ το κέντρο κατανομής παραμένει σταθερό [50].



Σχήμα 2.3: Αρνητική και θετική στρέβλωση [50]

2.1.1.4 Μετρώντας την κυρτότητα: Τέταρτη ροπή

Η τέταρτη ροπή, ή αλλιώς κυρτότητα, μετρά την καμπυλότητα της κατανομής. Το σχήμα 2.4 δείχνει την καμπυλότητα της κατανομής ενώ η διακεκομμένη γραμμή είναι κανονική κατανομή με κυρτότητα 3.0 [50].



Σχήμα 2.4: Καμπυλότητα της κατανομής [50]

2.1.2 Συσσωρεύτριες

Ορισμός συσσωρευτριών: *Οπώς βλέπουμε και στο [51] Εάν* $(Y_1, ..., Y_r)$ *με το* r *να είναι η τιμή της τυχαίας μεταβλητής, οι συσσωρεύτριες* $C[(Y_1, ..., Y_r)]$ *της* r^{th} *τάξεως υπολογίζονται ως*

$$C[(Y_1, ..., Y_r)] = (-1)p^{-1}(p-1)!(E\left\{\prod_{j\in v_1} y_j\right\})...(E\left\{\prod_{j\in v_p} y_j\right\}). \tag{1}$$

Οι συσσωρεύτριες προσφέρουν τρόπους ορισμού παραμέτρων ενδιαφέροντος με χρήσιμες μετρικές κοινής στατιστικής εξάρτησης (joint statistical dependence). Οι συσσωρεύτριες (Cumulants) επίσης ονομάζονται και semi-invariants. [1, 51].

2.1.3 Κυρία Χαρακτηριστικά των συσσωρευτριών

$C[Y_1...Y_r] = 0$ Εάν κάποια ομάδα από τα $(Y's)$ είναι ανεξάρτητη από τις υπόλοιπες $(Y's)$. Και πιο Συγκεκριμένα εάν δυο από τις πήγες S_1, S_2 είναι ανεξάρτητες τότε:

$$C[\underbrace{S_1, ..., S_1}_{p}, \underbrace{S_2, ..., S_2}_{n}] = C[S_1^p S_2^n] = 0, \forall (p, n) \in N^2 \tag{2}$$

Εάν τα S_1, S_2 είναι δυο μηδενικές μέσες γκαουσιανές διαδικασίες τότε ισχύει ότι

$$\begin{aligned} C[S_1^p] &= 0 \quad \mu\varepsilon \quad p > 2 \\ C[S_2^p] &= 0 \quad \mu\varepsilon \quad p > 2 \end{aligned} \tag{3}$$

Έτσι ώστε όλες οι στατιστικές ιδιότητες ορίζονται από της πρώτης και δεύτερης τάξης συσσωρεύτριες. Επιπλέον σε αυτή την περίπτωση η μη συσχέτιση συνεπάγεται ανεξαρτησία. Ωστόσο ακόμη και εάν οι

$$C[S_1^p] = 0(uncorrelation),\ but\ not\ necessarily \qquad (4)$$
$$C[S_1^p, S_2^n] = 0 \quad \forall n > 2\ (independce)$$

χρησιμοποιούμε τέταρτης τάξης συσσωρεύτριες δηλαδή την κύρτωση για διαχωρισμό τον πηγών $C[(S_1, S_2)] = 0$ άλλα $C[S_1^p, S_2^n] \neq 0$ για $n + p = 4$ και ανεξάρτητες πήγες $C[S_1^p, S_2^n] = 0$. Ο διαχωρισμός των πηγών θα γίνει με βάση τις πληροφορίες από την τιμή της τέταρτης τάξης συσσωρευτριών [1, 51].

2.2 Βραχυπρόθεσμος μετασχηματισμός Fourier (Short-term Fourier transform)

Η short-term Fourier transform (STFT) επιτρέπει την ανάλυση σημάτων σε επίπεδο χρονοσυχνότητας. Χρησιμοποιείται για την παραγωγή αναπαραστάσεων που εμπεριέχουν και τον τοπικό χρόνο άλλα και το πλαίσιο συχνότητας. Παρομοίως με την Fourier transform η STFT βασίζεται σε βάσεις πάγιων συναρτήσεων (fixed basis functions) ωστόσο χρησιμοποιεί σταθερού μεγέθους χρονική μετατόπιση παραθύρου $w(n)$ για να λάβει την μετατροπή του σήματος. Η short-term Fourier transform μπορεί να εκφραστεί ως: [52]:

$$X(k,m) = \sum_{n=0}^{N-1} x(n+m)w(n)W_N^{nk}; \quad k,m = 0,1,\cdot\cdot,N-1 \tag{5}$$

Σε αυτήν την εξίσωση, το $X(k,m)$ αντιπροσωπεύει την k-οστή συνιστώσα συχνότητας του (STFT) ενός σήματος x τη χρονική στιγμή m. Το STFT είναι ένας τρόπος ανάλυσης ενός σήματος στον τομέα της συχνότητας σε σύντομες χρονικές περιόδους. Ορίζεται από την εξίσωση η οποία περιλαμβάνει μια άθροιση των τιμών του σήματος x, σταθμισμένη από μια συνάρτηση παραθύρου w. Η μεταβλητή n αντιπροσωπεύει τον δείκτη του δείγματος στο σήμα x. Ο όρος W_N^{nk} είναι η μετατόπιση του σήματος στον τομέα συχνότητας. Ο εκθέτης είναι nk, που σημαίνει ότι θα μετατοπίσει το σήμα κατά n φορές την συχνότητα που αντιπροσωπεύεται από τον δείκτη k. Η σταθερά W_N χρησιμοποιείται για την κλιμάκωση της μιγαδικής εκθετικής έτσι ώστε να μετατοπίσει το σήμα κατά την επιθυμητή ποσότητα. Η συνάρτηση παραθύρου w χρησιμοποιείται για τη στάθμιση των δειγμάτων του σήματος x και μπορεί να έχει διάφορες μορφές ανάλογα με τις επιθυμητές ιδιότητες του STFT. Η λειτουργία παραθύρου επιλέγεται συνήθως να έχει καλή ανάλυση συχνότητας και χαμηλή φασματική διαρροή, πράγμα που σημαίνει ότι θα καταγράφει με ακρίβεια το περιεχόμενο συχνότητας του σήματος χωρίς να εισάγει τεχνουργήματα [52].

Ένα STFT φίλτρο περιέχει τα ακόλουθα 3 βήματα:

•Analysis (Ανάλυση): υπολογισμός του STFT για το σήμα εισόδου $x(t)$, $F_x^\gamma(t,f) = \int_{-\infty}^{\infty} x(t')\gamma_{t,f}^*(t')\ \mathrm{d}t'$ όπου $\gamma_{tf}(t') = \gamma(t'-t)\, e^{j\,2\pi f t'}$ με $\gamma(t)$ να είναι το παράθυρο ανάλυσης.

•Weighting(Διαβάθμιση): πολλαπλασιασμός STFT με (t,f) (συνάρτηση βαθμίδας)weight function $M(t,f)$, δηλαδή υπολογισμός του $M(t,f)\ F_x^\gamma(t,f)$..

•Synthesis (Σύνθεση): Το σήμα εξόδου $y(t)$ αποκτάται από ένα ανάστροφο STFT,

$$y(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [M(t',f')\ F_x^\gamma(t',f')] g_{t',f'}(t)\ \mathrm{d}t'\ \mathrm{d}f'.$$

εδώ, $g_{t',f'}(t)\ \mathrm{d}t'\ \mathrm{d}f'$. όπου $g(t)$ είναι το παράθυρο που υποθέτουμε ότι ικανοποιεί (για τέλια ανακατασκευή όταν $M(t,f) \equiv 1$)).

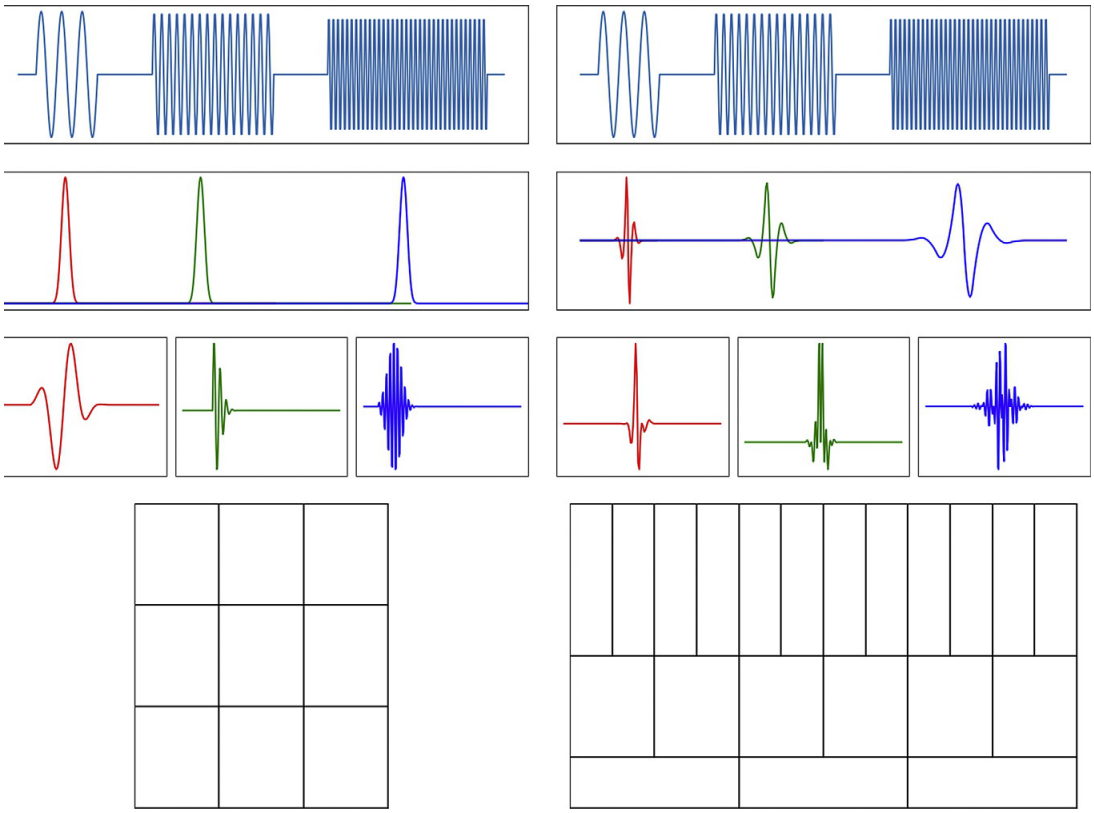
Αυτά τα βήματα εφαρμόζουν γραμμικό φίλτρο μεταβλητής χρόνου (Linear time-varying filter) που γράφεται ως $H_{\gamma,g}$ που εξαρτάται από t, f συνάρτηση βαθμίδας $M(t,f)$ και των παραθύρων $\gamma(t)$ και $g(t)$ [53]. Ωστόσο η STFT έχει καλύτερες ιδιότητες εντοπισμού χρόνου και συχνότητας σε σύγκριση με την Fourier transform. Αφού ο πολλαπλασιασμός της χρονικής ανάλυσης καθώς και της ανάλυσης της

συχνότητας είναι σταθερός (εξαιτίας της αρχής του Heisenberg για την αβεβαιότητα), τα παραγόμενα χαρακτηριστικά δεν μπορούν να πετύχουν άμεσο εντοπισμό του χρόνου μαζί με εντοπισμό της συχνότητας. Επιπλέον η χρήση αμετάβλητου μεγέθους παραθύρου καθώς και βάσης πάγιων συναντήσεων καθιστούν τα STFT να μην μπορούν να συλλάβουν περιπτώσεις με διαφορετική διάρκεια ή περιπτώσεις το σήμα έχει απότομες αλλαγές στα χαρακτηριστικά [52].Η ανάλυση στον χρόνο η την συχνότητα δεν μπορεί να είναι αυθαίρετα μικρή διότι ο πολλαπλασιασμός των δυο έχει κατώτατο όριο το οποίο πηγάζει από την αρχή της αβεβαιότητας του Heisenberg [54]. Αυτό σημαίνει ότι θα πρέπει να θυσιαστεί ανάλυσή χρόνου για ανάλυση συχνότητας η το ανάστροφο . Τα γκαουσιανά παράθυρα αντιμετωπίζουν τον περιορισμό με ισότητα [55].

Οι περισσότερες μέθοδοι χρησιμοποιούν STFT για προεπεξεργασία. Συγκεκριμένα η μίξη σημάτων ομιλίας πρώτα ”μεταμορφώνεται” από μονοδιάστατο σήμα στο τομέα του χρόνου σε δισδιάστατο φάσμα του τομέα χρονοσυχνότητας, στην πορεία η μείξη φάσματος διαχωρίζεται για να αποτελέσει φάσματα που αντιπροσωπεύουν διαφορετικές πηγές με την χρήση deep clustering (βαθιάς ομαδοποίησης) η mask estimation method (μέθοδος εκτίμησης μάσκας). Τέλος τα σαφώς διαχωρισμένα σήματα μπορούν να αποθηκευτούν με την χρήση ανάστροφου STFT στο κάθε φάσμα. Αυτή η μέθοδος έχει αρκετούς περιορισμούς. Αρχικά δεν γνωρίζουμε κατά ποσό η χρήση των STFT είναι ιδανική (ακόμα και εάν παράμετροι που επηρρεάζουν την λειτουργία του όπως ,μέγεθος , τύπος παραθύρου κ.α. είναι ιδανικές) για τον τομέα του διαχωρισμό πηγών. Επιπλέον οι μέθοδοι που χρησιμοποιούν STFT συχνά υποθέτουν λάθος ότι η φάση των διαχωρισμένων σημάτων είναι ίση με την φάση της μείξης πράγμα που επιπλέον προσθέτει ένα σαφές ανώτατο όριο στην απόδοση διαχωρισμού με την χρήση ideal masks (ιδανικών μασκών) [56].

Πριν την περίοδο της βαθιάς μηχανικής μάθησης πολλές παραδοσιακές μέθοδοι χρησιμοποιήθηκαν στο πρόβλημα του διαχωρισμού πηγών όπως non-negative matrix factorization (Παραγοντοποιήση μη αρνητικών πινάκων) [57] , computational auditory scene analysis (υπολογιστική ανάλυση ακουστικής σκηνής (CASA)) [58]. Ωστόσο αυτές οι μέθοδοι λειτουργούν για κλειστό σετ δεδομένων πράγμα που περιορίζει την πρακτικότητα τους. Με την εξέλιξη της βαθιάς μηχανικής μαθήσεως ξεκίνησε η δόμηση μοντέλων για διαχωρισμό σε άγνωστο αριθμό πηγών [37].

Ο μετασχηματισμός wavelet προσπαθεί να ελαχιστοποιήσει τους περιορισμούς των STFT και να κάνει καλύτερη δουλειά. Ξεκινά με τον ορισμό βασικών συναρτήσεων που καλούνται “mother wavelets” οι οποίες δεν είναι περιορισμένες σε μια οικογένεια συναρτήσεων. Επιπλέον οι βασικές συναρτήσεις έχουν ταυτόχρονα χρονικά τμήματα και τμήματα συχνοτήτων. Αυτό επιτρέπει την παραγωγή μιας σειράς από μεταβλητού μεγέθους wavelet συναρτήσεων. Το παρακάτω σχήμα δείχνει την διαφορά μεταξύ STFT και wavelet μετατροπής. Τα STFT διαιρούν τον τομέα της χρονοσυχνότητας σε ένα ίσου μεγέθους πλέγμα, ενώ για την μετατροπή wavelet βλέπουμε μια χονδροειδής προς λεπτή (δηλαδή ανομοιόμορφη) αναπαράσταση του σήματος [52].



Σχήμα 2.5: Wavelet VS STFT [52]

2.3 Παραγοντοποίηση μη αρνητικών πινάκων

Η παραγοντοποίηση μη αρνητικών πινάκων στοχεύει στην διάσπαση του πίνακα δεδομένων V σε έναν πολλαπλασιασμό πινάκων $(W * H)$. Μπορούμε να διαχωρίσουμε την ακριβή παραγοντοποίηση με παραγοντοποίηση εκτίμησης $V \approx R(WH)$. Για την παραγοντοποίηση με εκτίμηση θα πρέπει να οριστεί μια μετρική απόκλισης μεταξύ των δεδομένων V και του πίνακα ανακατασκευής R [59].

$$V \approx R(W, H) = WH \tag{6}$$

Η εξίσωση (6) μπορεί να γράφει στήλη προς στήλη ως:

$$V_i \approx R_i(W, H) = WH = \sum_j H_i^j W_j, \tag{7}$$

H_i^j είναι οι ενέργειές που χρησιμοποιήθηκαν για την κωδικοποίηση των δεδομένων. Το V_i είναι η i'οστή είσοδος και το R_i περιγράφει την i'οστό διάνυσμα ανακατασκευής. Ο δείκτης j περιγράφει το j'οστό διάνυσμα βάσης W_j που χρησιμοποιείται στην ανακατασκευή. Αυτό σημαίνει ότι κάθε διάνυσμα δεδομένων V_i που εκτιμάται από την ανακατασκευή R_i είναι ένας γραμμικός συνδυασμός από τα διανύσματα βάσεων W_j ρυθμιζόμενα από τις ενέργειες H_i^j. Γίνεται η χρήση της ευκλείδειας απόστασης ως κόστους για να ποσοτικοποιεί τον βαθμό εκτίμησης που έχει επιτευχθεί.

$$F(W, H) = 1/2 \, ||V - WH||^2 \tag{8}$$

Η αντιστοίχως

$$F(W, H) = 1/2 \sum_i \|V_i - \sum_j H_i^j W_j\|^2 \tag{9}$$

Φαίνεται στην (9) ότι οι ακολουθούμενοι τους κανόνες παραγοντοποίησης μη αρνητικών πινάκων μπορούμε να ελαχιστοποιήσουμε την (9) έως προς W και H, με διατήρηση της μη αρνητικότητας σε όλα τα τμήματα.

1. Υπολογισμός της ανακατασκευής:
$$\mathbf{R}_i = \sum_j H_i^j \mathbf{W}_j \, . \tag{10}$$

2. Ενημέρωση των ενεργειών
$$H_i^j \leftarrow H_i^j \odot \frac{\mathbf{V}_i^T \mathbf{W}_j}{\mathbf{R}_i^T \mathbf{W}_j} \, . \tag{11}$$

3. Εκ νέου εφαρμογή της (10) με τις καινούργιες ενέργειες Η

4. Ενημέρωση των βασικών διανυσμάτων

$$\mathbf{W}_j \leftarrow \mathbf{W}_j \odot \frac{\sum_i H_i^j N_i}{\sum_i H_i^j \mathrm{R}_i} \tag{12}$$

5. Επιστροφή πίσω στο βήμα 1 μέχρι να προκύψει σύγκλιση

Αυτοί οι κανόνες ενημέρωσης συνδυάζονται από την ομαλοποίηση των διανυσμάτων βάσης για την πρόληψη κοινών αριθμητικών μετατοπίσεων σε W, H σύμφωνα με:

$$\mathbf{W}_j \leftarrow \frac{\mathbf{W}_j}{\alpha_j} \tag{13}$$

Πρέπει να δοθεί προσοχή ώστε η 9 να μην επηρεαστεί από την κανονικοποίηση. Αυτό επιτυγχάνεται εάν ταυτοχρόνως κάνουμε επανακλιμάκωση των ενεργειών H_i^j με $1/\alpha_j$. Στην κλασσική περίπτωση η εξίσωση (11) φροντίζει για την επαναβάθμιση ωστόσο στην αραιά περίπτωσή τα πράγματα περιπλέκονται. Η παραγοντοποίηση μη αρνητικών πινάκων είναι μια μέθοδος gradient decent (κατάβαση κλίσης–δυναμικού) χωρίς παραμέτρους ενώ παράλληλα είναι υπολογιστικά φθηνό [60].

2.3.1 Αραιά παραγοντοποίηση μη αρνητικών πινάκων

Σε κάποιες περιπτώσεις βρίσκουμε παραγοντοποιήμενες λύσεις W, H στο πρόβλημα 6 που τοποθετούν αραιότητα στα τμήματα τους, Συγκεκριμένα στη συνάρτησή ενεργοποίησης H :. Επιπλέον η αραιότητα πλέον συμμετέχει στη δομή της συνάρτησης κόστους:

$$F(W,H) = \frac{1}{2} \sum_{\downarrow} \left\| \mathbf{V}_i - \sum_j H_i^j \mathbf{W}_j \right\|^2 + \lambda \sum_{i,j} g(H_i^j) \tag{14}$$

Βλέπουμε ότι η F είναι γραμμική σύνθεση ενός όρου ανακατασκευής και μιας ποινής αραιότητας. Από την (14) προκύπτει ένα πρόβλημα διαβάθμισης. Εφόσον είναι δυνατόν να διαβαθμίσουμε τα διανύσματα βάσεων W_j με τις σταθερές α_j και τις αντίστοιχες ενέργειες H_i^j με ενέργειες $1/\alpha_j$. Λαμβάνουμε το ίδιο κόστος ανακατασκευής άλλα διαφορετική ποινή αραιότητας. Αυτό αυτομάτως σημαίνει ότι με την μεγέθυνση των διανυσματικών βάσεων και σμίκρυνση των ενεργειών μπορούμε να λάβουμε μια μικρότερη συνάρτηση κόστους. Δηλαδή φτάνουμε στη σύγκλυση όταν η αραιότητα φτάνει στο 0 και τα διανύσματα βάσεων μεγαλώνουν χωρίς όριο. Έτσι οι λύσεις που βρίσκονται δεν εξαρτιούνται πλέον από την αραιότητα [60].

2.4 Independent component analysis (Ανάλυση ανεξάρτητων συνιστωσών) & Principal component analysis (Ανάλυση κύριων συνιστωσών)

2.4.1 Principal component analysis(Ανάλυση κύριων συνιστωσών)

Η ανάλυση κύριων συνιστωσών είναι μια κλασική στατιστική μέθοδος. Με την πρώτη εμφάνιση της το 1901 στη δουλεία του Pearson [61] ο οποίος πρότεινε μια μέθοδο γραμμικής παλινδρόμησης ελάχιστων τετράγωνων σε n διαστάσεις. Παρόλαυτα ο Hotelling [62] είναι αυτός που θεωρείτο ο ιδρυτής της μεθόδου με την εργασία του το 1933 πάνω στην ανάλυση της διακύμανσης πολυδιάστατων τυχαίων μεταβλητών [63].

Theorem (PCA) *Έστω ότι οι ιδιοτιμές $\lambda_1, \lambda_2, ..., \lambda_3$ του πίνακα αυτοσυσχέτισης R_x ($R_x = E\{xx^T\}$ με το x να είναι το διάνυσμα παρατήρησης και $E\{x\} = 0$ να είναι η μέση τιμή) να είναι κατανεμημένες σε φθίνουσα σειρά και τα $e_1, e_2, ..., e_n$ να είναι τα αντίστοιχα ορθοκανονικά ιδιοδιανύσματα. Τότε το μέσο τετραγωνικό σφάλμα J_{rec} ελαχιστοποιείται (με την διακύμανσή j_v να μεγιστοποιείται)*

$$W = W^* = TU \;\; \textit{όπου} \;\; U = [e_1, e_2, ..., e_m] \tag{15}$$

με n να είναι τυχαίες μεταβλητές, m βαθμοί ελευθέριας και W να είναι ο πίνακας απόμειξης με διαστάσεις $m * m$ ο οποίος μπορεί να γραφεί και ως γινόμενο ενός αντιστρέψιμου τετραγωνικού πίνακα $T \in \mathcal{R}^{m*m}$ και $V \in \mathcal{R}^{m*m}$. Μια ορθοκανονική βάση να αποτελεί τις γραμμές του έτσι ώστε το ελάχιστο τετραγωνικό σφάλμα να ορίζεται ως [64]:

$$\begin{aligned} \min J_{rec} &= \sum_{i=m+1}^{n} \lambda_i \\ \min J_v &= \sum_{i=1}^{m} \lambda_i \end{aligned} \tag{16}$$

Για τον μετασχηματισμό Karhunen-Loeve $y = U_x$ όπου y θεωρείτο το διάνυσμα χαρακτηριστικών υπάρχουν οι εξής ιδιότητές

- Οι κυρίες συνιστώσες είναι ανεξάρτητες μεταξύ τους
- Οι διακυμάνσεις του y ισούνται με τα λ του R_x
- Οι διακυμάνσεις του y έχουν φθίνουσα σειρά [65]

Για την χρήση του PCA καλό είναι τα δεδομένα να έχουν 5 βασικές υποθέσεις:

1. Οι λανθάνουσες μεταβλητές του διανύσματος πηγών να είναι στατιστικά ανεξάρτητες
2. Ο πίνακάς μείξης είναι τετραγωνικός πίνακάς δηλαδή ο αριθμός παρατηρήσεων είναι ίδιος με τον αριθμό πηγών
3. Το μοντέλο είναι απαλλαγμένο από θόρυβο δηλαδή η μονή στοχαστική παράμετρος είναι το διάνυσμα πηγών. Γιαυτό και πλέον δεν θεωρείται ιδανικό για πραγματικές περιπτώσεις

4. το διάνυσμα να έχει μηδενικό μέσο $E\{x\} = 0$

5. Οι μεμονωμένες συνιστώσες του διανύσματος παρατηρήσεων είναι μη-συχετιζόμενες άλλα όχι κατά ανάγκη ανεξάρτητες.Η λεύκανση μετασχηματίζει γραμμικά το διάνυσμα παρατηρήσεων με τρόπο ώστε ο πίνακας συσχέτισης R_x να ισούται με μοναδιαίο πίνακα [63, 64, 66].

2.4.2 Independent component analysis (Ανάλυση ανεξάρτητων συνιστωσών)

Η ανάλυση ανεξάρτητων συνιστωσών είναι μια εξέλιξη της ανάλυσης κύριων συνιστωσών που χρησιμοποιεί στατιστικές συναρτήσεις υψηλότερου βαθμού (higher order statistics). Το διάνυσμα παρατήρησης $x(k) = [x_1(k), x_2(k), \ldots, x_m(k)]^T$ που προκύπτει από τον γραμμικό μετασχηματισμό ενός διανύσματος στατιστικά ανεξάρτητων πηγαίων σημάτων είναι $\mathbf{s}(k) = [s_1(k), s_2(k), \ldots, s_n(k)]^T$

$$x(k) = As(k) \tag{17}$$

Έστω ότι ο πίνακας $A \in \mathcal{R}^{m*n}$ και πηγές $s_i(k)$ είναι άγνωστα ενώ παράλληλα

- Για κάθε ζευγάρι $i \neq j$ τα σήματα πηγών s_i, s_j είναι στατιστικά ανεξάρτητα .
- Ο πίνακας Α είναι είτε είναι τετραγωνικός η έχει παραπάνω γραμμές από στήλες. [63, 64, 66]

Η μέθοδος ICA είναι εφικτή όταν

- Γνωρίζουμε το πλήθος των πηγαίων σημάτων.
- Το πολύ ένα από τα σήματα πρέπει να ακολουθεί την γκαουσιανή κατανομή.
- Δεν μας ενδιαφέρει το πραγματικό πλάτος των σημάτων ούτε η σειρά τους.

Με την χρήση ενός γραμμικού μετασχηματισμού του διανύσματος παρατήρησης θα γίνει η ανάκτηση των άγνωστων σημάτων $y(k) = \mathbf{Wx(k)}$. Επειδή δεν γνωρίζουμε την σειρά καθώς και το πλάτος των σημάτων η σχέση μεταξύ διανύσματος εξόδου και διανύσματος πηγών θα δίνεται από τον γενικευμένο μετασχηματισμό μετάθεσης $y(k) = Ps(k)$. Όπου ο γενικευμένος πίνακας μετάθεσης P έχει όλα του τα στοιχεία μηδενικά έκτος από ένα σε κάθε στήλη [64, 66].

Κεφάλαιο 3ο: Νευρωνικά Δίκτυα

3.1 Εισαγωγή

Τα νευρωνικά δίκτυα εμφανίζουν μια σταθερή άνοδο τις τελευταίες δυο δεκαετίες με όλο και περισσοτέρους τομείς πλέον να έχουν απόδοση καλύτερη από αυτή του ανθρώπου. Συγκεκριμένα οι κυριότερες αρχιτεκτονικές ιστορικά είναι Multi-layer (πολλών στρωμμάτων) perceptron, Convolutional neural nets (συνελικτικά νευρωνικά δίκτυα), Recurrent neural nets (αναδρομικά νευρωνικά δίκτυα) και τα τελευταία 4 χρονιά υπάρχει μια άνοδος των Deep (βαθιών) transformers.

3.1.1 Perceptron

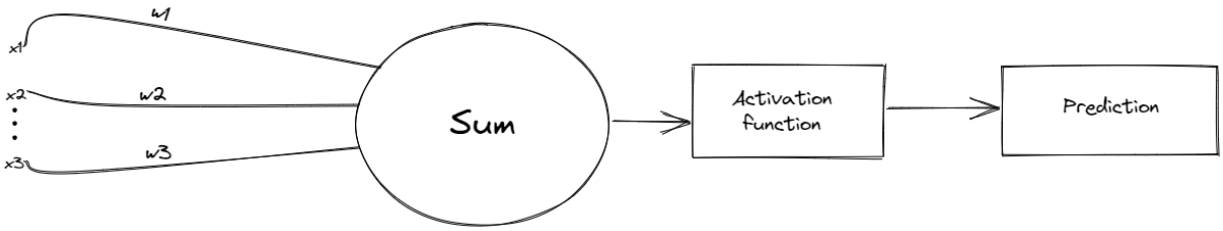
Το Perceptron είναι ένας νευρώνας με είσοδο είτε 0 είτε 1 δηλαδή $x_i \in 0, 1$ και με την συνάρτηση ενεργοποίησης να δίνεται από την Heaviside συνάρτηση:

$$\varphi(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0. \end{cases} \tag{18}$$

Η έξοδος του perceptron δίνεται ως μια πύλη κατωφλίου (threshold gate):

$$y = \varphi(x^T w - b) = \begin{cases} 0, & \text{if } \sum_{i=1}^{n} w_i x_i < b \\ 1, & \text{if } \sum_{i=1}^{n} w_i x_i \geq b. \end{cases} \tag{19}$$

Το κατώφλι b είναι μια μετρική του πόσο εύκολα πελήφθη η απόφαση της εξόδου $y = 0$. Για τις περισσότερες περιπτώσεις αυτό παρουσιάζεται ως βάρος w_0 και καλείται προκατάληψη (bias).



Σχήμα 3.1: Ενός στρώματος perceptron

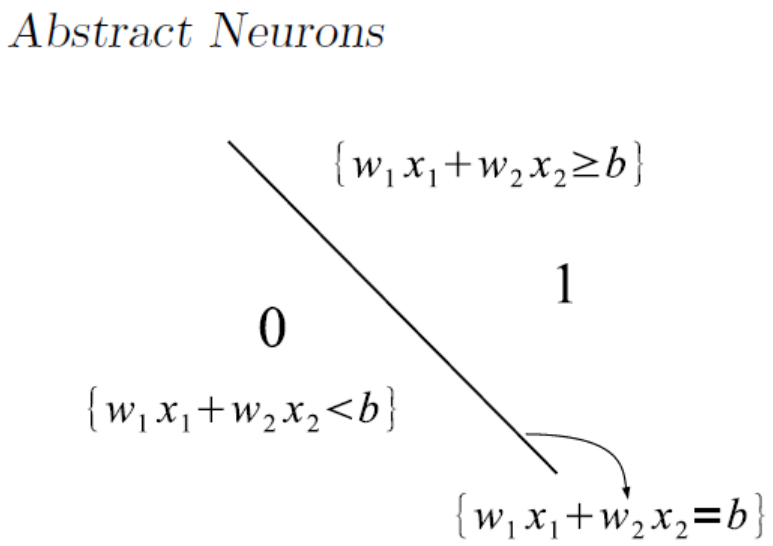
Στη συνέχεια θα δούμε γεωμετρική ανάλυση του Perceptron. Υποθέτοντας ένα υπερεπίπεδο διαστάσεων $(n-1)$ στους πραγματικούς αριθμούς $\mathbb{R}^n$ που ορίζεται από:

$$\mathcal{H} = \{(x_1\,, \cdot\,\cdot\,\cdot, x_n); \sum_{i=0}^{n} w_i x_i\} \tag{20}$$

Το κανονικό διάνυσμα N δίνεται σε σχέση με τα βάρη ως:

$$N^T = (\imath v_1 , \cdots, w_n). \tag{21}$$

Όπου T είναι το ανάστροφο διάνυσμα. Το υπερεπίπεδο περνάει από ένα σημείο p, το οποίο σχετίζεται με την προκατάληψη μέσω της σχέσης $b = p^T w$. Τότε η έξοδος y συνδυάζει την τιμή 0 ως ένα από τα δυο τμήματα που ορίζει το υπερεπίπεδο $\mathcal{H}$, και 1 ορίζει το υπόλοιπο υπερεπίπεδο. Για την περίπτωση 2 εισόδων $n = 2$, βλέπουμε το σχήμα 3.2 που μας δίνεται από [67]. Όπου δεδομένου ότι υπάρχει ένα σαφές όριο μεταξύ περιοχών, το Perceptron μπορεί να αποφασίσει σε ποία περιοχή ανήκει ένα σημείο. Με αυτόν τον τρόπο μπορούν να κωδικοποιηθούν οι λογικές πύλες AND & OR [67].



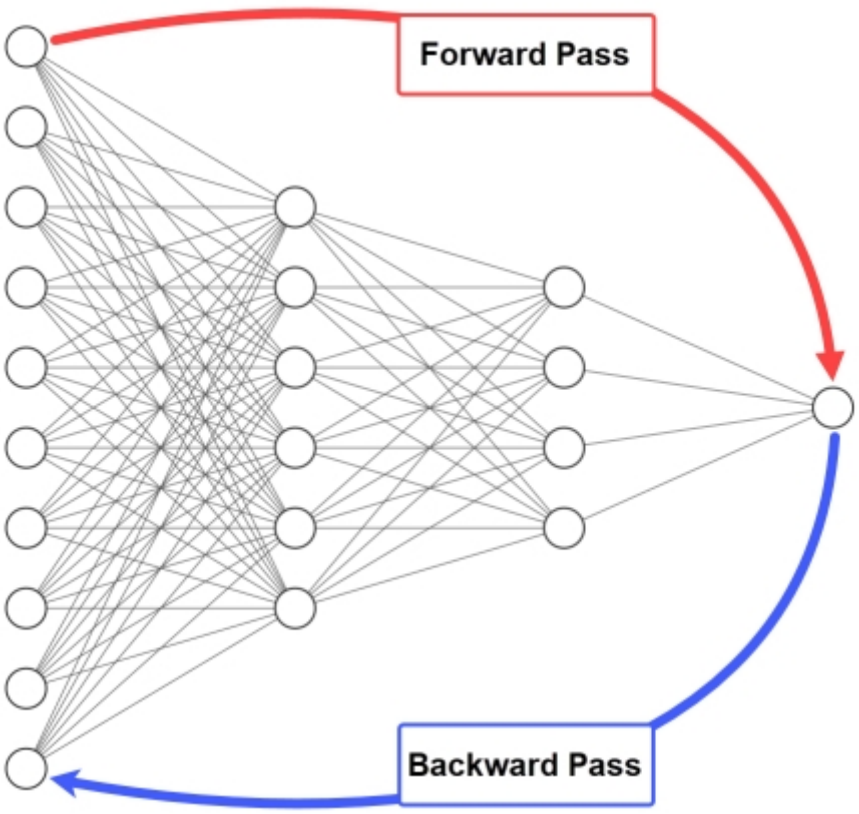
Σχήμα 3.2: Διαχωρισμός του υπερεπιπέδου [67].

3.1.2 Πολλαπλών στρωμάτων perceptron

Universal aproximator: *O George Cybenko το 1989 [68]αποδεικνύει ότι τα multi-layer perceptrons μπορούν να προσεγγίσουν οποιαδήποτε συνεχή συνάρτηση. Με την υπόθεση ότι το νευρωνικό δίκτυο χρησιμοποιεί την σιγμοειδή συνάρτηση ενεργοποίησης. Σήμερα έχει πάρει την θέση της η ReLU συνάρτηση ενεργοποίησης. Ωστόσο σε άλλες δημοσιεύσεις ([69], [70])φαίνεται πως δεν είναι απαραίτητη η χρήση της σιγμοειδής συνάρτησης ενεργοποίησης για να ισχύει το θεώρημα.*

Τα Perceptrons ενός στρώματος έχουν τον περιορισμό να μπορεί να αναπαραστήσει μόνο επίπεδες επιφάνειες [64]. Με την προσθήκη επιπλέον κρυφών στρωμάτων τα perceptrons μεταμορφώνουν ένα πρόβλημα μεγαλυτέρων διαστάσεων σε ένα πρόβλημα πολλών μικρών προβλημάτων επίπεδων επιφανειών. Για αυτή την περίπλοκη δομή χρησιμοποιείται ο αλγόριθμος Back-Propagation Learning [71, 72]που είναι άρρηκτα συνδεδεμένος με τα βαθιά νευρωνικά δίκτυα. Ο αλγόριθμος έχει θεσπιστεί ως βασικός πυλώνας για τον σχεδιασμό των multilayer perceptrons (MLP). Το στρώμα εισόδου και το στρώμα εξό-

δου συνδέουν το δίκτυο με τον εξωτερικό κόσμο. Επιπλέον από αυτά τα δυο στρώματα υπάρχουν και κρυφά στρώματα [71–73].



Σχήμα 3.3: backpropagation (εικόνα φτιάχθηκε με το NN-SVG [74] εργαλείο)

Η διαδικασία περιλαμβάνει τέσσερα βήματα:

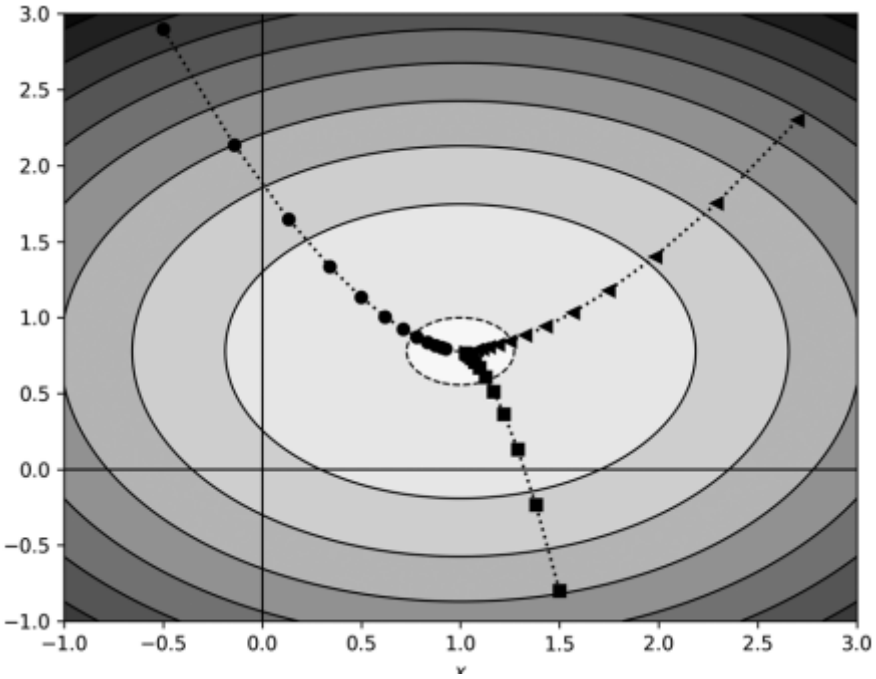
- feed foward (Εμπρόσθια τροφοδοσία): ένα διάνυσμα εισόδου παρουσιάζεται στο δίκτυο και αποθηκεύονται τα διανύσματα εξόδου και οι αξιολογημένες παράγωγοι των συναρτήσεων ενεργοποίησης.
- backward pass (Πίσω διάδοση): στο επίπεδο εξόδου: ακολουθείται η διαδρομή ανάστροφης διάδοσης από την έξοδο του δικτύου μέχρι τη μονάδα εισόδου και υπολογίζεται το σφάλμα ανάστροφης διάδοσης και η μερική παράγωγος.
- Διάδοση στα κρυφά στρώματα: το σφάλμα ανάστροφης διάδοσης υπολογίζεται για κάθε μονάδα στο κρυφό επίπεδο λαμβάνοντας υπόψη όλες τις πιθανές διαδρομές προς τα πίσω και υπολογίζεται η μερική παράγωγος.
- Ενημερώσεις βαρών: τα βάρη δικτύου ενημερώνονται στην κατεύθυνση αρνητικής κλίσης χρησιμοποιώντας μια σταθερά εκμάθησης. Είναι σημαντικό να κάνετε τις διορθώσεις στα βάρη μόνο αφού έχει υπολογιστεί το σφάλμα για όλες τις μονάδες του δικτύου [75].

3.1.3 Κατάβαση δυναμικού

Ο στόχος της εκπαίδευσης ενός νευρωνικού δικτύου είναι να προσαρμόσει τα βάρη και τις προκαταλήψεις του δικτύου έτσι ώστε να μπορεί να κάνει ακριβείς προβλέψεις σε αόρατα δεδομένα. Τα βάρη και οι προκαταλήψεις προσαρμόζονται για να ελαχιστοποιηθεί η διαφορά μεταξύ της πρόβλεψης και της πραγματικής εξόδου, η οποία μετράται από τη συνάρτηση απώλειας. Το Gradient descent χρησιμοποιείται για την εύρεση των τιμών των βαρών και των προκαταλήψεων που ελαχιστοποιούν το σφάλμα:

$$W \leftarrow W - \eta \, \Delta W$$
$$b \leftarrow b - \eta \, \Delta b \tag{22}$$

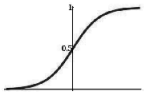
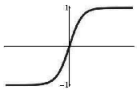
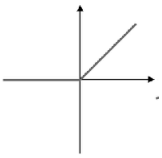
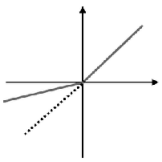
Σε αυτήν την εξίσωση 22, τα ΔW και Δb είναι σφάλματα που βασίζονται στις μερικές παραγώγους των βαρών και των προκαταλήψεων (bias), αντίστοιχα. Το η (eta) είναι μια παράμετρος κλίμακας ή ρυθμός εκμάθησης, μια τιμή που χρησιμοποιούμε για να προσαρμόσουμε τον τρόπο με τον οποίο κινούμαστε.



Σχήμα 3.4: Κατάβαση δυναμικού με 2 διαστάσεις [76]

Και οι τρεις διαδρομές κατάβασης που βλέπουμε στο 3.4 με κλίση συγκλίνουν στο ελάχιστο της συνάρτησης. Αυτό δεν προκαλεί έκπληξη δεδομένου ότι η συνάρτηση έχει μόνο ένα ελάχιστο. Η κατάβαση (δυναμικού)κλίσης θα βρει σίγουρα το ελάχιστο της συνάρτησης. Μπορεί να απαιτηθούν πολλά βήματα εάν το μέγεθος του βήματος είναι υπερβολικά μικρό. Η κάθοδος μπορεί να τείνει να κυμαίνεται γύρω από το ελάχιστο, αλλά να το ξεπερνάει συνεχώς εάν το μέγεθος του βήματος είναι πολύ μεγάλο [76].

3.2 Συναρτήσεις ενεργοποίησης

Name	Functions	Derivatives	Figure
Sigmoid	$\sigma(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1 - f(x))^2$	
tanh	$\sigma(x) = \frac{e^x - e^- x}{e^z + e^- z}$	$f'(x) = 1 - f(x)^2$	
ReLU	$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0. \end{cases}$	$f'(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases}$	
Leaky ReLU	$f(x) = \begin{cases} 0.01x & \text{if } x < 0 \\ x & \text{if } x \geq 0. \end{cases}$	$f'(x) = \begin{cases} 0.01 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases}$	
Softmax	$f(x) = \frac{e^x}{\sum_1^j e^x}$	$f'(x) = \frac{e^x}{\sum_1^j e^x} - \frac{(e^x)^2}{(\sum_1^j e^x)^2}$	

Σχήμα 3.5: Συναρτήσεις ενεργοποίησης [77]

3.2.1 Basic Rectified Linear Unit (ReLU)

Στην ηλεκτρονική αυτό αναφέρεται ως half-wave rectification. Ωστόσο, με σαφή βιολογικά κίνητρα και μαθηματικές αιτιολογήσεις, η συνάρτηση ενεργοποίησης εισήχθη για πρώτη φορά σε ένα δυναμικό δίκτυο [78, 79]. Επιπλέον, επιτρέπει την πιο αποτελεσματική εκπαίδευση βαθύτερων δικτύων σε σύγκριση με τις κοινώς χρησιμοποιούμενες συναρτήσεις ενεργοποίησης. Σύμφωνα με έρευνα του 2018 [80] , η RELU ήταν η πιο δημοφιλής συνάρτηση ενεργοποίησης για βαθιά νευρωνικά δίκτυα. Τέλος, επειδή forward και back propagation (εμπρόσθια και πίσω διάδοση) είναι γρήγορες διαδικασίες, ο υπολογισμός του αποτελέσματος της συνάρτησης και της κλίσης είναι απλός. Σύμφωνα με τα ευρήματα της μελέτης, η ReLU είναι έξι φορές ταχύτερη από άλλες γνωστές συναρτήσεις ενεργοποίησης [81, 82].

3.3 Αναδρομικά νευρωνικά δίκτυα

3.3.1 RNNs (Recurrent Neural Networks)

Η βασική αρχή στα αναδρομικά δίκτυα είναι ότι το διάνυσμα εισόδου και κάποια πληροφορία από το προηγούμενο βήμα χρησιμοποιούνται για να υπολογιστεί η έξοδος όπου μαζί με την πληροφορία τροφοδοτούνται στο επόμενο βήμα. Επιπλέον ένα μοντέλο μπορεί να διαχειριστεί ακολουθίες με διαφορετικά μεγέθη. Γενικότερα οι φόρμουλες που χρησιμοποιούνται για τον υπολογισμό των τιμών εξόδου για κάθε βήμα ονομάζονται blocks. Έτσι για το απλούστερο αναδρομικό δίκτυο, ένα μπλοκ μπορεί να οριστεί ως εξής:

$$
\begin{aligned}
a^{<t>} &= f_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) \\
\hat{y}^{<t>} &= f_2(W_{ya}a^{<t>} + b_y),
\end{aligned}
\tag{23}
$$

Όπου $x^{<t>}$ είναι ένα διάνυσμα ή ένας αριθμός που είναι μέρος μιας ακολουθίας εισόδου. Το t σηματοδοτεί το βήμα των επαναλαμβανόμενων υπολογισμών, W_{aa}, W_{ax}, W_{ya}, b_a , b_y είναι οι πίνακές βαρών και διανύσματά με δεδομένη διάστασή. Τα f_1 και f_2 είναι οι συναρτήσεις ενεργοποίησης. Συνήθως για f_1, χρησιμοποιούμε την tanh ή την ReLU, ενώ για f_2, αφού υπολογίσει την τιμή της εξόδου χρησιμοποιούμε είτε σιγμοειδή είτε softmax [83].

3.3.2 Δίκτυο μακροπρόθεσμης μνήμης (Long short-term memory)

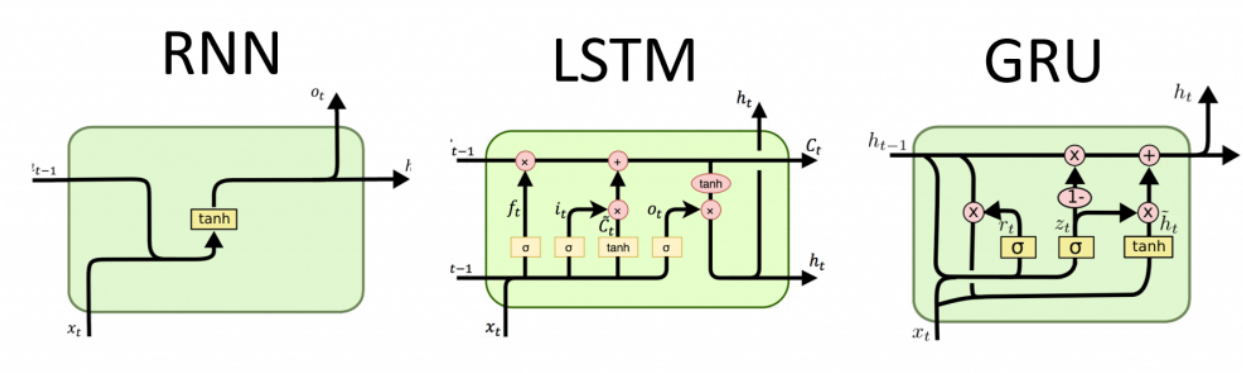
το Δίκτυο μακροπρόθεσμης μνήμης είναι ένας ειδικός τύπος από αναδρομικά νευρωνικά δίκτυά. Συγκεκριμένα αυτή η αρχιτεκτονική ανακαλύφθηκε για την επίλυσή του προβλήματος vanishing και exploding gradients (Οι Yoshua Bengio, Patrice Simard και Paolo Frasconi στην εργασία τους το 1994 [84]. Παρατήρησαν ότι όταν εκπαιδεύονται βαθιά νευρωνικά δίκτυα, οι διαβαθμίσεις των βαρών μπορεί να γίνουν είτε πολύ μικρές (εξαφανιζόμενες κλίσεις) είτε πολύ μεγάλες (τεράστιες κλίσεις)). Επιπλέον αυτός ο τύπος δικτύων είναι καλύτερος για συγκράτηση συνδέσεων-εξαρτήσεων μεγάλων αποστάσεων στα δεδομένα. Αυτό επιτυγχάνεται αναγνωρίζοντάς την σχέση μεταξύ τιμών στην αρχή και στο τέλος της σειράς των δεδομένων. Το δίκτυο μακροπρόθεσμης μνήμης εισαγάγει όρους όπως πύλες ειδικότερα πύλες τριών ειδών:

- Forget gate (Πύλη υπενθύμισης): ελέγχει την διάχυση πληροφορίας από το προηγούμενο βήμα δηλαδή το πόσο θυμάται το μπλοκ τις προηγούμενες καταστάσεις γιαυτό και ονομάζεται πύλη μνήμης.
- Update (input) gate (Πύλη ανανέωσης): αποφασίζει την αλλαγή που θα πραγματοποιηθεί στο κελί και ρυθμίζει πόση πληροφορία το παρόν κελί θα λάβει από μελλοντικά κελιά.
- Output gate (Πύλη εξόδου):ελέγχει την τιμή της επομένης κρυφής κατάστασης.

Μαθηματικά ένα LSTM ορίζεται ως:

$$
\begin{aligned}
\Gamma_u &= \sigma(W_{uu}a^{<t-1>} + W_{ux}x^{<t>} + b_u) \\
\Gamma_f &= \sigma(W_{ff}a^{<t-1>} + W_{fx}x^{<t>} + b_f) \\
\Gamma_o &= \sigma(W_{oo}a^{<t-1>} + W_{ox}x^{<t>} + b_o) \\
\hat{c}^{<t>} &= tanh(W_{cc}a^{<t-1>} + W_{cx}x^{<t>} + b_c) \\
c^{<t>} &= \Gamma_u \odot \hat{c}^{<t>} + \Gamma_f \odot c^{<t-1>} \\
a^{<t>} &= \Gamma_o \odot \tanh(c^{<t>}),
\end{aligned}
\tag{24}
$$

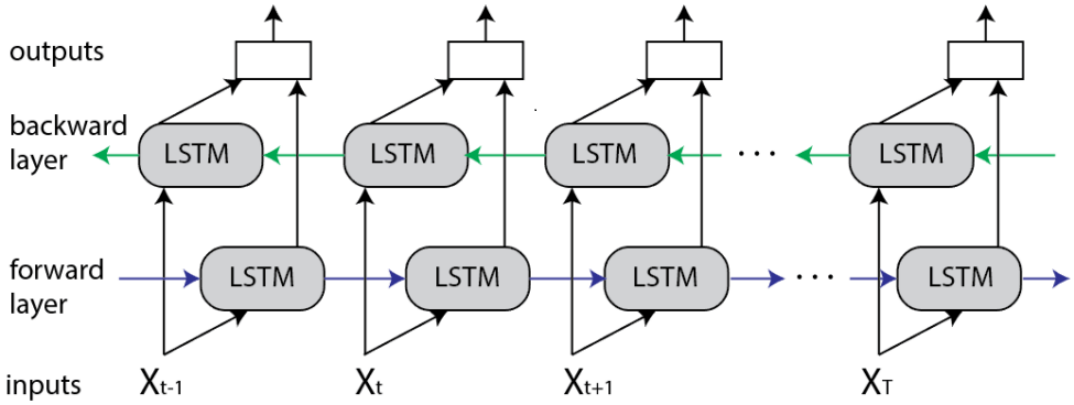
W και b είναι οι πίνακες και διανύσματα βαρών, t είναι η παρούσα επανάληψη του επαναλαμβανόμενου υπολογισμού στο νευρωνικό δίκτυο, Γ_u είναι η update gate, Γ_f είναι η forget gate, Γ_o είναι η output gate, $\hat{c}^{<t>}$ είναι η πιθανή τιμή του κελιού, $c^{<t>}$ είναι η τωρινή τιμή του κελιού, και $a^{<t>}$ είναι η τιμή εξόδου ή αλλιώς η κρυφή κατάσταση. Η Αρχιτεκτονική μπορεί να φανεί στο παρακάτω σχήμα 3.6 [83]:



Σχήμα 3.6: Αρχιτεκτονικές αναδρομικών νευρωνικών δικτύων [85]

3.3.3 Διπλής κατεύθυνσης LSTM

Η πρώτη εμφάνιση των Bidirectional LSTM έγινε στην δουλεία των Mike Schuster και KuldipK. Paliwal το 1997 [86]. Αποδεικνύεται ότι το πλαίσιο ολόκληρης της εκφοράς χρησιμοποιείται για να ερμηνεύσει αυτό που λέγεται και όχι μια γραμμική ερμηνεία. Με την δουλεία των Alex Graves και Jurgen Schmidhuber το 2005 [87] η χρήση των Bidirectional LSTM έγινε γνωστή. Τα Bidirectional LSTM (BiLSTM) είναι αναδρομικά νευρωνικά δίκτυά που χρησιμοποιούνται κυρίως στον τομέα της ανάλυσης φυσικής γλωσσάς. Αντιθέτως με τα κλασσικά LSTM, η είσοδος διαχέεται και από τις δυο κατευθύνσεις και είναι ικανά να εκμεταλλευτούν πληροφορία και από τις δυο πλευρές. Επιπλέον είναι ένα ισχυρό εργαλείο για σειριακές εξαρτήσεις μεταξύ λέξεων και προτάσεων με οποιαδήποτε κατεύθυνση στη σειρά δεδομένων. Για παράδειγμά ένα αναδιπλωμένο BiLSTM παρουσιάζεται στο σχήμα 3.7:



Σχήμα 3.7: Αρχιτεκτονική διπλής κατεύθυνσης αναδρομικά δικτύων [83]

Τα BiLSTM θα έχουν διαφορετική έξοδο για κάθε τμήμα της σειράς δεδομένων. Ως αποτέλεσμα τα BiLSTM έχουν πλεονέκτημα σε προβλήματα ανάλυσής φυσικής γλωσσάς όπως κατηγοριοποίηση προτάσεων και μετάφρασή. Τέλος όσο αναφορά τα μειονεκτήματα σε σχέση με τα κλασσικά LSTM είναι σαφώς πιο αργά μοντέλα και χρειάζεται πολύς περισσότερος χρόνος εκπαίδευσης [83].

3.4 Transformers και attetion (Προσοχή)

Η αρχιτεκτονική των Transformer πρωτοεμφανίστηκε με τον Ashish Vaswani το 2017 [88,89] έπειτα ακολούθησε ανοδική πορεία σε τομείς όπως αναγνώριση ομιλίας, παράγωγη ομιλίας , όραση υπολογιστή και άλλα. Προτού μπούμε στις τεχνικές λεπτομέρειές της αρχιτεκτονικής των Transformer θα πρέπει να αναλύσουμε πρώτα ένα βασικό πυλώνα της αρχιτεκτονικής που είναι η προσοχή (Self-attention) [88,89].

3.4.1 Self-attention (αυτοπροσοχή)

Self-attention: *Η αυτοπροσοχή είναι μια ενέργεια όπου μια σειρά διανυσμάτων εισάγεται και μια σειρά διανυσμάτων εξάγεται. Με τα διανύσματα εισόδου να είναι τα $x_1, x_2, ..., x_t$ και τα αντίστοιχα διανύσματα εξόδου να είναι τα $y_1, y_2, ..., y_t$ με όλα τα διανύσματα να έχουν το ίδιο μέγεθός (διάσταση)k.*

Για την παράγωγη του διανύσματος εξόδου y_i, η αυτοπροσοχή απλώς παίρνει ένα διαβαθμισμένο μέσο για όλα τα διανύσματα εισόδου [88, 89].

$$\mathbf{y}_{\mathrm{i}} = \sum_{\mathrm{i}} w_{\mathrm{ij}} \mathbf{x}_{\mathrm{j}} \tag{25}$$

Όπου j είναι ο δείκτης για όλη την σειρά. Τα βάρη w_ij δεν είναι παράμετρος όπως σε κανονικά νευρωνικά δίκτυά αλλά εξάγονται από μια συνάρτηση για x_i και x_j. Η απλούστερη μορφή μιας τέτοιας συνάρτησης είναι το dot product (πολλαπλασιασμό διανυσμάτων) [88, 89].

$$w'_{\mathrm{ij}} = {\chi_{\mathrm{i}}}^{\mathrm{T}} \chi_{\mathrm{j}} \tag{26}$$

Σημείωση ότι x_i είναι το διάνυσμα εισόδου με ίδιο δείκτη με το παρόν διάνυσμα εξόδου y_i. Για το επόμενο διάνυσμα εξόδου λαμβάνουμε ολοκληρωτικά καινούργιες σειρές από dot products (πολλαπλασιασμό διανυσμάτων), και διαφορετικά διαβαθμισμένο άθροισμα. Ο πολλαπλασιασμός διανυσμάτων δίνει τιμές μεταξύ θετικού απείρου και αρνητικού απείρου $\cdot \in (\infty, -\infty)$, γιαυτό εφαρμόζεται ένα softmax για να αντιστοιχίζει τις τιμές σε εύρος $[0, 1]$ για να εξασφαλιστεί ότι το άθροισμα οδηγεί στην μονάδα για όλη την σειρά [88, 89]:

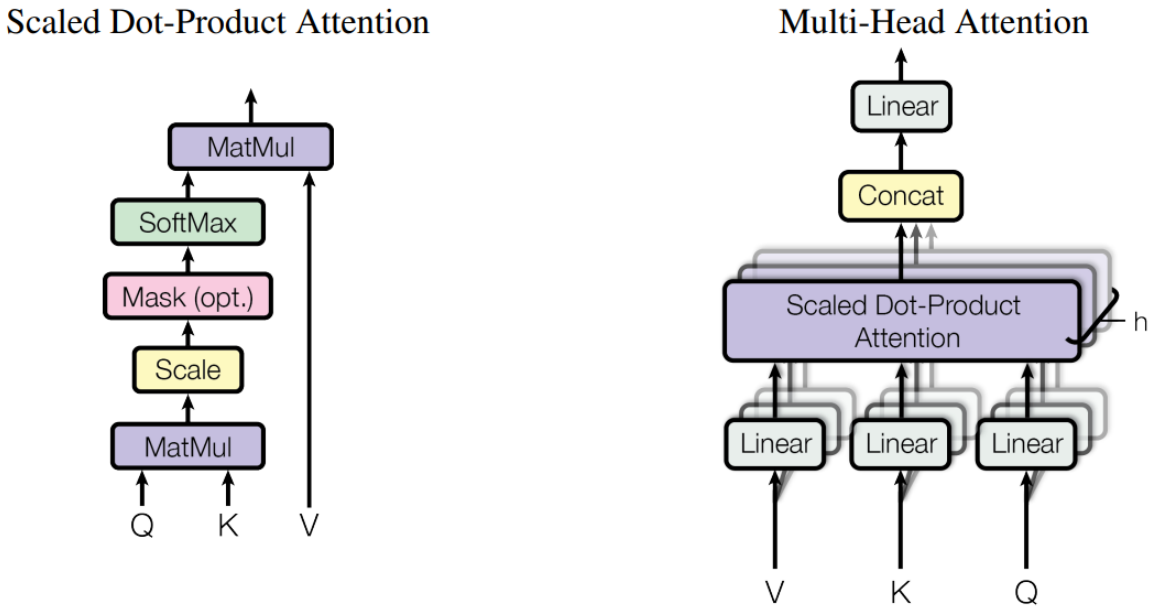
$$w_{\mathrm{ij}} = \frac{\exp\ w'_{\mathrm{ij}}}{\sum_{\mathrm{i}} \exp\ w'_{\mathrm{ii}}}. \tag{27}$$

Αυτό αποτελεί την βασική ιδέα πίσω από την αυτοπροσοχή. Ο πολλαπλασιασμός διανυσμάτων εκφράζει την συσχέτιση μεταξύ δυο διανυσμάτων στην σειρά εισόδου. Η συσχέτιση ορίζεται κάθε φορά από το πρόβλημα . Τα διανύσματα εξόδου είναι αναβαθμισμένες αθροίσεις καθόλη την σειρά εισόδου. Επιπλέον τα βάρη ορίζονται από των πολλαπλασιασμό διανυσμάτων. Πρώτου προχωρήσουμε πρέπει να αναφερθούν κάποιες ασυνήθιστες ιδιότητές:

- *η αυτοπροσοχή καθορίζεται πλήρως από τον μηχανισμό δημιουργίας σειρών εισόδου. Μηχανισμοί*

μεγέθυνσης (upstream), όπως ένα embedding στρώμα, οδηγούν τον αυτοπροσοχής με την εκμάθησή αναπαραστάσεων με συγκεκριμένα dot products.

- *Η αυτοπροσοχή αντιλαμβάνεται την είσοδο της ως ένα σετ και όχι ως μια σειρά - ακολουθία. Εάν μεταθέσουμε την σειρά εισόδου η σειρά εξόδου θα παραμείνει ακριβώς ίδια πράγμα που προβλέπει η αρχιτεκτονική του transformer άλλα η αυτοπροσοχή αγνοεί πλήρως την σειριακή φύση της ακολουθίας (σειράς) [88, 89].*



Σχήμα 3.8: Διαβαθμισμένο προϊόν πολλαπλασιασμού διανυσμάτων & Προσοχή πολλαπλών κεφαλών [88]

Η πραγματική χρήση της αυτοπροσοχή βασίζεται σε 3 επιπλέον πυλώνες.

1. Queries, keys και values κάθε διάνυσμα εισόδου x_i χρησιμοποιείται με 3 διαφορετικούς τρόπους.
   - Συγκρίνεται με τα υπόλοιπά διανύσματά για να υπολογίσει τα βάρη για την έξοδο του y_i.
   - συγκρίνεται με τα υπόλοιπά διανύσματά για να υπολογίσει τα βάρη για την έξοδο του $j-$ οστού διανύσματος y_j
   - χρησιμοποιείται ως μέρος των διαβαθμισμένων αθροίσεων για των υπολογισμό κάθε διανύσματος εξόδου εφόσον τα βάρη έχουν εδραιωθεί.

   Στην βασική αυτοπροσοχή που είδαμε μέχρι τώρα κάθε διάνυσμα εισόδου πρέπει να υιοθετήσει και τους 3 ρόλους. Μπορούμε να εξάγουμε καινούρια διανύσματα για κάθε ρόλο, με μια εφαρμογή γραμμικού μετασχηματισμού στο αρχικό διάνυσμα εισόδου. Με άλλα λογία προσθέτουμε 3 πίνακές βαρών $k * k$ διαστάσεων W_q, W_k, W_v αντιστοίχως και υπολογίζουμε 3 γραμμικούς μετασχηματισμούς δια κάθε x_i για τα 3 διαφορετικά μέρη της αυτοπροσοχής [88, 89]:

$$\mathbf{q}_\mathrm{i} = W_\mathbf{q}\mathbf{x}_\mathrm{i} \qquad \mathbf{k}_\mathrm{i} = W_\mathrm{k}\mathbf{x}_\mathrm{i} \qquad \mathbf{v}_\mathrm{i} = W_\nu\mathbf{x}_\mathrm{i} \tag{28}$$

$$
\begin{aligned}
w'_{\mathrm{ij}} &= \mathbf{q_i}^{\mathsf{T}} \mathbf{k_j} \\
w_{\mathrm{ij}} &= \mathbf{sofmax}(w'_{\mathrm{ij}}) \\
\mathrm{y_i} &= \sum_j w_{\mathrm{ij}} v_{\mathrm{j}} \,.
\end{aligned}
$$

Αυτό επιτρέπει στο στρώμα αυτοπροσοχής κάποιες ρυθμίσιμες παραμέτρους και επιπλέον την επεξεργασία εισερχομένων διανυσμάτων για να εφαρμόσει τους 3 ρόλους.

2. Διαβάθμιση του πολλαπλασιασμού διανυσμάτων

   Η συνάρτησή softmax μπορεί να γίνει πολύ ευαίσθητη σε μεγάλους αριθμούς εισόδου. Ως αποτέλεσμα να καταστρέφονται οι gradients (κλίσεις) και να καθυστερείται η εκπαίδευση μέχρι και την διακοπή της. Αφού η μέση τιμή των πολλαπλασιασμένων διανυσμάτων μεγαλώνει με την διάσταση των embeddings κ η διαβάθμιση αποτρέπει τις εισόδους από το να μεγαλώνουν υπερβολικά προτού εισέλθουν στο softmax [88, 89]:

$$
w'_{\mathrm{ij}} = \frac{\mathbf{q_i}^{\mathsf{T}} \mathbf{k}_j}{\sqrt{\mathbf{k}}} \tag{29}
$$

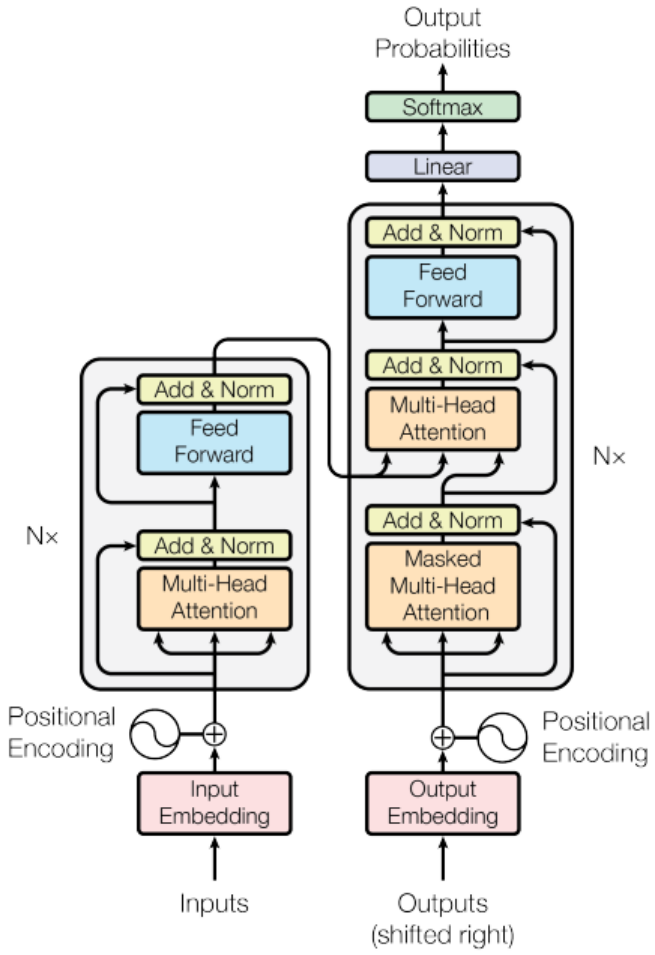
   Γιατί όμως επιλέχθηκε η $\sqrt{\mathbf{k}}$? Φανταζόμαστε ένα διάνυσμα στους $\mathbb{R}^k$ με όλες τις τιμές του να είναι c. Το ευκλείδειο μήκος είναι $\sqrt{\mathbf{k}}c$. Διαιρούμε το ποσό κατά το οποίο η αύξηση της διάστασης αυξάνει το μήκος των μέσων διανυσμάτων.

3. Πολλών κεφαλών προσοχή: Μπορούμε να έχουμε μεγαλύτερη διαχωριστική δύναμη με τον συνδυασμό πολλών μηχανισμών αυτοπροσοχής όπου θα έχουν τον δείκτη r. Κάθε μηχανισμός θα έχει τους αντίστοιχους W_k^r, W_q^r, W_v^r πίνακες. Καθεμιά αυτοπροσοχή ονομάζεται κεφαλή προσοχής. Για είσοδο x_i κάθε κεφαλή προσοχής παράγει ένα διαφορετικό διάνυσμα εξόδου y_i^r. Με απλή σύνθεση αυτών των κεφαλών αυτοπροσοχής και στη συνέχεια με έναν γραμμικό μετασχηματισμό ελαττώνουμε την διάσταση πίσω στον αριθμό κ

4. Αποδοτική Πολλών κεφαλών προσοχή: Ο ποιο απλός τρόπος να κατανοήσουμε την πολλών κεφαλών προσοχή είναι να δούμε ένα μικρό αριθμό τμημάτων αυτοπροσοχής να εφαρμόζεται παράλληλα με επιπλέον χρονική πολυπλοκότητά με r φορές. Ωστόσο υπάρχει ένας τρόπος έτσι ώστε να μειώσουμε την χρονική πολυπλοκότητά σχεδόν στα επίπεδά μονής κεφαλής κρατώντας παράλληλα τα πλεονεκτήματά των παράλληλων υπολογισμών. Για να το πετύχουμε αυτό τμηματοποιούμε κάθε εισερχόμενο διάνυσμα και για κάθε τμήμα παράγουμε W_k^r, W_q^r, W_v^r [88, 89].

3.4.2 Αρχιτεκτονική transformer block

Το block εφαρμόζει σειριακά τα εξής: Ένα στρώμα self attention (αυτοπροσοχής), ένα στρώμα normalization (ομαλοποίηση), ένα feed forward (εμπρόσθιας τροφοδοσίας) στρώμα και ένα επιπλέον στρώμα normalization (ομαλοποίηση). Υπολειμματικές συνδέσεις προστίθενται πριν το normalization (ομαλοποίηση), η σειρά τον τμημάτων δεν είναι οριστικοποιημένη άλλα η χρήση των τμημάτων θεωρείται σταθερή. Normalization (Ομαλοποίηση) και υπολειμματικές συνδέσεις θεωρούνται βασικά τεχνάσματα για να βοηθήσουν τα νευρωνικά δίκτυά στο να εκπαιδευτούν γρηγορότερα και με μεγαλύτερη ακρίβεια [88, 89].



Σχήμα 3.9: Αρχιτεκτονική μπλοκ προσοχής [88, 89]

Ο πιο συνήθης τρόπος να φτιάξεις έναν κατηγοριοποιητή σειράς από στρώματα ακολουθίας σε ακολουθία είναι η εφαρμογή global average pooling (παγκόσμια μέση συγκέντρωση) στην τελική ακολουθία εξόδου και να αντιστοιχίσεις το αποτέλεσμα σε ένα διάνυσμα κλάσης που έχει περαστεί από softmax. Μια ματιά σε ένα απλό κατηγοριοποιητή αποτελούμενο από transformer με απλές ακολουθίες. Η σειρά εξόδου παράγει κατά μέσο όρο ένα διάνυσμα αναπαράστασης όλης της σειράς. Αυτό το διάνυσμα προβάλλεται σε ένα διάνυσμα με ένα στοιχείο ανά κλάση και περνά από ένα softmax για τον υπολογισμό πιθανοτήτων [88, 89].

3.4.2.1 Encoding problem (Πρόβλημα κωδικοποίησης)

Εάν ανακαλέσουμε τα δεδομένα θα θέλαμε το μοντέλο να έχει μια ευαισθησία στην σειρά των δεδομένων. Αυτό αντιμετωπίζεται εύκολα δημιουργώντας ένα διάνυσμα με ίδιο μήκος που αντιπροσωπεύει την θέση του στοιχείου στην σειρά. Το διάνυσμα αυτό θα προστεθεί στα αντίστοιχά embeddings [88, 89].

3.4.2.2 Position encoding (Κωδικοποίηση θέσης)

Η κωδικοποίησή θέσης λειτουργεί με τον ίδιο τρόπο όπως τα embeddings μόνο που δεν μαθαίνονται τα

διανύσματα θέσης άλλα ανταυτού διαλέγεται μια συνάρτηση $f : \mathbb{N} \rightarrow \mathbb{R}k$ για να αντιστοιχίσει τις θέσεις με πραγματικών τιμών διανύσματα και να αφήσει το νευρωνικό δίκτυο να αποφασίσει πως θα αντιληφθεί αυτή την κωδικοποίηση. Το πλεονέκτημα για μια συνάρτηση που ταιριάζει στο πρόβλημα είναι ότι το μοντέλο θα μπορεί να ανταπεξέλθει σε σειρές μεγαλύτερες από αυτές που έχει δει κατά την διάρκεια της εκπαίδευσης. Το κύριο μειονέκτημα είναι ότι η απόφαση της κατάλληλης συνάρτησης είναι περίπλοκη υπόθεση και περιπλέκει κατά πολύ την υλοποίηση [88, 89].

3.4.2.3 Αδυναμία των κλασσικών αναδρομικών νευρωνικών δικτύων

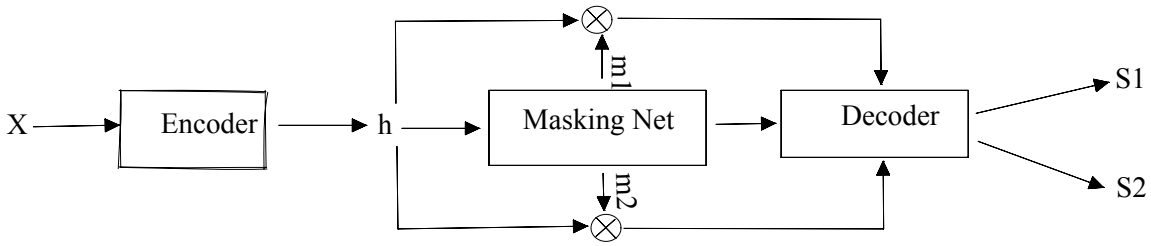
Η μεγάλη αδυναμία των κλασσικών αναδρομικών νευρωνικών δικτύων είναι οι αναδρομικές συνδέσεις. Παρόλο που επιτρέπει την πληροφορία να διαχέεται σε όλη την ακολουθία επίσης σημαίνει ότι δεν μπορούμε να υπολογίσουμε το κελί στο χρονικό βήμα i μέχρι να έχουμε υπολογίσει το προηγούμενο βήμα, αντιθέτως το 1D convolution (συνέλιξη μιας διάστασης) έχει μεγαλύτερη ελευθερία στους υπολογισμούς. Το μειονέκτημα των convolutions (συνελίξεων) είναι ότι έχουν σοβαρό περιορισμό σχετικά με την μοντελοποίηση εξαρτήσεων με μεγάλη απόσταση μεταξύ τους. Η αρχιτεκτονική των transformers είναι μια προσπάθεια για να έχουμε τα καλά των δυο κόσμων. Μπορούν να μοντελοποιήσουν εξαρτήσεις με μεγάλη απόσταση μεταξύ τους στη σειρά καθώς κρατούν των παράλληλο υπολογισμό [88, 89].

Κεφάλαιο 4ο: Αξιόλογοι αλγόριθμοι

Σε αυτό το κεφάλαιο θα εμβαθύνουμε σε 5 βασικούς αλγορίθμους που είτε είναι αυτοί με τα καλύτερα αποτελέσματα στον κλάδο είτε η σχέση απόδοσης και υπολογιστικού κόστους είναι εντυπωσιακή. Επίσης θα εστιάσουμε σε αλγορίθμους που πλησιάζουν περισσότερο σε πραγματικές συνθήκες.

4.1 Sepformer

Το προτεινόμενο μοντέλο [44] βασίζεται στην εκμάθησή μιας μάσκας και την χρησιμοποίηση 3 δικτύων: κωδικοποιητή, αποκωδικοποιητή και δίκτυο μάσκας όπως βλέπουμε στο σχήμα (4.1). Ο κωδικοποιητής είναι πλήρως συνελiκτικός ενώ το δίκτυο μάσκας χρησιμοποιεί δυο Transformers ενσωματωμένους σε ένα dual-path (διπλής κατεύθυνσης) μπλοκ επεξεργασίας που προτάθηκε στο [36]. Ο αποκωδικοποιητής ανακατασκευάζει τα διαχωρισμένα σήματα στον τομέα του χρόνου με την χρήση των μασκών που προβλέφθηκαν από το αντίστοιχο δίκτυο.



Σχήμα 4.1: Αρχιτεκτονική μοντέλου sepformer

4.1.1 Κωδικοποιητής

Ο κωδικοποιητής παίρνει την μια μείξη στο τομέα του χρόνου $x \in \mathbb{R}^T$ ως είσοδο που περιέχει πολλαπλούς ομιλητές. Μαθαίνει μια παρόμοια με Short-term-fourier-transform αναπαράσταση $h \in \mathbb{R}^{F \times T'}$ με ένα στρώμα συνέλιξης.

$$h = \text{ReLU}(\text{convld}(x)). \tag{30}$$

Με το βήμα της συνέλιξης να παίζει μεγάλο ρολό στην ακρίβεια, ταχύτητα και μνήμη του μοντέλου.

4.1.2 Δίκτυο μάσκας

Το γράφημα 4.1 δείχνει την αναλυτική αρχιτεκτονική του δικτύου μάσκας. Στο οποίο λαμβάνει τις κωδικοποιημένες αναπαραστάσεις $h \in \mathbb{R}^{F \times T'}$ και υπολογίζει την μάσκα $\{m_1, ..., m_{Ns}\}$ για καθένα από τους N_s ομιλητές. Η κωδικοποιημένη είσοδος h διοχετεύεται σε ένα στρώμα normalization (ομαλοποίηση) και επεξεργάζεται από ένα γραμμικό στρώμα όπως πρώτο-προτάθηκε στο [35]. Στη συνέχεια διασπάται το h βάσει τον άξονα του χρόνου σε κομμάτια μεγέθους C και με παράμετρο επικάλυψης (overlap factor) 50%. $h' \in \mathbb{R}^{F \times C \times Nc}$ είναι η έξοδος της πράξης τμηματοποίησης και N_c το πλήθος των τμημάτων. Η αναπαράσταση h' τροφοδοτεί το SepFormer block, που αποτελεί το βασικό κομμάτι του

δικτύου μάσκας. Αυτό το κομμάτι εμπεριέχει δυο transformers ικανούς να μάθουν μακροπρόθεσμες και βραχυπρόθεσμες εξαρτήσεις. Η έξοδος αυτού του τμήματος $h'' \in \mathbb{R}^{F\times C\times Nc}$ επεξεργάζεται από PReLU συνάρτηση ενεργοποίησης. Έπειτα ακολουθείτε από ένα γραμμικό στρώμα όπου το σηματοδοτούμε ως $h''' \in \mathbb{R}^{(F\times Ns)\times C\times Nc}$ όπου Nc είναι ο αριθμός των ομιλητών. Στη συνέχεια εφαρμόζεται η μέθοδος του overlap add (Η τεχνική περιλαμβάνει τη διαίρεση του σήματος σε επικαλυπτόμενα τμήματα, την επεξεργασία κάθε τμήματος ανεξάρτητα, και στη συνέχεια την προσθήκη των επικαλυπτόμενων τμημάτων ξανά μαζί για την ανακατασκευή του αρχικού σήματος) [36] και λαμβάνουνε το $h'''' \in \mathbb{R}^{F\times Ns\times T'}$. Έπειτα αυτή η αναπαράσταση περνάει από δυο εμπρόσθιας τροφοδοσίας (feed-forward) στρώματα και μια συνάρτηση ενεργοποίησης Relu με αυτόν τον τρόπο λαμβάνουμε την μάσκα m_k για καθένα από τους ομιλητές [44].

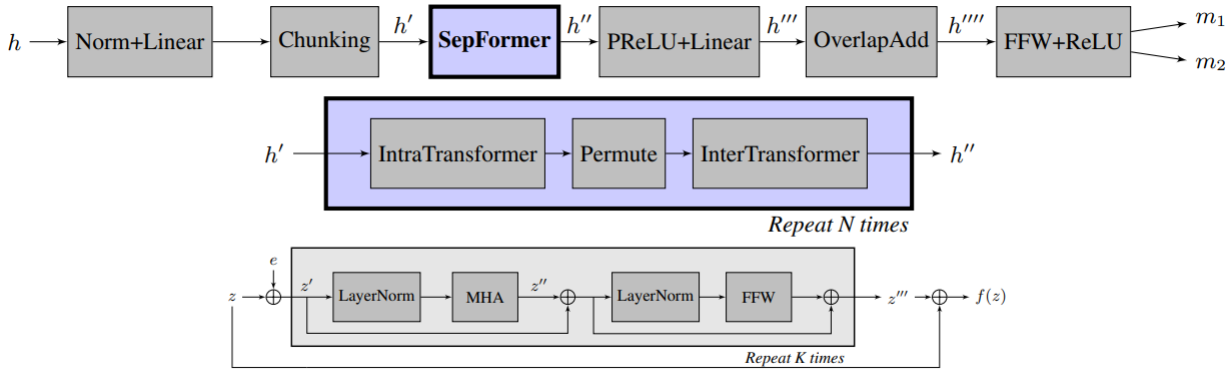
4.1.3 SepFormer μπλοκ

Το τμήμα του μπλοκ που μοντελοποιεί την βραχυπρόθεσμες εξαρτήσεις καλείται IntraTransformer (IntraT) ενώ το τμήμα που μοντελοποιεί την μακροπρόθεσμες εξαρτήσεις καλείται InterTransformer (InterT). Το IntraT κομμάτι επεξεργάζεται την δεύτερη διάσταση του h' οπότε δρα ανεξάρτητα σε κάθε κομμάτι (chunk), μοντελοποιώντας τις βραχυπρόθεσμες εξαρτήσεις μέσα σε κάθε κομμάτι. Έπειτα μετατίθενται οι τελευταίες δυο διαστάσεις (τις οποίες σηματοδοτούν με το p) και το InterT εφαρμόζεται για να μοντελοποιήσει τις μεταβάσεις μεταξύ κομματιών (Chunks). Η μετατροπή του SepFormer ορίζεται ως εξής:

$$h'' = f_{\text{inter}}(\mathcal{P}(f_{\text{intra}}(h'))) \tag{31}$$

Όπου τα IntraT,InterT σηματοδοτούνται ως $f_{\text{inter}}(\cdot)$,$f_{\text{intra}}(\cdot)$ και το SepFormer block επαναλαμβάνεται N φορές.

4.1.3.1 Intra και Inter Transformers



Σχήμα 4.2: Επιμέρους κομμάτια sepformer [44]

Επάνω η αρχιτεκτονική του δικτύου μάσκας. Στο κέντρο Το SepFormer μπλοκ. Κάτω Η αρχιτεκτονική του δικτύου των transformer $f(\cdot)$ που χρησιμοποιείται και στο IntraTransformer μπλοκ και

στο InterTransformer μπλοκ επιπλέον η ημιτονοειδής κωδικοποίηση θέσης e προστίθεται στην είσοδο ώστε [44]:

$$z' = z + e. \tag{32}$$

Η κωδικοποίηση θέσης εμπλουτίζει την είσοδο με την σειρά των στοιχείων που αποτελούν την σειρά και με αυτόν τον τρόπο επιτυγχάνεται βελτίωση στην αποτελεσματικότητα διαχωρισμού. Χρησιμοποιείται η κωδικοποίηση θέσης που πρωτοεμφανίστηκε στο [44]. Στην συνέχεια εφαρμόζονται πολλαπλά στρώματα Transformer. Μέσα σε κάθε Transformer στρώμα $g(\cdot)$ πρώτα εφαρμόζεται ένα στρώμα ομαλοποίησης και στην πορεία ένα στρώμα προσοχής πολλών κεφαλών:

$$z'' = \text{MultiHeadAtention}(\text{LayerNorm}(z')) \tag{33}$$

Όπως προτείνεται στο [44] κάθε κεφαλή προσοχής υπολογίζει την βαθμισμένη dot product (μέσω πολλαπλασιασμό διανυσμάτων) προσοχή μεταξύ όλων των στοιχείων στην ακολουθία. Στο τέλος του Transformer υπάρχει ένα εμπρόσθιας τροφοδοσίας δίκτυο (feed-forward network), το οποίο εφαρμόζεται σε κάθε θέση ανεξάρτητα:

$$z''' = \text{FeedForward}(\text{LayerNorm}(z'' + z')) + z'' + z' \tag{34}$$

Ολικά το transformer block ορίζεται ως:

$$f(z) = g^K(z + e) + z \tag{35}$$

Όπου $g^K(.)$ σηματοδοτεί K transformer στρώματά $g(\cdot)$. Χρησιμοποιούνται $K = Nintra$ στρώματα για το IntraT, και $K = Ninter$ στρώματα για το InterT. Όπως φαίνεται στο σχήμα 4.2 (κάτω) και στην εξίσωση (35) προσθέτουμε υπολειπόμενες διασυνδέσεις (residual connections) στα transformer στρώματα και στην αρχιτεκτονική transformer για καλύτερη μεταφορά των gradient (κλίσεων) μέσω backpropagation.

4.1.3.2 Αποκωδικοποιητής

Ο αποκωδικοποιητής χρησιμοποιεί ένα ανάστροφης συνέλιξης (transposed convolution) στρώμα με ίδιο βήμα συνέλιξης και μέγεθος πυρήνα με τον κωδικοποιητή. Η είσοδος του αποκωδικοποιητή είναι πολλαπλασιασμός μάσκας ομιλητή m_k επι την εξόδο του κωδικοποιητή h. Η μετατροπή του αποκωδικοποιητή μπορεί να γράφει ως εξής:

$$\hat{s}_k = \text{convld} - \text{transpose}(m_k * h), \tag{36}$$

Όπου $\hat{s}_k \in \mathbb{R}^T$ σηματοδοτεί την διαχωρισμένη πηγή k.

4.1.3.3 Training details (Ρυθμίσεις εκπαίδευσης)

Σε αυτή την ενότητα θα αναλύσουμε κάποιες τεχνικές λεπτομέρειες σχετικά με τις τιμές των παραμέτρων καθώς και τυχόν μεταβλητές που χρειάζονται fine-tuning (πειραματισμό). Ο κωδικοποιητής έχει 256 φίλτρα συνέλιξης με μέγεθος πυρήνα 16 στοιχεία και βήμα συνέλιξης 8 στοιχεία. Οι καλύτερες αποδόσεις επιτεύχθηκαν με το SepFormer δίκτυο μάσκας να λαμβάνει κομμάτια (chunks) μεγέθους $C = 250$ με $1/2$ επικάλυψη μεταξύ των κομματιών, 8 στρώματα από transformers (IntraT και InterT). Η IntraT-InterT dual-path διασωλήνωση επεξεργασίας να επαναλαμβάνετε $\mathrm{N} = 2$ φορές. Χρησιμοποιούνται 8 παράλληλες κεφαλές προσοχής καθώς και 1024-διαστάσεων feed-forward δίκτυο θέσεων μέσα σε κάθε στρώμα Transformer. Το μοντέλο έχει 26 εκατομμύρια παραμέτρους. Επιπλέον ερευνάται η χρήση dynamic mixing (DM) για εμπλουτισμό των δεδομένων [90] που περιέχει σύνθεση νέων μείξεων από πηγές με 1 ομιλητή. Σε αυτήν την τεχνική προστέθηκε επιπλέον διαταραχή ταχύτητας στις πηγές προτού ενωθούν. Η ταχύτητα αλλάζει τυχαία μεταξύ 0.95 και 1.05. Χρησιμοποιείται ο αλγόριθμος Adam [91] με βήμα βελτίωσης $15e$. Μετά από 65 εποχές εκπαίδευσης (100 με δυναμική μείξη) το βήμα βελτίωσης μειώνεται στο μισό εάν δεν υπάρξει κάποια βελτίωση στην απόδοση επιβεβαίωσης για 3 συνεχόμενες εποχές (5 για δυναμική μείξη). Το ψαλίδισμα των gradient (κλίσεων) εφαρμόζεται για να περιορίσει την $L2$ νόρμα από gradients των 5. Κατά την διάρκεια της εκπαίδευσης το μέγεθος ομάδας (batch size) είναι 1 , και για αξιολόγηση χρησιμοποιείτε η αναλογία σήματος προς θόρυβο αμετάβλητης κλίμακας (SI-SNR) [92] μέσω utterance-level permutation invariant loss [32] με ψαλίδισμα στα $30d_B$ [90]. Γίνεται χρήση αυτόματου υπολογισμού αναγκαίας ακρίβειας υπολογισμών (automatic mixed-precision) για γρηγορότερη εκπαίδευση. Το μοντέλο εκπαιδεύτηκε για 200 εποχές με κάθε εποχή να παίρνει κατά μέσο όρο 1.5 σε μια NVIDIA V100 GPU με 32 GB μνήμη [44].

4.2 Convtasnet

4.2.1 Εισαγωγή

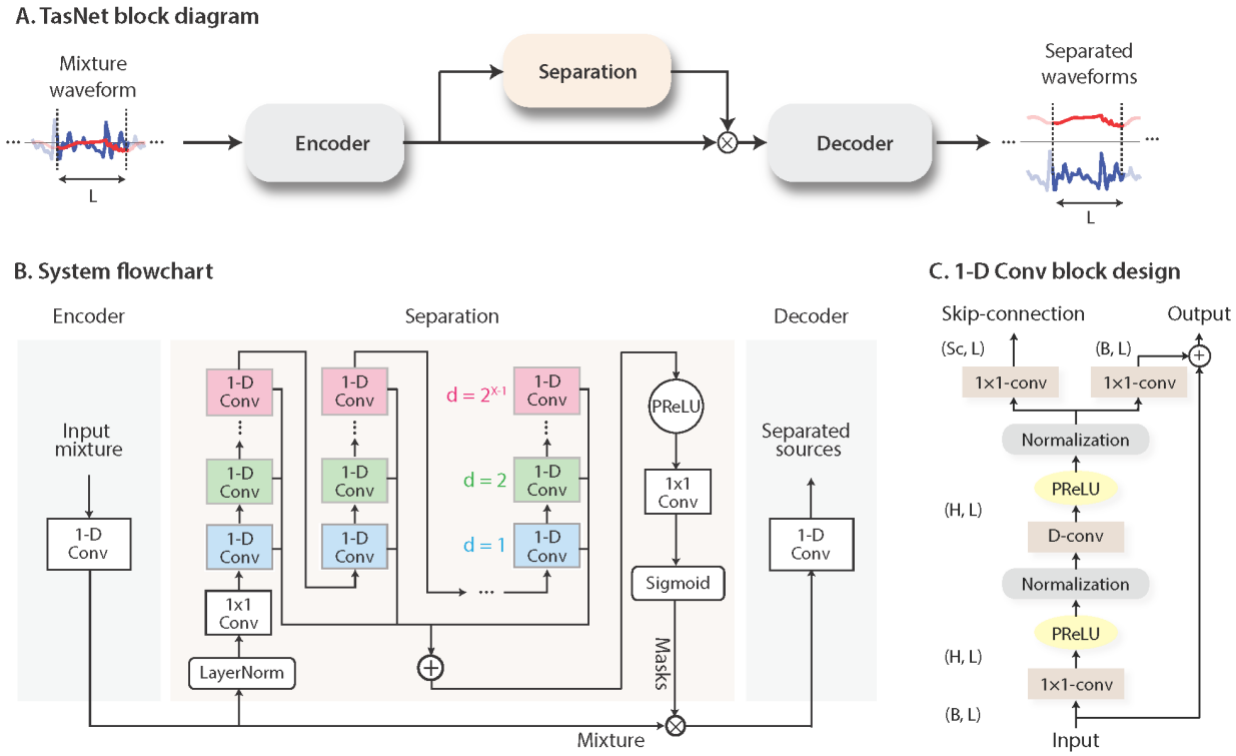
Για τον διαχωρισμό ομιλίας οι συγγραφείς προτείνουν ένα πλήρως συνελικτικό δίκτυο διαχωρισμού ήχου σε τομέα χρόνου (Conv-TasNet). Ένας γραμμικός κωδικοποιητής χρησιμοποιείται για τη δημιουργία αναπαράστασης της κυματομορφής ομιλίας βελτιστοποιημένη για το διαχωρισμό μεμονωμένων ήχων. Ένα προσωρινό συνελικτικό δίκτυο (Temporal Convolution Net) χρησιμοποιείται για την εύρεση συναρτήσεων στάθμισης (μάσκες) που θα εφαρμοστούν στην έξοδο του κωδικοποιητή. Στη συνέχεια, χρησιμοποιείται ένας γραμμικός αποκωδικοποιητής για την αναστροφή των τροποποιημένων παραστάσεων του κωδικοποιητή σε κυματομορφές [35].

Προτείνεται μια μέθοδος διαχωρισμού ομιλίας χρησιμοποιώντας έναν συνελικτικό κωδικοποιητή που χρησιμοποιεί μια λειτουργία συνέλιξης 1-D για την μετατροπή της εισόδου σε μια αναπαράσταση N-διαστάσεων. Έναν αποκωδικοποιητή που αναδομεί την κυματομορφή από αυτήν την αναπαράσταση χρησιμοποιώντας μια 1-D αντίστροφη λειτουργία συνέλιξης. Για τη δημιουργία των τελικών κυματομορφών, τα επικαλυπτόμενα ανακατασκευασμένα τμήματα προστίθενται μαζί. Ο όρος ”συνελικτικός αυτόματος κωδικοποιητής” χρησιμοποιείται επειδή, στην πράξη, τα συνελικτικά και τα αντίστροφα συνελικτικά στρώματα χρησιμοποιούνται για τον χειρισμό της επικάλυψης τμημάτων, επιτρέποντας ταχύτερη εκπαίδευση και καλύτερη σύγκλιση [35].

4.2.2 Αρχιτεκτονική Convtasnet

Για την εκτίμηση των διανυσμάτων μάσκας C για τις πηγές στόχου C, η έξοδος του προσωρινού συνελικτικού δικτύου διέρχεται μέσω ενός συνελικτικού μπλοκ 1x1. Για να μειωθεί ο αριθμός των παραμέτρων, τα συνελικτικά μπλοκ 1-διάστασης χρησιμοποιούν μια υπολειπόμενη διαδρομή (residual connection), μια διαδρομή σύνδεσης παράκαμψης (skip connection) και μια διαχωρίσιμη κατά βάθος συνέλιξη (D-conv). Για να βελτιωθεί η ακρίβεια διαχωρισμού, οι συγγραφείς προτείνουν επίσης έναν περιορισμό αθροίσματος ενότητας για τα διανύσματα μάσκας. Για να ληφθεί η αναπαράσταση κάθε πηγής, τα διανύσματα μάσκας πολλαπλασιάζονται με την έξοδο του κωδικοποιητή. Στη συνέχεια η αναπαράσταση ανακατασκευάζεται από τον αποκωδικοποιητή για να ληφθεί η κυματομορφή για κάθε πηγή [35].

Στο σχήμα 4.3 (A):Μια ξεχωριστή ενότητα υπολογίζει μια μάσκα για κάθε μια από τις πηγές-στόχους , ενώ ένας κωδικοποιητής χαρτογραφεί ένα τμήμα της κυματομορφής του μείγματος σε μια αναπαράσταση υψηλών διαστάσεων. Οι κυματομορφές πηγής ανακατασκευάζονται από έναν αποκωδικοποιητή από τα καλυμμένα χαρακτηριστικά. (B): Το διάγραμμα ροής του προτεινόμενου συστήματος. Οι κυματομορφές αντιπροσωπεύονται από έναν 1-D συνελικτικό κωδικοποιητή και ένα προσωρινό συνελικτικό δίκτυο. Οι μάσκες εκτιμώνται από τη μονάδα διαχωρισμού TCN με βάση την έξοδο του κωδικοποιητή. Τα διαφορετικά χρώματα στα 1-D συνελικτικά μπλοκ του TCN αντιπροσωπεύουν διάφορους παράγοντες διαστολής. (Γ): Η κατασκευή ενός 1-D συνελικτικού μπλοκ. Κάθε μπλοκ αποτελείται από μια λειτουργία μετατροπών που ακολουθείται από μια λειτουργία Dconv, με μια μη γραμμική συνάρτηση ενεργοποίησης και κανονικοποίηση που προστίθεται μεταξύ κάθε δύο συνελίξεις. Δύο γραμμικά μπλοκ 1×1-conv χρησιμεύουν ως υπολειπόμενες συνδέσεις και ως συνδέσεις παράκαμψης [35].



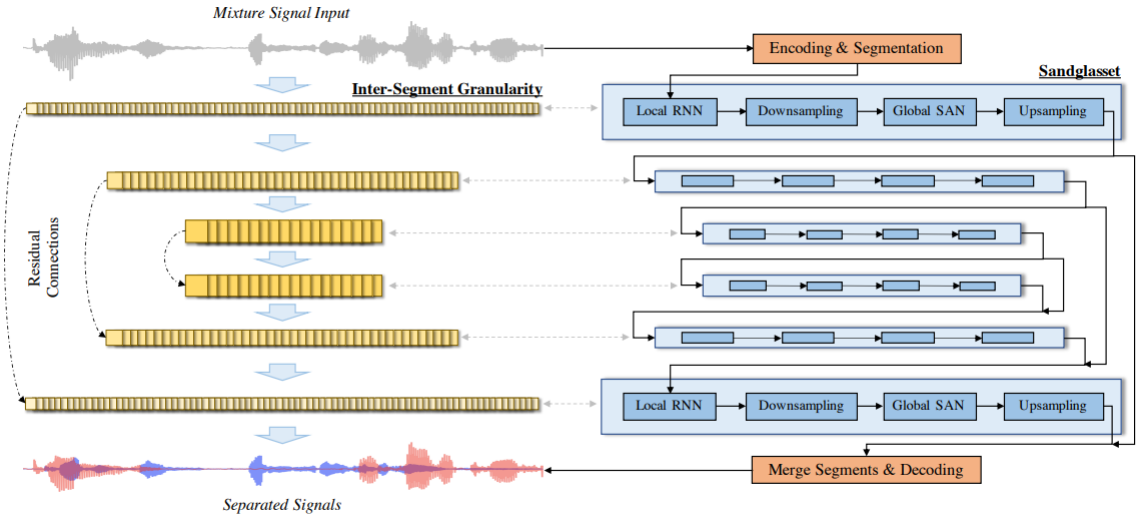
Σχήμα 4.3: Επιμέρους τμήματα Convtasnet [35]

4.2.3 Αποδόσεις

Δοκιμάστηκε το σύστημα με έναν επεξεργαστή σε μια CPU Intel Core i7-5820K για τη διαμόρφωση της CPU. Όσον αφορά την GPU είναι μια GPU Nvidia Titan Xp. Σε εφαρμογές όπου είναι διαθέσιμη μόνο μια πιο αργή CPU, το LSTM-TasNet με διαμόρφωση CPU έχει ένα TPF (time per frame) κοντά στο μήκος του πλαισίου του (5 ms), το οποίο είναι οριακά αποδεκτό. Επιπλέον, επειδή η επεξεργασία στο LSTM-TasNet γίνεται διαδοχικά, κάθε χρονικό πλαίσιο πρέπει να περιμένει την ολοκλήρωση του προηγούμενου χρονικού πλαισίου, αυξάνοντας τον συνολικό χρόνο επεξεργασίας ολόκληρης της διαδικασίας. Επειδή το Conv-TasNet αποσυνδέει την επεξεργασία των διαδοχικών καρέ, τα επόμενα καρέ δεν χρειάζεται να περιμένουν να ολοκληρωθεί το τρέχον πλαίσιο. Επειδή το Conv-TasNet απομονώνει τον υπολογισμό των διαδοχικών πλαισίων, τα επόμενα καρέ δεν χρειάζεται να περιμένουν να ολοκληρωθεί το τρέχον πλαίσιο, επιτρέποντας παράλληλους υπολογισμούς. Αυτή η διαδικασία έχει ως αποτέλεσμα ένα TPF που είναι πέντε φορές μικρότερο από το μήκος του πλαισίου (2 ms) στη διαμόρφωση της CPU μας. Ως αποτέλεσμα, το Conv-TasNet μπορεί ακόμα να πραγματοποιεί διαχωρισμό σε πραγματικό χρόνο σε πιο αργές CPU [35].

4.3 Sandglasset

4.3.1 Αρχιτεκτονική Sandglasset



Σχήμα 4.4: Επιμέρους κομμάτια sandglasset

Το προτεινόμενο μοντέλο Sandglasset [48] αποτελείται από N μπλοκς όπως φαίνεται στο γράφημα στο 4.4 στα δεξιά. Εάν η ροή της πληροφορίας είναι από πάνω προς τα κάτω τα πρώτα $N/2$ μπλοκς αποτελούν μια ανάστροφη πυραμίδα όπου τα καρέ σημάτων λαμβάνουν μέρος σε μια υποδειγματοληψία. Αυτό συνεπάγεται μεγάλη κοκκοποίηση καθώς και ανωτέρου επίπεδου αφηρημένα χαρακτηριστικά. Τα τελευταία $N/2$ μπλοκς αποτελούν μια κανονικής ροής πυραμίδα όπου σε αυτά τα ανωτέρου επίπεδου χαρακτηριστικά εφαρμόζεται υπερδειγματοληψία (upsampling). Αυτό κάνει την δομή τους να επιστρέψει πίσω σε μεγαλύτερες ακολουθίες χαρακτηριστικών από μικρότερα κομμάτια σε λεπτομερές χρονικές κλίμακες, δηλαδή λεπτή κοκκοποίηση, σε κατωτέρου επίπεδου χαρακτηριστικά. Για την διατήρηση της πληροφορίας τα δείγματα που υπέστησαν υπερδειγματοληψια δηλαδή τα τελευταία $N/2$ μπλοκς προστίθενται με τα προηγουμένως υπολογισμένα χαρακτηριστικά με την ίδια κοκκοποίηση χρησιμοποιώντας υπολειπόμενες συνδέσεις (residual connections). Αυτή η επεξεργασία είναι χρήσιμη για καλύτερη ανακατασκευή σήματος ενώ παράλληλα αποφεύγουμε την εξάλειψη των gradient (κλίσεων διόρθωσης). Αυτό το κλεψυδροειδές σχήμα είναι ικανό να μοντελοποιεί πολλαπλές βαθμίδες χρονικής κοκκοποίησης για την επεξεργασία του σήματος εισόδου ιεραρχικά.

4.3.2 Κωδικοποίησή και τμηματοποίηση

4.3.2.1 Κωδικοποιητής TasNet

Αρχικά το σήμα εισόδου είναι μια κυματομορφή στον τομέα του χρόνου $x \in R^T$ βασισμένο στο Tasnet [33, 36]. Η μείξη εισόδου κωδικοποιείται σε μια ακολουθία με $1/2$ επικαλυπτόμενα καρέ σηματοδοτούμενα από $\tilde{X} = [\tilde{x_1}, ..., \tilde{x_L}] \in R^{M \times L}$. Όπου M είναι μια υπερπαράμετρος που αναφέρεται ως μέγεθος παραθύρου $L = \lceil 2T/M \rceil$. Στο Tasnet χρησιμοποιείται ένα στρώμα με συνέλιξη πύλης μιας διάστασης και συνάρτηση ενεργοποίησης RELU. Αυτο γίνεται για να αντικαταστήσει την παραδοσιακή μέθοδο short-term Fourier transform (STFT) για κωδικοποίηση σημάτων:

$$\hat{\mathbf{X}} = \text{ReLU}\left(\text{ConvlD}\left(\tilde{\mathbf{X}}; \mathbf{U}\right)\right) \tag{37}$$

Όπου ConvlD $\left(\tilde{\mathbf{X}}; \mathbf{U}\right)$ είναι η μονοδιάστατη συνέλιξη που εφαρμόζεται στο $\tilde{\mathbf{X}}$ που παραμετροποιείται από ένα εκπαιδεύσιμο βάρος $U \in R^{E \times M}$ με 1×1 πυρήνα. Το $ReLU\cdot$ είναι η διορθωμένη γραμμική μονάδα που χρησιμοποιείται στα [36,37] για την βεβαίωση ότι δεν θα υπάρχουν αρνητικά αποτελέσματα. Το E είναι οι διαστάσεις από το κάθε κωδικοποιημένο καρέ. Άντι για την απευθείας χρήση του $\tilde{\mathbf{X}}$ για την μεταγενέστερη επεξεργασία, αντιστοιχίζεται γραμμικά ο πίνακας σε bottleneck (σμίκρυνση) $X = \mathbf{B}\tilde{\mathbf{X}} \in R^{D \times L}$, όπου $\mathbf{B} \in R^{D \times E}$ και $D < E$.

4.3.2.2 Τμηματοποίηση

Δεδομένης μιας ακολουθίας από καρέ στον πίνακα με μορφή $X \in R^{D \times L}$, χρησιμοποιείται μια ενότητα τμηματοποίησης για διαχωρισμό του X σε $S = 1/2$ επικαλυπτόμενα τμήματα. Το καθένα μεγέθους K .Το πρώτο και το τελευταίο τμήμα είναι γεμάτα με μηδενικά για την δημιουργία $S = \lceil 2L/K \rceil$ ίσου μεγέθους τμημάτων. Αυτά τα τμήματα στην συνέχεια μπορούν να συναχθούν μαζί σε έναν τρισδιάστατο τένσορα που σηματοδοτείται από $\mathcal{X} \in R^{D \times K \times S}$. Πρέπει να σημειωθεί ότι το μέγεθος των τμημάτων είναι υπερπαράμετρος που χρησιμοποιείται για τον έλεγχο της διαβάθμισης της τοπικότητας. Αυτά τα τμήματα $\mathcal{X}$ μεταβιβάζονται σε μια στοίβα από Sandglasset μπλοκς.

4.3.3 Sandglasset μπλοκς

Για το ν-οστό μπλοκ, δίνεται ένας τρισδιάστατος τένσορας εισόδου $\mathcal{X}_b \in R^{D \times K \times S}$, που θα εμπεριέχει S τμήματα καθένα από K καρέ των D διαστάσεων. Για να γίνουν οι ακόλουθες επαναληπτικές μαθηματικές σχέσεις κατανοητές ορίζεται το $\mathcal{X}_1 = \mathcal{X}$.Καθένα από τα Sandglasset μπλοκ περιέχει 2 κύρια τμήματα. Πρώτα επεξεργάζεται το intra-τμήμα ακολουθίας χρησιμοποιώντας ένα αναδρομικό νευρωνικό δίκτυο (RNN) για να μοντελοποιηθεί η τοπικότητα όπως στο [36], και έπειτα να μοντελοποιηθεί το inter-τμήμα ακολουθίας χρησιμοποιώντας SAN (ένα δίκτυο αυτοπροσοχής με επίγνωση σε μεταβλητό πλαίσιο) για την καταγραφή των μακροχρόνιων εξαρτήσεων. Μια υποδειγματοληψια (downsampling) και μια υπερδειγματοληψια (upsampling) αλλάζουν την κοκκοποίηση για την μακροπρόθεσμη ακολουθία που θα επεξεργαστεί από το SAN (ένα δίκτυο αυτοπροσοχής με επίγνωση σε μεταβλητό πλαίσιο).

4.3.3.1 Αναδρομικά νευρωνικά δίκτυά για τοπική επεξεργασία ακολουθίας

Τα intra-τμήματα ακολουθιών είναι οι τοπικές ακολουθίες με μέγεθος K , που περιέχουν μικρές τοπικές λεπτομέρειες ,χρονική ή φασματική συνέχεια, δομή φάσματος, χροιά, κλπ. Πράγματα τα οποία είναι άχρηστα σε μακροχρόνιο πλαίσιο. Το πρόβλημα των τοπικών ακολουθιών επεξεργάζεται ένα στρώμα απο αναδρομικά νευρωνικά δίκτυα. Ειδικότερα σε κάθε Sandglasset μπλοκ ο τρισδιάστατος τένσορας $\mathcal{X}_b^{LR} = \mathcal{X}$ που λαμβάνεται από το κομμάτι της τμηματοποίησης μεταβιβάζεται σε ένα bi-directional LSTM με H κρυμμένους νευρώνες . Για ευκολία αναφοράς, χρησιμοποιούμε $\mathcal{X}_b^{LR}, \mathcal{Y}_b^{LR}$ για να σηματοδοτήσουμε τις εισόδους, εξόδους στο τοπικό αναδρομικό νευρωνικό δίκτυο. Ο εκθέτης LR χρησιμοποιείται για να διαφοροποιήσει από τις εισόδους-εξόδους από το μοντέλο μακροχρόνιων εξαρτήσεων.

$$y_b^{LR} = [\mathrm{M}_b \cdot \mathrm{BiLSTM}_b\left(\chi_b^{LR}[\ :, s,\ :]\right) + \mathbf{c}_b, s = 1, ..., S) \tag{38}$$

Όπου $\cdot$ είναι ο πολλαπλασιασμός πινάκων $\chi_b^{LR}[\ :, s, :\ ] \in R^{D\times K}$ αναφέρεται στη τοπική ακολουθία μέσα στο s-οστό τμήμα (chunk), $\mathbf{M}_b \in R^{D\times 2H}$ $\kappa\alpha\iota$ $\mathbf{c}_b \in R^D$ είναι οι παράμετροι της γραμμικής μεταμορφώσεις.

4.3.3.2 Αυτοπροσεχούμενο δίκτυο για πολυκοκκοποίηση

Αφού επεξεργάζονται τα intra-τμήματα των ακολουθιών το καθένα μεγέθους K , στοχεύουν στην μοντελοποίηση των inter-τμήματα ακολουθιών το καθένα με μέγεθος S. Σημειωτέο ότι τα inter-τμήματα ακολουθιών είναι πιθανό να κωδικοποιήσουν το πλαίσιο της πληροφορίας μέσα στο σήμα ομιλίας. Στο Sandglasset εφαρμόζεται ένα δίκτυο αυτοπροσοχής με επίγνωση σε μεταβλητό πλαίσιο (variable-context-aware self-attentive network (SAN)) για την αποθήκευση των μακροχρονίων εξαρτήσεων σε διαφορετικές χρονικές κλίμακες. Άντι για την απευθείας χρήση y_b^{LR} ως είσοδο στο SAN πρώτα εφαρμόζεται ένα στρώμα ομαλοποίησης (normalization) $LN(\cdot)$ στην έξοδο LR στρώματος και προστίθενται υπολειπόμενες συνδέσεις στο μπλοκ εισόδου:

$$\chi_b^{GA} = \mathrm{LN}\left(\ y_b^{LR}\right) + \chi_b, \tag{39}$$

Τα οποία resampling (ξανά δειγματολείπτονται) για να επεξεργαστούν την χρονική κλίμακα για μακροχρόνια επεξεργασία κατά μήκος όλων των τμημάτων:

$$y_b^{GA} = \mathrm{US}_b\left(\mathrm{SAN}_b\left(\mathrm{DS}_b\left(\mathcal{X}_b^{GA}\right)\right)\right) \tag{40}$$

Όπου $US_b(\cdot), DS_b(\cdot)$ είναι η υπερδειγματοληψία και η υποδειγματοληψία , που ορίζονται ως εξής:

$$\mathrm{US}_b\left(\chi\right) = \begin{cases} \mathrm{ConvTranslD}_K\left(\chi; 4^b\right) & \text{if} \quad \mathrm{b} \leq \mathrm{N/2}; \\ \mathrm{ConvTranslD}_K\left(\chi; 4^{N-b-1}\right) & \text{if} \quad \mathrm{b} > \mathrm{N/2}; \end{cases} \tag{41}$$

$$\mathrm{DS}_b\left(\chi\right) = \begin{cases} \mathrm{ConvTranslD}_K\left(\chi; 4^b\right) & \text{if} \quad \mathrm{b} \leq \mathrm{N/2}; \\ \mathrm{ConvTranslD}_K\left(\chi; 4^{N-b-1}\right) & \text{if} \quad \mathrm{b} > \mathrm{N/2}; \end{cases} \tag{42}$$

Όπου $Conv1DA(\cdot; B)$ και $ConvTran1DA(\cdot; B)$ σηματοδοτούν την μονοδιάστατη και την αντίστροφη μονοδιάστατη συνέλιξη ως προς τον άξονα A. Το μέγεθος πυρήνα B έτσι ώστε το προκύπτουν μέγεθος να είναι $\lfloor A/B \rfloor$ (στην υποδειγματοληψία) ή $\lfloor AB \rfloor$ (στην υπερδειγματοληψία). $SANb(\cdot)$ είναι το δίκτυο αυτοπροσοχής με επίγνωση σε μεταβλητό πλαίσιο που έχει τροποποιηθεί από το [88] γενικότερα το δίκτυο SAN ορίζεται για κάθε είσοδο $\mathcal{X} \in \mathcal{R}^{D\times S\times K}$ ως :

$$SAN(X) = \left[SelfAttn\left(LN\left(\mathcal{X}[:, :, k]\right) + \mathbf{P}\right), k = 1, \ldots, K\right], \tag{43}$$

Όπου P είναι ο πίνακας κωδικοποίησης θέσης όπως εμφανίζεται στο [88] $D \times S$ να αναφέρεται στο intertμήμα της ακολουθίας. Το $SelfAttn(\cdot)$ είναι ένα τυπικό πολλαπλών κεφαλών αυτοπροσοχής δίκτυο που γραμμικά προβάλλει έναν πίνακα εισόδου $\mathbf{X} \in R^{D \times S}$ σε 3 μορφές από πίνακες. Με αντίστοιχες σημάνσεις ως $query = Q_j, key = K_j, value = V_j$ για τον υπολογισμό της διαβαθμιζόμενης dot-product (πολλαπλασιασμό διανυσμάτων) προσοχής. Αυτό συμβαίνει για όλες τις διαφορετικές κεφαλές $j = 1, ..., J$ που στο τέλος ενώνονται με συνένωση και έναν πολλαπλασιασμό πινάκων

$$[\mathbf{Q}_j \ \mathbf{K}_j \ \mathbf{A}_j]^{\mathsf{T}} = \left[\mathbf{W}_j^{\mathsf{Q}} \ \mathbf{W}_j^{\mathsf{K}} \ \mathbf{W}_j^{\mathsf{V}}\right]^{\mathsf{T}} \mathbf{X} + \left[\mathbf{b}_j^{\mathsf{Q}} \ \mathbf{b}_j^{\mathsf{K}} \ \mathbf{b}_j^{\mathsf{V}}\right]^{\mathsf{T}} \tag{44}$$

$$\mathbf{A}_j = \text{Sofmax}\left(\frac{\mathbf{Q}_j^{\mathsf{T}} \mathbf{K}_j}{\sqrt{D/J}}\right) \mathbf{V}_j \tag{45}$$

$$\mathbf{A} = \mathbf{W} \cdot \text{Concat}(\mathbf{A}_1, ..., \mathbf{A}_J) \tag{46}$$

$$SelfAttn(\mathbf{X}) = LN(\mathbf{X} + DROP(\mathbf{A}) \tag{47}$$

όπου $DROP(\cdot)$ σηματοδοτεί την τεχνική [93], και $\mathbf{W} \in \mathbb{R}^{D \times D}$, $\mathbf{W}_j^{\mathbf{Q}}$, $\mathbf{W}_j^{\mathbf{K}}$, $\mathbf{W}_j^{\mathbf{V}} \in \mathbb{R}^{D/J \times D}$ και $\mathbf{b}_j^{\mathrm{Q}}, \mathbf{b}_j^{\mathrm{K}}, \mathbf{b}_j^{\mathrm{V}} \in \mathbb{R}^{D/J}$ είναι οι παραμέτροι του SAN δικτύου.

4.3.3.3 Υπολειμματικές συνδέσεις για την πρόληψη πληροφοριών

Ένα από τα στιγμιότυπα του Sandglasset είναι η εισαγωγή υπολειπόμενων συνδέσεων μεταξύ ζευγαριών από Sandglasset μπλοκς που έχουν την ίδια κοκκοποίηση. Αυτή η τεχνική χρησιμοποιείται για να αποτρέψει την απώλεια πληροφορίας αφού μεταβιβαστεί από τα κεντρικά μπλοκς όπου η κοκκοποίηση είναι πιο τραχεία και ορίζεται μαθηματικά ως:

$$\mathcal{X}_{b+1}^{LR} = \begin{cases} y_b^{GA} & \text{if} \quad \text{b} \leq \text{N/2;} \\ y_b^{GA} + y_{b-N/2}^{GA} & \text{if} \quad \text{b} > \text{N/2;} \end{cases} \tag{48}$$

Το οποίο ορίζει την επαναληπτική σχέση μεταξύ του b-οστό και το b+1-οστού Sandglasset μπλοκ. Έχει αποδειχτεί πειραματικά ότι οι υπολειπόμενες συνδέσεις είναι ζωτικής σημασίας για διόρθωση ακατέργαστων λεπτομερειών επιπέδου σήματος για βελτίωσή της ανακατασκευής του σήματος και την αποφυγή της εξάλειψης των gradients. Μια παρόμοια δουλειά ως προς την επανα-δειγματοληψία είναι το U-Net [94, 95]που επιπλέον συνδυάζει τα χαρακτηριστικά σε διαφορετικές χρονικές κλίμακες. Ωστόσο υπάρχουν κάποιες κύριες διαφορές: (1) Η υποδειγματοληψία και υπερδειγματοληψία γίνονται στο ίδιο μπλοκ (2) Τα χαρακτηριστικά πολυκοκκοποίησης επεξεργάζονται από τα SAN μέσα σε κάθε μπλοκ (3) Οι υπολειπόμενες συνδέσεις γίνονται μέσω πρόσθεσης αντί για συνένωσης.

4.3.4 Ένωση τμημάτων και αποκωδικοποίηση

4.3.4.1 Εκτίμηση μάσκας

Αφού μεταβιβαστούν μεταξύ N Sandglass μπλοκς ,λαμβάνουμε έναν τρισδιάστατο τένσορα $\mathcal{X}_{N+1}^{LR} \in \mathbb{R}^{D\times K\times S}$, Πράγμα που μπορεί να εκτιμήσει μάσκες για C ομιλητές. Πρώτα πραγματοποιείται μια μεταμόρφωση στην έξοδο του τελευταίου μπλοκ ,χρησιμοποιώντας ένα PReLU δισδιάστατο στρώμα πύλης (PReLU-gated 2D convolutional layer) για να λάβουν ένα τένσορα τεσσάρων διαστάσεων με σχήμα $C \times E \times K \times S$:

$$\mathcal{V} = \text{Conv2D}\left(\text{PReLU}\left(\mathcal{X}N + 1^{LR}\right); \mathbf{C}\right), \tag{49}$$

Όπου το Conv2D($\mathcal{Y}$; $\mathbf{C}$) σηματοδοτεί την δισδιάστατη συνέλεξη που εφαρμόζεται στο Υ που παραμετροποιείται από το βάρος $\mathbf{C} \in \mathbb{R}^{CE\times D}$ με 1×1 πυρήνα και $PReLU(\cdot)$. Μετά ενώνονται τα τμήματα εξόδου $\mathcal{Y}$ με την τη διαίρεση του σήματος σε επικαλυπτόμενα τμήματα, την επεξεργασία κάθε τμήματος ανεξάρτητα και, στη συνέχεια, την προσθήκη των επικαλυπτόμενων τμημάτων μαζί για την ανακατασκευή του αρχικού σήματος όπως στο [36]. Αυτό γίνεται για να ταιριάξει στο σχήμα των καρέ των μείξεων $\hat{\mathbf{X}} \in \mathbb{R}^{E\times L}$ ενώ για τις μάσκες ισχύει:

$$\text{M} = \text{ReLU}\left(\text{OverlapAdd}\left(\text{Y}\right)\right) \tag{50}$$

Όπου το $\odot$ είναι ο πολλαπλασιασμός στοιχείο προς στοιχείο.

4.3.4.2 Αποκωδικοποιητής για ανακατασκευή κυματομορφής

Τέλος το c-οστό σήμα πηγής ανακατασκευάζεται με την εφαρμογή της c-οστής εκτιμώμενης μάσκας στα αρχικά υπολογισμένα καρέ μίξης $\tilde{\mathbf{X}}$ και μετά την χρήση της OverlapAdd για την μετατροπή των καρέ σε κυματομορφή.

$$\hat{\mathbf{s}}_c = \text{OverlapAdd}(\tilde{\mathbf{X}} \odot \text{M}_c). \tag{51}$$

Τέλος δεδομένου ότι C είναι οι εκτιμώμενες πήγες, το scale-invariant source-to-noise ratio (SI-SNR) loss [92] χρησιμοποιείται με u-PIT [31]για να μάθει τις παραμέτρους του δικτύου και την επίλυση του προβλήματος μετάθεσης.

4.3.5 Ρυθμίσεις εκπαίδευσης

4.3.5.1 Παράμετροι

Η Σύνθεση κωδικοποιητή-αποκωδικοποιητή των [33,35] χρησιμοποιείται και το τμήμα τμηματοποίησης που περιγράφεται στο [36]. Με $\text{M} = 4, \text{E} = 256, D = 128$.Γίνεται η χρήση $\text{N} = 6$ Sandglasset μπλοκς, όπου στο πρώτο μπλοκ το αρχικό μέγεθος του τμήματος είναι $\kappa = 256$ που θα μειωθεί/επεκταθεί κατά 4 στα πρώτα/τελευταία μπλοκς όπως περιγράφεται στα (41-42). Μέσα σε κάθε sandglasset

μπλοκ χρησιμοποιείται ένα Bi-LSTM με 128 κρυφούς νευρώνες το global SAN (ένα δίκτυο αυτοπροσοχής με επίγνωση σε μεταβλητό πλαίσιο) είναι με $J = 8$ κεφαλές προσοχής με 0.1 ρυθμό dropout [93]. Για εκπαίδευση επιλέχθηκε ο αλγόριθμος adam [91] με αρχικό βήμα εκμάθησης $= 0.001$ και με ρυθμό μειώσεις στο 0.98. Η εκπαίδευση σταματά όταν δεν υπάρχει βελτίωση για 10 συνεχόμενες εποχές.

4.3.5.2 Βελτίωση με δυναμική ανακατασκευή δεδομένων

Με περαιτέρω ανάλυση περιπτώσεων κακού διαχωρίσμού φάνηκε το μοντέλο να έχει μια αδυναμία διαχωρισμού ομιλητών με παρόμοια φωνητική χροιά. Για την αντιμετώπιση αυτού του θέματος αφού έχει ήδη προκληθεί σύγκλιση στην εκπαίδευση επεκτείνεται η εκπαίδευση με δυναμική ανάμειξη του ίδιου ομιλητή με 1:1 αναλογία στο σετ δεδομένων,

4.4 The hungarian PIT

4.4.1 Εισαγωγή

Αυτή η προσέγγιση [46] είναι η πιο κοντινή σε πραγματικές συνθήκες καθώς μπορεί να διαχειριστεί περισσοτέρους ομιλητές. Με αυτή την μέθοδο οι αρθρογράφοι ισχυρίζονται ότι μπορούν να διαχωρίσουν έως και 20 ομιλητές από 1 μικρόφωνο. Χτίζοντας πάνω στην έρευνά του με το PIT σφάλμα [31] και την χρήση του αλγορίθμου που προτάθηκε το 1957 από τον James Munkres [45] γνωστό και ως Hungarian Algorithm. Οι ερευνητές καπηλεύονται την οικονομική φύση του αλγορίθμου για να βρουν την βέλτιστη μετάθεση και η σύγκλιση αυτή εξαρτάται από της λιγότερες επαναλήψεις του αλγορίθμου.

4.4.2 Αρχιτεκτονική μοντέλου Hungarian PIT

4.4.2.1 Hungarian Loss (Σφάλμα βασισμένο στον αλγόριθμο Hungarian)

Ένα δίκτυο διαχωρισμού ομιλίας με ένα κανάλι ομιλίας δέχεται ένα ηχητικό σήμα που περιέχει μια μίξη από C ομιλητές και C σήματα εξόδου. Κατά την διάρκεια της εκπαίδευσης το δίκτυο εξάγει τα διαχωρισμένα ηχητικά σήματα με αυθαίρετη σειρά. Έτσι για τον υπολογισμό του σφάλματος μια ευθυγράμμιση δηλαδή μια μετάθεση πρέπει να ανακτηθεί μεταξύ της εξόδου του δικτύου και των πραγματικών σημάτων. Ένας τρόπος για να βρεθεί η σωστή μετάθεση είναι ο έλεγχος όλων των πιθανών μεταθέσεων $C!$ και να επιλεχθεί η μετάθεση με το μικρότερο μέσο σφάλμα στο ζευγάρι (PIT) [31]. Το υπολογιστικό κόστος του PIT είναι αμελητέο όταν το C είναι μικρό ωστόσο κάνει την εκπαίδευση σε μεγάλο αριθμό ομιλητών αδύνατη. Για παράδειγμα για 20 ομιλητές χρειάζεται ο έλεγχος $20! \approx 2.4 \times 1018$ μεταθέσεων. Η προσέγγιση των ερευνητών είναι η θεώρηση του προβλήματος ως πρόβλημα ανάθεσης γραμμικού αθροίσματος. Εάν έχουμε C σήματα εξόδου και C πραγματικά σήματα υπολογίζεται το σφάλμα κατά ζεύγη (si, sj) σε κάθε ζευγάρι με έξοδο (sj) και στόχο (si) όπου έχει ως αποτέλεσμα έναν πίνακα σφαλμάτων $\mathrm{M} = C * C$. Έπειτα κάθε έξοδος αναθέτετε με ένα μοναδικό στόχο και αντίστροφα. Μια βέλτιστη ανάθεση ελαχιστοποιήσει το άθροισμά των τιμών των επιλεγμένων στοιχείων. Το PIT σφάλμα μπορεί να θεωρηθεί έως μια προσπάθεια τυχαίων δοκίμων που ελέγχει όλες τις πιθανές λύσεις:

$$\ell(s, \hat{s}) = \min_{\pi \in \Pi_C} \frac{1}{C} \sum_{i=1}^{C} \hat{\ell}(s_i, \hat{s}_{\pi(i)}) \tag{52}$$

Εφαρμόζοντας τον Hungarian αλγόριθμο στον πίνακα M βρίσκουμε την ιδανική μετάθεση σε πολυωνυμικό χρόνο αντί για τυχαίες δοκιμές. Εδώ πρέπει να σημειωθεί ότι η διαδικασία εύρεσης μεταθέσεων μπορεί να γίνει διαχωριστά από τον ανάστροφο υπολογισμό των gradients (διορθώσεων).

4.4.2.2 Trainnig details (Ρυθμίσεις εκπαίδευσης)

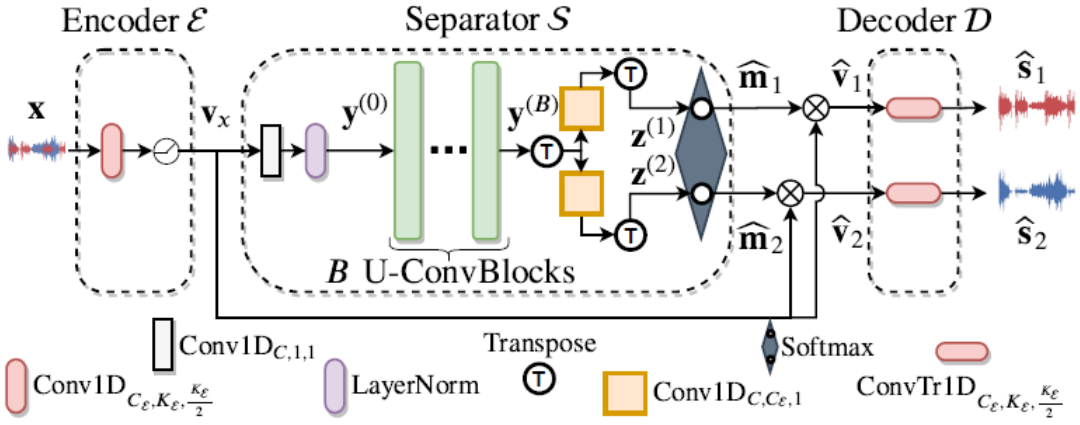
30 ώρες από ομιλίες χρησιμοποιήθηκαν για την δημιουργία του πακέτου εκπαίδευσης και του πακέτου επιβεβαίωσης. Οι 5 ομιλητές επιλέχθηκαν τυχαία με Signal to noise ratio μεταξύ 0-5 dB. Το πακέτο επιβεβαίωσης δημιουργήθηκε από si_et_s και si_dt_s με 16 ομιλητές, που δεν βρίσκονται στο πακέτο εκπαίδευσής. Η εκπαίδευση έγινε με Adam optimizer [91] και μέγεθος πακέτου (batch size) 32 με βήμα εκμάθησης $1e-3$ που πολλαπλασιάζεται με 0.95 κάθε 2 εποχές και κάθε σήμα διασπάται σε 4 σημεία [46].

4.5 SuDORMRF

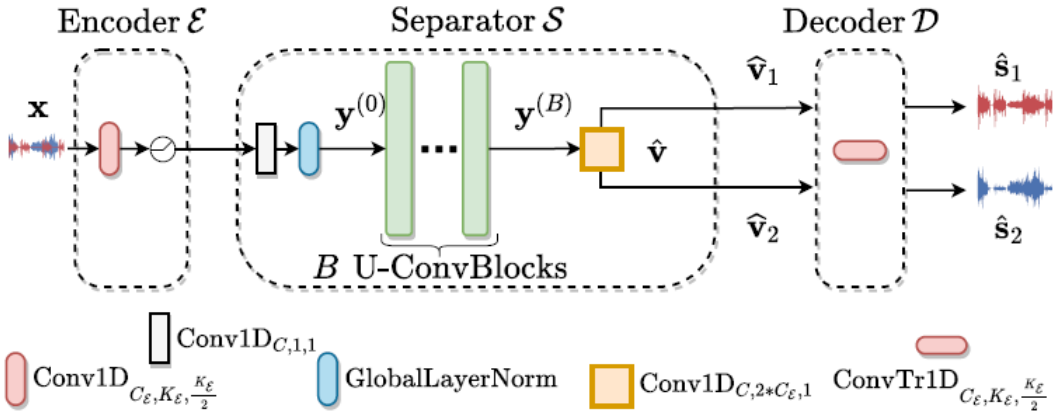
4.5.1 Εισαγωγή SuDoRM-RF

Η προτεινόμενη μέθοδος [96, 97] πραγματοποιεί επαναληπτικές υποδειγματολειψίες και δειγματοληψίες πολλαπλών αναλύσεων χαρακτηριστικών (SuDoRM-RF) με την χρήση συνέλιξης κατά βάθος (depth-wise convolutions). Με αυτόν τον τρόπο η μέθοδος εκμεταλλεύεται την αποδοτικότητα των επαναλαμβανόμενων χρονικών δειγματοληψιών [98] και αποφεύγει την ανάγκη πολλαπλών στοιβαγμένων διασταλμένων στρωμάτων συνέλιξης.

4.5.2 Αρχιτεκτονική SuDoRM-RF - SuDORMRFImprovednet



(a) SuDoRM-RF architecture.



(b) SuDoRM-RF++ architecture.

Σχήμα 4.5: Στο γράφημα παρουσιάζονται και οι δύο αρχιτεκτονικές (SuDoRM-RF - SuDORMRFImprovednet) [96]

Η είσοδος είναι ένα ανεπεξέργαστο σήμα από μια μείξη $\mathbf{x} \in \mathbb{R}^T$ με T δείγματα στο τομέα του χρόνου. Πρώτα προωθείται η είσοδος x σε έναν κωδικοποιητή $\mathcal{E}$ με σκοπό να ληφθεί μια λανθάνουσα αναπαράσταση για την μείξη $v_x = \mathcal{E}(x) \in \mathbb{R}^{C_\mathcal{E} \times L}$. Επιπλέον η λανθάνουσα αυτή αναπαράσταση προωθείται σε ένα τμήμα διαχωρισμού S που υπολογίζει τις αντίστοιχες μάσκες $\hat{m}_i \in \mathbb{R}^{C_\mathcal{E} \times L}$ για καθένα από τους ομιλητές N $s_1 \ldots s_N \in \mathbb{R}^T$ που αποτελούν την μείξη. Η εκτιμώμενη λανθάνουσα αναπαράσταση για κάθε πηγή είναι ο λανθάνων χώρος $\hat{v}_i$ που λαμβάνεται με τον πολλαπλασιασμό βάση μιας εκτιμώμενης μάσκας $\hat{m}_i$ και των κωδικοποιημένων αναπαραστάσεων της μείξης v_x. Τέλος η ανακατασκευή για

κάθε πηγή $\hat{s_i}$ λαμβάνεται με την χρήση ενός αποκωδικοποιητή $\mathcal{D}$ για την μετάθεση του λανθάνων χώρου $\hat{v_i}$ πίσω στο τομέα του χρόνου $\hat{s_i} = \mathcal{D}(\hat{v_i})$ [96].

Definition (1) $\text{Conv1D}_{C.K.S} : \mathbb{R}^{C_i n \times L_i n} \rightarrow \mathbb{R}^{C \times L}$ *ορίζει ένα πυρήνα (kernel)* $\boldsymbol{W} \in \mathbb{R}^{C \times C_{i\,n} \times K}$ *και ένα βάρος προκατάληψης (bias vector)* $\mathbf{b} \in \mathbb{R}^{\mathbb{C}}$. *Όταν εφαρμόζεται σε μια είσοδο* $\mathbf{x} \in \mathbb{R}^{\mathbf{C_{in}} \times \mathbf{L_{in}}}$ *πραγματοποιεί μια συνέλιξη μιας διάστασης με βήμα συνέλιξης* S *όπως ορίζεται παρακάτω:*

$$\text{ConvlD}_{C,K,S}\left(\mathbf{x}\right)_{i,l} = \mathbf{b}_i + \sum_{j=1}^{C_{in}} \sum_{k=1}^{K} \mathbf{V}_{i,j,k} \cdot \mathbf{x}_{j,S \cdot l - k} \tag{53}$$

Όπου οι σημάνσεις i, j, k, l *σηματοδοτούν το κανάλι εξόδου , το κανάλι εισόδου το δείγμα του πυρήνα και το χρονικό βήμα. Να σημειωθεί ότι χωρίς το σφάλμα γενικότητας και του κατάλληλου γεμίσματος/μετάθεσης (padding) η τελευταία διάσταση της αναπαράστασης της εξόδου θα ήταν* $L = \lceil L_{in} \big/ S \rceil$

Definition (2) $\text{ConvTr1D}_{C.K.S} : \mathbb{R}^{C_i n \times L_i n} \rightarrow \mathbb{R}^{C \times L}$ *ορίζει την μονοδιάστατη ανάστροφη συνέλιξη, εφόσον κάθε συνέλιξη μπορεί να εκφραστεί ως πολλαπλασιασμός πινάκων. Η ανάστροφη συνέλιξη είναι ταυτόσημη με τον υπολογισμό της κλίσης για την κανονική συνέλιξη σύμφωνα με [99].*

Definition (3) $\text{DWConv1D}_{C.K.S} : \mathbb{R}^{C_i n \times L_i n} \rightarrow \mathbb{R}^{C \times L}$ *σηματοδοτεί την μονοδιάστατη κατά-βάθος συνελίξεις. Στην πράξη αυτός ο τελεστής σηματοδοτείται ως* $G = C_{in}$ *σε διαφορετικές μονοδιάστατες συνέλιξης* $\mathcal{F}_i = [Conv1D_{C_G,K,S}]_i$ *με* $i \in \{1, \ldots, G\}$ *όπου* $C_G = \lfloor C/G \rfloor$. *Δεδομένης της εισόδου* $\boldsymbol{x} \in \mathbb{R}^{C_{in} \times L_{in}}$ *η ι-οστή μονοδιάστατη συνέλιξη συμβάλει στα* $C_G = \lfloor C/G \rfloor$ *κανάλια εξόδου αναγνωρίζοντας ως είσοδο μόνο την ι-οστή σειρά από την είσοδο που παρουσιάζεται από κάτω [96]:*

$$\text{DWConv1D}_{C,K,S}\left(\mathbf{x}\right) = \text{Concat}\left(\{\mathcal{F}_i\left(\mathbf{x}_i\right), \quad \forall i\}\right) \tag{54}$$

Όπου Concat$(\cdot)$ *πραγματοποιεί την συνένωση με όλες τις εξόδους από τις μεμονωμένες μονοδιάστατες συνελίξεις κατά δια μέσου της διάστασης του καναλιού.*

4.5.2.1 Κωδικοποιητής

Η αρχιτεκτονική του κωδικοποιητή $\mathcal{E}$ της μιας διάστασης συνέλιξης με μέγεθος πυρήνα $\text{K}\mathcal{E}$ και βήμα συνέλιξης ίσο με $\text{K}\mathcal{E} = 2$ παρομοίως με [35]. Κάθε ηχητικό τμήμα που έχει υποστεί συνέλιξη από $\text{K}\mathcal{E}$ δείγματα μετατρέπεται σε μια αναπαράσταση διανύσματος $C\mathcal{E}$-διαστάσεων όπου $C\mathcal{E}$ είναι ο αριθμός των καναλιών εξόδου από την μονοδιάστατη συνέλιξη .Επιβάλλεται η έξοδος του κωδικοποιητή να είναι αυστηρά μη αρνητική με την χρήση της Relu πάνω στην έξοδο της μονοδιάστατης συνέλιξης έτσι η κωδικοποιημένη αναπαράσταση της μείξης μπορεί να εκφραστεί ως [96]:

$$\mathbf{v}_{\mathbf{x}} = \mathcal{E}\left(\mathbf{x}\right) = \text{ReLU}\left(\text{ConvlD}_{C_\varepsilon, K_\varepsilon, K_\varepsilon/2}\left(\mathbf{x}\right)\right) \in \mathbb{R}^{C_\varepsilon \times L}, \tag{55}$$

Με την $\text{ReLU}(\cdot)$ να εφαρμόζεται στοιχείο προς στοιχείο.

4.5.2.2 Διαχωριστής

Στην πράξη ο διαχωριστής S πραγματοποιεί τις παρακάτω μεταμορφώσεις στην κωδικοποιημένη αναπαράσταση $\mathrm{v}_{\times} \in \mathbb{R}^{C_{\varepsilon} \times L}$:

1. προβάλλει την κωδικοποιημένη αναπαράσταση μείξης $\mathrm{v}_{\times} \in \mathbb{R}^{C_{\varepsilon} \times L}$ σε ένα νέο χωρικό κανάλι μέσω μιας layer-normalization (LN) [100] και στην συνέχεια μιας συνέλιξης σημείο προς σημείο όπως παρουσιάζεται παρακάτω:

$$\mathbf{y}_0 = \mathrm{ConvlD}_{C,1,1}\left(\mathrm{LN}\left(\mathbf{v}_{\mathbf{x}}\right)\right) \in \mathbb{R}^{C \times L} \tag{56}$$

   Το $LN(v_x)$ σηματοδοτεί ένα στρώμα κανονικοποίησης όπου οι ροπές εξάγονται δια μέσου της χρονικής διάστασης για κάθε κανάλι ξεχωριστά.

2. Εκτελεί μια επαναλαμβανόμενη μη γραμμική μεταμόρφωση που παρέχεται από B U-convolutional στην ενδιάμεση αναπαράσταση v_0. Δηλαδή η έξοδος του ι-οστού U-ConvBlock θα μπορούσε να εμφανιστεί ως $\mathbf{y}i \in \mathbb{R}^{C \times L}$ και θα χρησιμοποιηθεί για είσοδο στο $\iota - o\sigma\tau o + 1$ μπλοκ. Κάθε U-ConvBlock εξάγει και συσσωρεύει πληροφορία από πολλαπλές αναλύσεις.

3. Συσσωρεύει την πληροφορία από πολλαπλά κανάλια με την εφαρμογή μιας μονοδιάστατης συνέλεξης για κάθε πηγή στην ανάστροφη αναπαράσταση χαρακτηριστικών $\mathbf{y}_B^T \in \mathbb{R}^{L \times C}$. Ουσιαστικά για την ι-οστή πηγή λαμβάνεται μια ενδιάμεση λανθάνουσα αναπαράσταση [96]:

$$z_i = \mathrm{ConvlD}_{C,C_E,1}\left(\mathbf{y}_B^T\right)^T \in \mathbb{R}^{C_E \times L} \tag{57}$$

   Αυτό το βήμα πρωτοεμφανίστηκε στο [101]. Εμπειρικά φάνηκε ότι κάνει την εκπαίδευση πιο σταθερή παρά την χρησιμοποίηση συναρτήσεων ενεργοποίησης του τελευταίου μπλοκ Y_b για την εκτίμηση μασκών.

4. Συνδυάζει την προαναφερθείσες λανθάνουσες αναπαραστάσεις για όλες τις πήγες $z_i \forall i \in \{1, \cdots, N\}$ με την χρήση ενός softmax [102] για να λάβουμε τις εκτιμήσεις των μασκών $m_i \in [0,1]^{C\varepsilon \times L}$ που η πρόσθεση τους θα πρέπει να ισούται με 1. Με αυτόν τον τρόπο η αντίστοιχη εκτίμηση μάσκας για την ι-οστή πηγή θα είναι:

$$\hat{m}_i = \mathrm{vec}^{-1}\left(\frac{\exp\left(\mathrm{vec}(\mathbf{z}_i)\right)}{\sum_{j=1}^{N} \exp\left(\mathrm{vec}(\mathbf{z}_j)\right)}\right) \in \mathbb{R}^{C\varepsilon \times L}, \tag{58}$$

   Όπου $\mathrm{vec}\left(\cdot\right) : \mathbb{R}^{K \times N} \longrightarrow \mathbb{R}^{K \times N}$ και $\mathrm{vec}^{-1}\left(\cdot\right) : \mathbb{R}^{K \cdot N} \longrightarrow \mathbb{R}^{K \times N}$ σηματοδοτούν την διανυσματοποίηση του τένσορα εισόδου και την αντίστροφη πράξη αντίστοιχα.

5. υπολογίζεται μια λανθάνων μια αναπαράσταση $\hat{\mathbf{v}}_i \in \mathbb{R}^{C\varepsilon \times L}$ για κάθε πηγή με τον πολλαπλασιασμό κατά στοιχεία την αναπαράσταση κωδικοποιημένης μείξης v_x με την αντίστοιχη μάσκα $\hat{m}_i$:

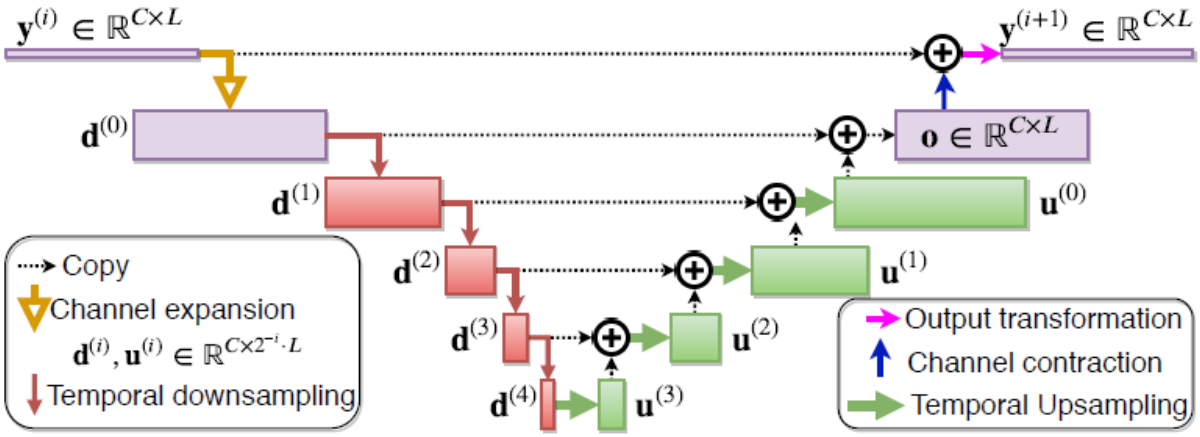
$$\hat{\mathbf{v}}_i = \mathbf{v}_{\mathbf{x}} \odot \hat{m}_i \in \mathbb{R}^{C_{\varepsilon} \times L} \tag{59}$$

   όπου $\alpha \odot \beta$ είναι ο πολλαπλασιασμός των τενσόρων κατά στοιχεία υποθέτοντας ότι έχουν ίδιο μέγεθος

4.5.2.3 U-συνελικτικά μπλοκ

Τα U-ConvBlock χρησιμοποιούν μια δομή με υπολειπόμενες σύνδεσης όπως στο [35]. Τα U-ConvBlock εξάγουν πληροφορία από πολλαπλές αναλύσεις με την χρήση Q συνεχόμενων χρονικών υποδειγματοληψιών και Q υπερδειγματοληψίες όπως στην αρχιτεκτονική του U-net [94]. Η έξοδος κάθε μπλοκ αφήνει την χρονική ανάλυση άθικτη ενώ αυξάνει την αποτελεσματικότητα του δεκτικού πεδίου πολλαπλασιαστικά με κάθε χρονική υποδειγματοληψία [103]. Μια εικόνα της αρχιτεκτονικής μπορεί να βρεθεί στο σχήμα [96].

4.5.3 Αρχιτεκτονική U-ConvBlock βασισμένο στο U-net



Σχήμα 4.6: Στο γράφημα παρουσιάζεται η αρχιτεκτονική του U-ConvBlock [96]

Definition (4) $\mathrm{PReLU}_C : \mathbb{R}^{C\times L} \to \mathbb{R}^{C\times L}$ *σηματοδοτεί μια PReLU [104] με* C *εκπαιδεύσιμες παραμέτρους με* $\boldsymbol{a} \in \mathbb{R}^C$ *.Όταν εφαρμόζεται σε ένα πίνακα εισόδου* $\mathbf{y} \in \mathbb{R}^{C\times L}$ *η μη γραμμική μετάθεση μπορεί να εκφραστεί κατά στοιχείο ως:*

$$\mathrm{PReLU}_C\left(y\right)_{i,j} = \max\left(0, y_{i,j}\right) + a_i \cdot \min\left(0, y_{i,j}\right) \tag{60}$$

Definition (5) $I_M \mathbb{R}^{C\times L} \to \mathbb{R}^{C\times M\cdot L}$ *ορίζει την κοντινότερη γειτονική χρονική παρεμβολή με συντελεστή* M. *Όταν εφαρμόζεται σε έναν πίνακα εισόδου αυτή η υπερδειγματοληπτική πράξη μπορεί να εκφραστεί ως:* $I_M\left(\mathbf{u}\right)_{i,j} = \mathbf{u}_{i,\lfloor j/M \rceil}$

Definition (6) $\mathrm{LN} : \mathbb{R}^{C\times L} \longrightarrow \mathbb{R}^{C\times L}$ *σηματοδοτεί ένα παραμετρικό στρώμα κανονικοποίησης [100] με εκπαιδεύσιμες παραμέτρους* $\gamma \in \mathbb{R}^C$ και $\beta \in \mathbb{R}^C$*.Όταν εφαρμόζεται σε ένα πίνακα εισόδου* $\mathbf{y} \in \mathbb{R}^{C\times L}$ *η κανονικοποίηση κατά στοιχείο μπορεί να γράφει ως:*

$$\mathrm{LN}\left(\mathbf{y}\right)_{i,j} = \frac{\mathbf{y}_{i,j} - \mu_i}{\sigma_i}\gamma_i + \beta_i, \quad \mu_i = \sum_j \mathbf{y}_{i,j}, \quad \sigma_i = \sqrt{\sum_j \left(\mathbf{y}_{i,j} - \mu_i\right)^2} \tag{61}$$

Definition (7) $\mathrm{GLM} : \mathbb{R}^{C\times L} \to \mathbb{R}^{C\times L}$ *σηματοδοτεί ένα παραμετρικό στρώμα κανονικοποίησης με εκπαιδεύσιμες παραμέτρουςγ* $\in \mathbb{R}^C$ και $\beta \in \mathbb{R}^C$*.Όταν εφαρμόζεται σε ένα πίνακα εισόδου* $\mathbf{y} \in \mathbb{R}^{C\times L}$ *η κανονικοποίηση κατά στοιχείο μπορεί να γράφει ως:*

$$\mathrm{GLN}\left(\mathrm{y}\right)_{i,j} = \frac{\mathrm{y}_{i,j} - \mu}{\sigma}\gamma_i + \beta_i,\ \mu = \sum_{i,j}\mathrm{y}_{i,j},\ \sigma = \sqrt{\sum_{i,j}\left(\mathrm{y}_{i,j} - \mu\right)^2} \tag{62}$$

4.5.3.1 Αποκωδικοποιητής

Ο αποκωδικοποιητής $\mathcal{D}$ είναι το τελευταίο βήμα για την μετάθεση την λανθάνουσας αναπαράστασης χώρου $\hat{\mathbf{v}}_i$ για κάθε πηγή πίσω στον τομέα του χρόνου. Κάθε λανθάνωνουσα αναπαράσταση πηγής $\hat{\mathbf{v}}_i$ τροφοδοτείται μέσα σε ένα διαφορετικό κωδικοποιητή ανάστροφης συνέλεξης $\mathrm{ConvTr1D}_{C_\varepsilon, K_\varepsilon, K_\varepsilon/2}$.Αγνοώντας το πρόβλημα των μεταθέσεων για την ι-οστή πηγή έχουμε την παρακάτω ανακατασκευή στο χρόνο:

$$\hat{s_i} = \mathcal{D}_i\left(\hat{\mathbf{v}}_i\right) = \mathrm{ConvTr1D}_{C_\varepsilon, K_\varepsilon, K_\varepsilon/2}\left(\hat{\mathbf{v}}_i\right) \tag{63}$$

4.5.4 Βελτιωμένη έκδοση χωρίς εκτίμηση μάσκας SuDoRMRFImprovedNet

Στην βελτιωμένη έκδοση το μοντέλο εκτιμά κατευθείαν την λανθάνουσα αναπαράσταση για κάθε σήμα $\hat{\mathbf{v}}_i \in \mathbb{R}^{C_\varepsilon \times L}$ και έπειτα χρησιμοποιεί των αποκωδικοποιητή. Η βασική υπόθεση αυτής της οπτικής είναι ότι ένα υψηλά παραμετροποιημένο νευρωνικό δίκτυο μπορεί να εκτιμήσει τους στόχους χωρίς την χρήση σκληρά-κανονικοποιημένου κατά στοιχεία πολλαπλασιασμού για την εύρεση μασκών $\mathbf{v}_x \in \mathbb{R}^{C_\varepsilon \times L}$. Το SuDoRMRFImprovedNet που παρουσιάζεται στο σχήμα 4.5 μπορεί να ληφθεί με τις ακόλουθες αλλαγές στην αρχιτεκτονική [96]:

- Μετά την τελευταία έξοδο του μοντέλου $y^(B)$ η εκτίμηση της μάσκας και ο κατά στοιχεία υπολογισμός εναλλάσσονται με απευθείας εκτίμηση των στόχων λανθάνουσα σημάτων $\hat{\mathbf{v}}_\mathbf{i}$. Έχει αποδειχθεί πειραματικά ότι η αφαίρεση της εκτίμησης της μάσκας οδηγεί σε παρόμοια η καλύτερα αποτελέσματα.
- Χρησιμοποιείται ένας εκπαιδεύσιμος αποκωδικοποιητής για την μετάθεση της λανθάνουσας αναπαράστασης πίσω στο τομέα του χρόνου αντί για δυο διαφορετικούς $\hat{s_i} = D_i\left(\hat{v}_i\right) = \mathrm{ConvTr1D}_{C_\varepsilon, K_\varepsilon, K_\varepsilon/2}\left(\hat{v}_i\right)$
- Το στρώμα κανονικοποίησης 61 αντικαθίσταται με ένα μακροχρόνιο στρώμα κανονικοποίησης 62. Αυτή η αλλαγή βελτιώνει σημαντικά την σύγκλιση του μοντέλου κυρίως λόγω των βραχυπροθέσμων εξαρτήσεων των στατιστικών της κλίσης μεταξύ των καναλιών.
- Για καθεμία αναπαράσταση με C κανάλια απλοποιείται η συνάρτηση ενεργοποίησης με μόνο μια παράμετρο εκμάθησης. Αυτό γίνεται για λιγότερες παραμέτρους.

4.5.5 Παραλλαγή ομαδικής επικοινωνίας

Προτείνεται μια νέα παραλλαγή C-SuDoRMRFImprovedNet που συνδυάζει την ομαδική επικοινωνία με το βελτιωμένο μοντέλο SuDoRMRFImprovedNet. Η ομαδική επικοινωνία είναι ένας τρόπος για σημαντική μείωση των παραμέτρων από ένα δίκτυο ηχητικής επεξεργασίας που προτάθηκε προσφάτως στο [105]. Οι ενδιάμεσες αναπαραστάσεις επεξεργάζονται σε τμήματα από υποζώνες καναλιών. Διαχωρίζουμε τα κανάλια με κάθε $1\chi1$ μπλοκ συνέλιξης σε 16 ομάδες και επεξεργάζονται πρώτα ανεξάρτητα

με τον διαμοιρασμό των παραμέτρων για όλες της ομάδες και υποζώνες των καναλιών. Πράγμα που οδηγεί σε σημαντική μείωση παραμέτρων στο μοντέλο [96].

4.5.6 C-SuDoRMRFImprovedNet

Η τελευταία εξέλιξη στο μοντέλο είναι η ικανότητα να τρέχει την εφαρμογή σε πραγματικές συνθήκες. Προτείνεται το C-SuDoRMRFImprovedNet το οποίο είναι ένα πιο ρηχό δίκτυο.

- Εναλλάσσονται όλες οι μη συνηθισμένες συνελίξεις με συνηθισμένες. Με αυτό τον τρόπο η αρχιτεκτονική δεν εξαρτάται από μελλοντικά στοιχεία για την εκτίμηση του σήματος μέχρι το παρόν καρέ.
- Όλα τα στρώματά κανονικοποίησης απαλείφονται για μικρότερη χρήση μνήμης.

4.5.7 Προεπεξεργασιά και αναπαραγωγή δεδομένων

Ακολουθείται η ίδια προεπεξεργασία με [101] Επίσης κανονικοποιούνται τα ηχητικά τμήματα με την αφαίρεση του μέσου και την διαίρεση με την τυπική απόκλιση.

4.5.8 Σταθερός αριθμός ομιλητών

Για την διαδικασία παραγωγής των δεδομένων εκπαίδευσης πραγματοποιούνται τα παρακάτω :

1. Τυχαία επιλογή δυο ομιλητών.
2. Τυχαία τμηματοποίηση στα 4 δευτερόλεπτα .
3. Τυχαία επιλογή ανάλογιας σημάτος προς θόρυβο για κάθε εποχή 20000 μείξης παράγονται και 3000 για επιβεβαίωσή κάθε ηχητικό τμήμα έχει συχνότητα 8kHz.

4.5.9 Ρυθμίσεις εκπαίδευσης

Όλα τα μοντέλα εκπαιδεύτηκαν για 120 εποχές με batch size 4. Για υπολογισμό του σφάλματος χρησιμοποιήθηκαν τα negative permutation-invariant [31] scale-invariant signal to distortion ratio (SI-SDR) [92]. Το συνολικό σφάλμα για N πηγές υπολογίζεται ως το μέσο σφάλμα κατά μήκος κάθε πηγής. Για την ι-οστή πηγή ορίζεται το σφάλμα μεταξύ του καθαρού σήματος και των εκτιμήσεων ως:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \text{SI} - \text{SDR}(\text{s}_i^*, \hat{\mathbf{s}}_i) = -\frac{1}{N}\sum_{i=1}^{N} 10\log_{10}\left(\frac{||\alpha_i \mathbf{s}_i^*||^2}{||\alpha \text{s}_i^* - \hat{\mathbf{s}}_i||^2}\right). \tag{64}$$

Όπου s^* ορίζει την μετάθεση στις πηγές που μεγιστοποιεί το SI-SDR και μια διαβάθμισή $\alpha_i = \hat{\mathbf{s}}_i^\top \mathbf{s}_i^* / \mathbf{s}_i||^2$ χρησιμοποιείται για να κάνει το σφάλμα αμετάβλητο με την βαθμίδα της ι-οστής πηγής $\hat{\mathbf{s}}_i$. Κατά την διάρκειά της εκπαίδευσής χρησιμοποιείται ο αλγόριθμος Adam με αρχικό βήμα εκμάθησης 0.001 και μειώνεται κατά 5 κάθε 50 εποχές [96]

4.5.9.1 Προτεινόμενες ρυθμίσεις για κάθε έκδοση

Παρακάτω περιγράφονται οι προεπιλεγμένες τιμές για τις αρχιτεκτονικές SuDoRM-RF, SuDoRMRFImprovedNet , C-SuDoRMRFImprovedNet. Για των κωδικοποιητή E και των αποκωδικοποιητή χρησιμοποιείται μέγεθος πυρήνα 21 για την μείξη εισόδου και η συχνότητα ορίζεται στα $8kHz$, Η βάση ορίζεται στα 512. Για κάθε U-ConvBlock ο αριθμός των καναλιών εισόδου είναι $C = 128$. Και ο αριθμός των επαναλαμβανόμενων επαναδειγματοληψιών είναι $Q = 4$. Ο διασταλμένος αριθμός καναλιών ανέρχεται στα $CU = 512$. Σε κάθε υποδειγματοληψία μειώνεται η χρονική διάσταση κατά βαθμό 2 και όλες οι κατά βάθος ξεχωριστές συνελίξεις έχουν μέγεθος πυρήνα $KU = 5$ και βήμα συνέλιξης $SU = 2$. Μόνο για την παραλλαγή C-SuDoRMRFImprovedNet αυξάνεται ο αριθμός των καναλιών σε $C = 256$ και το μέγεθος πυρήνα $KU = 11$. Αυτό συμβαίνει για να αυξηθεί το λαμβάνον πεδίο για ρηχότερες και αποτελεσματικότερες αρχιτεκτονικές με απαιτήσεις πραγματικών καταστάσεων. Για ευκολία χρησιμοποιείται η ακόλουθη σήμανση βασισμένη στον αριθμό B από U-ConvBlocks μέσα στο τμήμα του διαχωριστή S. Τα SuDoRM-RF 2.0x, SuDoRM-RF 1.0x, SuDoRM-RF 0.5x, SuDoRM-RF 0.25x περιέχουν από $32, 16, 8$ και 4 μπλοκς αντίστοιχα. Το ίδιο ισχύει και για την βελτιωμένη έκδοση του SuDoRMRFImprovedNet και C- SuDoRMRFImprovedNet [96].

Κεφάλαιο 5ο: Σετ δεδομένων και πειράματα

Το σετ δεδομένων Wall Street Journal 2 mix (WSJ2mix) [106] είναι μια συλλογή από προφορικές ηχογραφήσεις της Wall Street Journal. Αναπτύχθηκε ως πρότυπο για την εκπαίδευση και την αξιολόγηση μοντέλων αναγνώρισης ομιλίας. Το σύνολο δεδομένων περιέχει 2.000 ώρες ήχου και περισσότερες από 2 εκατομμύρια λέξεις μεταγραμμένου κειμένου. Οι επαγγελματίες ηθοποιοί κατέγραψαν τον ήχο και οι μεταγραφές ελέγχθηκαν χειροκίνητα για ακρίβεια. Το σύνολο δεδομένων χωρίζεται σε τρεις ενότητες: εκπαίδευση, ανάπτυξη και δοκιμή, τα οποία μπορούν να χρησιμοποιηθούν για την αξιολόγηση της απόδοσης των μοντέλων αναγνώρισης ομιλίας σε διάφορους τύπους ομιλίας. Ωστόσο η πρόσβαση του δεν είναι δωρεάν. Τα μοντέλα Sepformer [44],DualpathRnn [36] έχουν εκπαιδευτεί από άλλους ερευνητές πάνω στο WSJ2mix.

5.1 Wall Street Journal 2 mix (WSJ2mix)

Rank	Model	SI-SDRi	SDRi	Extra Training Data	Paper	Code	Result	Year	Tags
1	SepIt	22.4		×	SepIt: Approaching a Single Channel Speech Separation Bound			2022	
2	SepFormer	22.3	22.4	×	Attention is All You Need in Speech Separation			2020	Transformer
3	Wavesplit v2	22.2	22.3	×	Wavesplit: End-to-End Speech Separation by Speaker Clustering			2020	CNN
4	DPTNet (Libri1Mix speech enhancement pre-trained)	21.3	21.5	✓	Stabilizing Label Assignment for Speech Separation by Self-supervised Pre-training			2020	Transformer
5	Sandglasset	21.0		×	Sandglasset: A Light Multi-Granularity Self-attentive Network For Time-Domain Speech Separation			2021	multiscale Transformer LSTM
6	GALR	20.3		×	Effective Low-Cost Time-Domain Audio Separation Using Globally Attentive Locally Recurrent Networks			2021	LSTM
7	DPTNet	20.2		×	Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation			2020	Transformer
8	Gated DualPathRNN	20.12		×	Voice Separation with an Unknown Number of Multiple Speakers			2020	
9	Sudo rm -rf (U=36)	19.5		×	Compute and memory efficient universal sound source separation			2021	
10	Wavesplit v1	19.0		×	Wavesplit: End-to-End Speech Separation by Speaker Clustering			2020	CNN
11	Sudo rm -rf XL	18.9		×	Sudo rm -rf: Efficient Networks for Universal Audio Source Separation			2020	CNN
12	Dual-path RNN	18.8		×	Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation			2019	LSTM
13	DeepCASA	17.7		×	Divide and Conquer: A Deep CASA Approach to Talker-independent Monaural Speaker Separation			2019	
14	IAC-PIT Tasnet	17.5		×	Interrupted and cascaded permutation invariant training for speech separation			2019	
15	Deformable TCN + Dynamic Mixing	17.2	17.4	×	Deformable Temporal Convolutional Networks for Monaural Noisy Reverberant Speech Separation			2022	
16	Hybrid-Tasnet	16.6		×	Improved Speech Separation with Time-and-Frequency Cross-domain Joint Embedding and Clustering			2019	
17	Two-step Conv-TasNet	16.1		×	Two-Step Sound Source Separation: Training on Learned Latent Targets			2019	CNN
18	Conv-TasNet	15.3		×	Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation			2018	CNN

Σχήμα 5.1: Πίνακας κατάταξης wsj2mix

Ο πίνακας κατάταξης 5.1 για το σύνολο δεδομένων WSJ2MIX μπορεί να βρεθεί παρακάτω υπερσύνδεσμο `https://paperswithcode.com/sota/speech-separation-on-wsj0-2mix`.

5.2 Libri2mix

Το Libri2mix αποτελείται από καθαρά και θορυβώδη δεδομένα που περιέχουν μείξεις δύο και τριών ομιλητών. Μιας και αντλεί δεδομένα από το LibriSpeech [107] σετ δεδομένων ακολουθεί και την οργάνωση του. Έχουν δύο σύνολα εκπαίδευσης (train-100, train-360), ένα σύνολο επικύρωσης (dev) και ένα σύνολο δοκιμών (test). Για να καλύψουμε το υποσύνολο train-360 του LibriSpeech χωρίς επανάληψη, τα δεδομένα θορύβου εκπαίδευσης έχουν προσαρμοστεί με συντελεστές ταχύτητας 0,8 και 1,2 όπως περιγράφεται στο [108, 109].

5.2.1 Τρόπος μείξης

`https://github.com/ShakedDovrat/LibriMix/blob/master/scripts/create_librimix_from_metadata.py#L375` στον παραπάνω υπερσύνδεσμο μπορεί να βρεθεί ο πηγαίος κώδικας για την παραγωγή των δεδομένων με την ακόλουθη συνάρτηση να κάνει την μείξη (Ουσιαστικά είναι απλό overlap δηλαδή στοίβαξη της μιας πηγής πάνω στην άλλη).

```python
def mix(sources_list):
    """ Do the mixing """
    # Initialize mixture
    mixture = np.zeros_like(sources_list[0])
    for source in sources_list:
        mixture += source
    return mixture
```

Σχήμα 5.2: Συνάρτηση μίξεως librimix

Rank	Model	SI-SDRi	SDRi	Extra Training Data	Paper	Code	Result	Year	Tags
1	**TDANet Large**	17.4		×	An efficient encoder-decoder architecture with top-down attention for speech separation			2022	
2	**TDANet**	16.9		×	An efficient encoder-decoder architecture with top-down attention for speech separation			2022	
3	**Conv-Tasnet** (Libri1Mix speech enhancement pre-trained)	14.1	14.6	✓	Stabilizing Label Assignment for Speech Separation by Self-supervised Pre-training			2020	
4	**Conv-Tasnet** (Libri1Mix speech enhancement multi-task)	13.7	14.1	✓	Stabilizing Label Assignment for Speech Separation by Self-supervised Pre-training			2020	
5	**Conv-Tasnet**	13.2	13.6	×	Stabilizing Label Assignment for Speech Separation by Self-supervised Pre-training			2020	

Σχήμα 5.3: Πίνακας κατάταξης Libri2mix

Στο παρακάτω υπερσύνδεσμο μπορεί να βρεθεί ο πίνακας κατάταξης 5.3 για το Libri2mix `https://paperswithcode.com/sota/speech-separation-on-Libri2mix`

5.3 Μετρικές

5.3.1 SI-SDR

Το SI-SDR είναι ένα μαθηματικό μέτρο της ποιότητας ενός αλγόριθμου διαχωρισμού πηγών. Αντιπροσωπεύει ”αναλογία σήματος προς παραμόρφωση αμετάβλητης κλίμακας” και χρησιμοποιείται για την αξιολόγηση της απόδοσης αλγορίθμων που προσπαθούν να διαχωρίσουν ένα μείγμα ηχητικών σημάτων στα επιμέρους στοιχεία τους. Το SI-SDR υπολογίζεται αρχικά υπολογίζοντας τον λόγο σήματος προς παραμόρφωση (SDR) για τα διαχωρισμένα σήματα και, στη συνέχεια, κλιμακώνοντας το αποτέλεσμα με μια σταθερή τιμή. Όσο υψηλότερο είναι το SI-SDR, τόσο καλύτερη είναι η ποιότητα του διαχωρισμού. Το SI-SDR χρησιμοποιείται συχνά σε συνδυασμό με άλλες μετρήσεις, όπως ο λόγος σήματος προς παρεμβολή (SIR) και ο λόγος σήματος προς atrifact (SAR), για να παρέχει μια πιο ολοκληρωμένη αξιολόγηση του αλγόριθμου διαχωρισμού [92]. Περισσότερες πληροφορίες μπορείτε να δείτε εδώ (64).

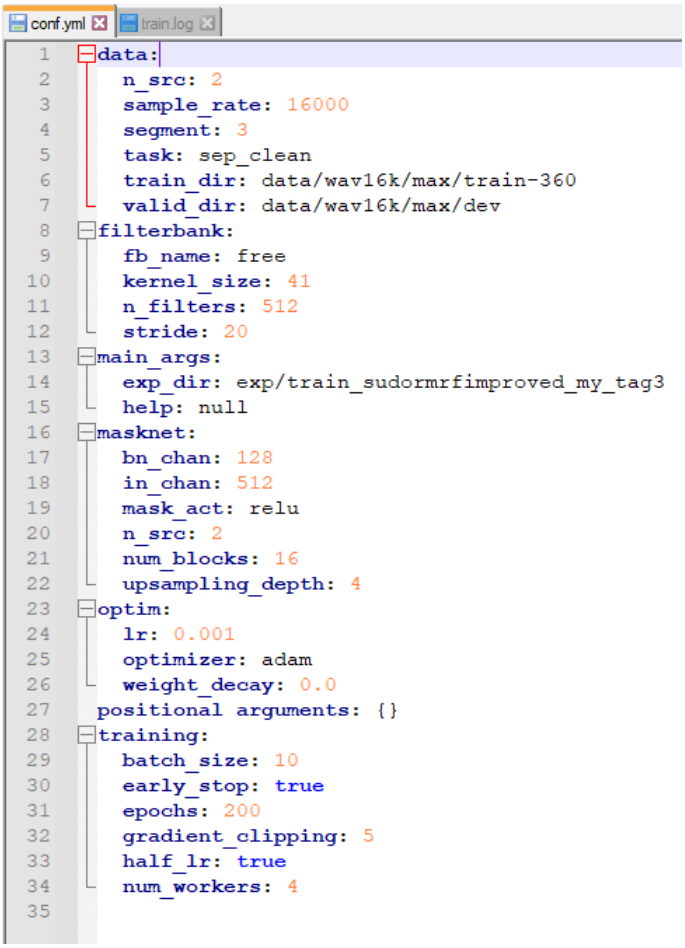
5.4 Πειράματα και συγκρίσεις

Τα πειράματα πραγματοποιήθηκαν με το framework της pytorch [110] και την επέκταση του με το Pytorch lighting [111]. Το Pytorch lighting είναι μια απλοποιημένη έκδοση του pytorch που όπως υποστηρίζει αφαιρεί τον περιττό κώδικα. Επιπλέον χρησιμοποιείται η δομή της Asteriod [112] που είναι μια βιβλιοθήκη που έπειτα από συνένωση με την speechbrain [113] είναι μια από τις δημοφιλέστερες βιβλιοθήκες για πολλά προβλήματα όπως ενδυνάμωση ομιλίας , αναγνώριση ομιλίας , διαχωρισμό ομιλίας και άλλα. Για τα γραφήματα έγινε η χρήση του tensorboard (εργαλείο εικονοποίησης του tensorflow [114]). Επιπλέον χρησιμοποιήθηκε το Libri2mix [108] χωρίς επιπλέον θόρυβο για όλα τα πειράματα.

5.4.1 SuDORMRFImprovednet

5.4.1.1 Ρυθμίσεις

Σε αυτή την υποενότητα θα αναφέρουμε τις διάφορες ρυθμίσεις καθώς και το πως επηρεάζει η καθεμία την μάθηση.



Σχήμα 5.4: Ρυθμίσεις εκπαίδευσης SuDORMRFImprovednet

Στη πρώτη υποομάδα βρίσκονται οι ρυθμίσεις του σετ δεδομένων

- n_src: αριθμός ομιλητών (πιθανές τιμές 2,3)
- sample_rate: ρυθμός δειγματοληψίας ενός ηχητικού σήματος είναι ο αριθμός των δειγμάτων ακουστικής κυματομορφής που λαμβάνονται ανά δευτερόλεπτο, μετρημένοι σε hertz (Hz) ή kHz. Ένας υψηλότερος ρυθμός δειγματοληψίας παράγει μια εγγραφή υψηλότερης ποιότητας, αλλά παράγει επίσης μεγαλύτερο μέγεθος αρχείου. Τα 44,1 kHz, 48 kHz και 96 kHz είναι συνήθεις ρυθμοί δειγματοληψίας για εγγραφές ήχου πιθανές τιμές είναι 8kHz , 16kHz.
- segment: η δυνατότητα αποκοπής εγγραφών μικρότερου από του αναγραφομένου μεγέθους σε δευτερόλεπτα.
- task : μια από τις πιθανές τιμές από (’enh_single’, ’enh_both’, ’sep_clean’, ’sep_noisy’) είναι για την ενίσχυση ενός , μείξη ομιλίας , διαχωρισμό μείξη καθαρού ή θορυβώδους μείγματος αντίστοιχά.

Στη δεύτερη υποομάδα βρίσκονται οι ρυθμίσεις των filterbanks (Οι τράπεζες φίλτρων χρησιμοποιούνται για την εξαγωγή χαρακτηριστικών από το σήμα ήχου που μπορούν να χρησιμοποιηθούν για την αναπα-

ράσταση του σήματος με πιο συμπαγή και ουσιαστικό τρόπο και συχνά χρησιμοποιούνται ως είσοδος σε μοντέλα μηχανικής εκμάθησης για εργασίες όπως η αναγνώριση ομιλίας και η ταξινόμηση ήχου):

- fb_name: μέθοδος εξαγωγής χαρακτηριστικών (melgram,stft,encoder decoder,free).
- kernel_size: μέγεθος πυρήνα (Μια τράπεζα φίλτρων με μικρό μέγεθος πυρήνα θα έχει υψηλή ανάλυση χρόνου αλλά ανάλυση χαμηλής συχνότητας, και το αντίστροφο).
- n_filters: αριθμός φίλτρων (χρησιμοποιούνται για τη διαίρεση του σήματος ήχου σε διαφορετικές περιοχές συχνοτήτων).
- stride: ρυθμός βήματος του πυρήνα πάνω στον αριθμό των φίλτρων.

Στη τρίτη υποομάδα βρίσκονται οι ρυθμίσεις του δικτύου παράγωγης μασκών:

- bn_chan: Αριθμός περιοχών τοποθέτησης στο στρώμα σμίκρυνσης και Unet μπλοκ.
- in_chan: είναι ο αριθμός των καναλιών εισόδου συνήθως ίσος με τον αριθμό των n_filters.
- mask_act: είναι η συνάρτηση ενεργοποίησης που χρησιμοποιείται στο masknet.
- n_src: αριθμός ομιλητών (πιθανές τιμές 2,3).
- num_blocks: είναι ο αριθμός των UBlocks στο δίκτυο.
- upsampling_depth: είναι ο αριθμός των φορών που γίνεται upsampling (υπερδειγματοληψία) του σήματος.

Στη τέταρτη υποομάδα βρίσκονται οι ρυθμίσεις του αλγορίθμου βελτιστοποίησης.

- lr: το βήμα εκμάθησης ελέγχει το μέγεθος του βήματος των ενημερώσεων κατά τη διάρκεια της εκπαίδευσης, ένα μικρό θα οδηγήσει σε ακριβείς λύσεις αλλά αργή σύγκλιση, ενώ ένα μεγάλο θα οδηγήσει σε ταχύτερη σύγκλιση αλλά σε κίνδυνο υπέρβασης ή απόκλισης από τη βέλτιστη λύση.
- optimizer: ο αλγόριθμος που θα χρησιμοποιηθεί για την βελτιστοποίηση μπορεί να είναι adam , στοχαστική κατάβαση κλίσης (stohastic gradient decent).
- weight_decay: Η ιδέα πίσω από τη μείωση του βάρους είναι να προστεθεί ένας όρος ποινής στη συνάρτηση σφάλματος που αποθαρρύνει το μοντέλο να έχει μεγάλα βάρη, κάτι που μπορεί να βοηθήσει στην αποφυγή υπερβολικής προσαρμογής (overfitting).

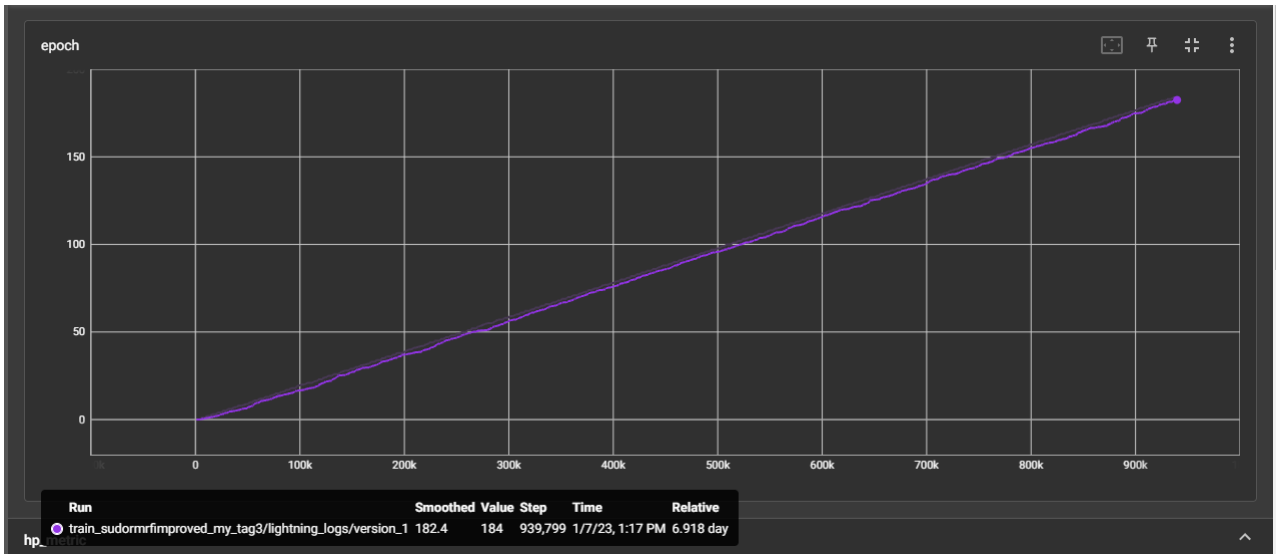
Στη τελευταία υποομάδα βρίσκονται περαιτέρω ρυθμίσεις εκπαίδευσης:

- batch_size: είναι η ομαδοποίηση των δεδομένων όταν τροφοδοτούνται για εκπαίδευση, Οσο μεγαλύτερη ομαδοποίηση τόσο γρηγορότερη η εκπαίδευση.

- early_stop είναι μια ρύθμιση που ανάλογα με την θέληση μας για κάποιο κριτήριο μπορεί να σταματήσει την εκπαίδευση.
- epochs: πόσες εποχές θα τρέξει ο αλγόριθμος
- gradient_clipping: είναι μια τεχνική που χρησιμοποιείται για να αποτρέψει να γίνουν πολύ μεγάλες οι κλίσεις ενός μοντέλου.
- half_lr: σε περίπτωση αποτυχίας κάποιου κριτηρίου απόδοσης μειώνεται το βήμα εκπαίδευσης στη μέση.
- num_workers: Όταν χρησιμοποιείται μεγαλύτερος αριθμός η φόρτωση και η προεπεξεργασία δεδομένων θα είναι ταχύτερη, καθώς μπορεί να γίνει παράλληλα.

Επιπλέον μερικές διευκρινήσεις σχετικά με τα δεδομένα, κριτήρια εφαρμογής, συναρτήσεις σφαλμάτων. Για όλα τα παρακάτω πειράματα χρησιμοποιήθηκαν δεδομένα 212 ηχητικών ωρών Libri2mix/train-360 με 50800 στιγμιότυπα για εκπαίδευση και τεστ καθώς 11 ώρες και 3000 στιγμιότυπα για επικύρωση. Για την παράμετρο half_lr το κριτήριο εφαρμογής είναι η μη συνεχόμενη βελτίωση του validation σφάλματος για 5 εποχές. Ενώ το early_stop έχει κριτήριο εφαρμογής τη μη συνεχόμενη βελτίωση του validation σφάλματος για 30 εποχές. Το σφάλμα που χρησιμοποιήθηκε για όλα τα πειράματα είναι η (64). Ενώ υπό φυσιολογικές συνθήκες τα σφάλματα θέλουμε να τα ελαχιστοποιήσουμε στο 0 σε αυτή την περίπτωση επειδή το σφάλμα ορίζεται ως το αρνητικό του SI-SDR, μια χαμηλότερη τιμή στην πραγματικότητα αντιστοιχεί σε υψηλότερη τιμή του SI-SDR, πράγμα που σημαίνει ότι η απόδοση αυξάνεται. Επίσης εδώ να αναφέρω ότι όλα τα πειράματα γίνανε με την χρήση μια κάρτας γραφικών Nvidia rtx A4000 και διαθέσιμη μνήμη ram 32 gb.

5.4.1.2 Γραφήματα και αποτελέσματα

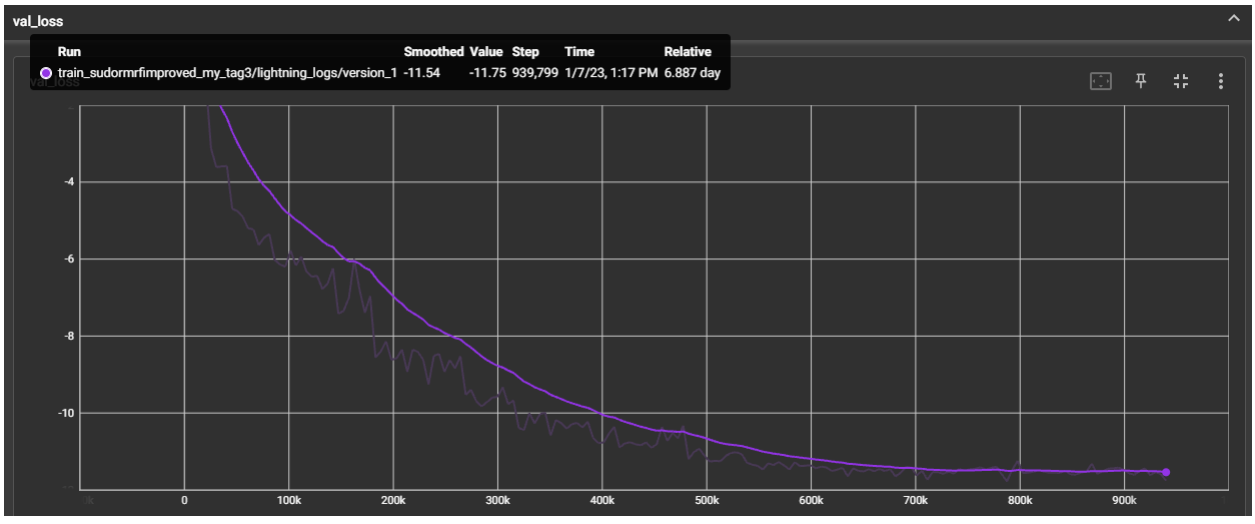


Σχήμα 5.5: Χρόνος εκτέλεσης SuDORMRFImprovednet

Σε αυτό το σχήμα βλέπουμε τον χρόνο εκτέλεσης του πειράματος με 184 εποχές και περίπου 7 ημέρες εκπαίδευσης Στους άξονες είναι οι εποχές και τα αντίστοιχά συνολικά βήματα.

Σχήμα 5.6: Σφάλμα εκπαίδευσης SuDORMRFImprovednet

Σε αυτό το γράφημα 5.6 βλέπουμε ότι το σφάλμα εκπαίδευσης που στόχος είναι η ελαχιστοποίηση του. Βλέπουμε ότι την περισσότερη επίδραση έχουν οι πρώτες 100 εποχές σύμφωνα με τα αντίστοιχα βήματα ωστόσο η ελάχιστη τιμή έρχεται κοντά στην εποχή 160.



Σχήμα 5.7: Σφάλμα επικύρωσης SuDORMRFImprovednet

Στο γράφημά 5.7 βλέπουμε ότι στο σφάλμα επικύρωσης έχει την περισσότερη επίδραση έχουν οι πρώτες 100 εποχές σύμφωνα με τα αντίστοιχα βήματα.

```
Overall metrics :
{'sar': 14.852375228381717,
 'sar_imp': -54.042857616536175,
 'sdr': 14.067047784979584,
 'sdr_imp': 13.999191188306305,
 'si_sdr': 13.668311493335292,
 'si_sdr_imp': 13.66798521975511,
 'sir': 23.45611526413674,
 'sir_imp': 23.388257057039596,
 'stoi': 0.9262572733907176,
 'stoi_imp': 0.16302127914861234}

Don't forget to share your pretrained models at https://zenodo.org/communities/asteroid-models/! =)
You can directly use our CLI for that, run this:
`asteroid-upload exp/train_sudormrfimproved_my_tag3/publish_dir --uploader "Your name here"`
```

Σχήμα 5.8: Τα αποτελέσματα SuDORMRFImprovednet

Τα αποτελέσματα όπως φαίνονται στο σχήμα 5.8 στη κυρία μετρική δείχνουν πολύ υποσχόμενα καθώς σύμφωνα με την βιβλιογραφία (si-sdr) το μοντέλο Convtasnet με διπλάσιες περίπου εκπαιδεύσιμες παραμέτρους έχει ελαφρώς χειρότερη απόδοση στην περίπτωση που δεν έχει χρησιμοποιηθεί το librimix για ενίσχυση.

5.4.2 SuDORMRFImprovednetweighted

5.4.2.1 Ρυθμίσεις

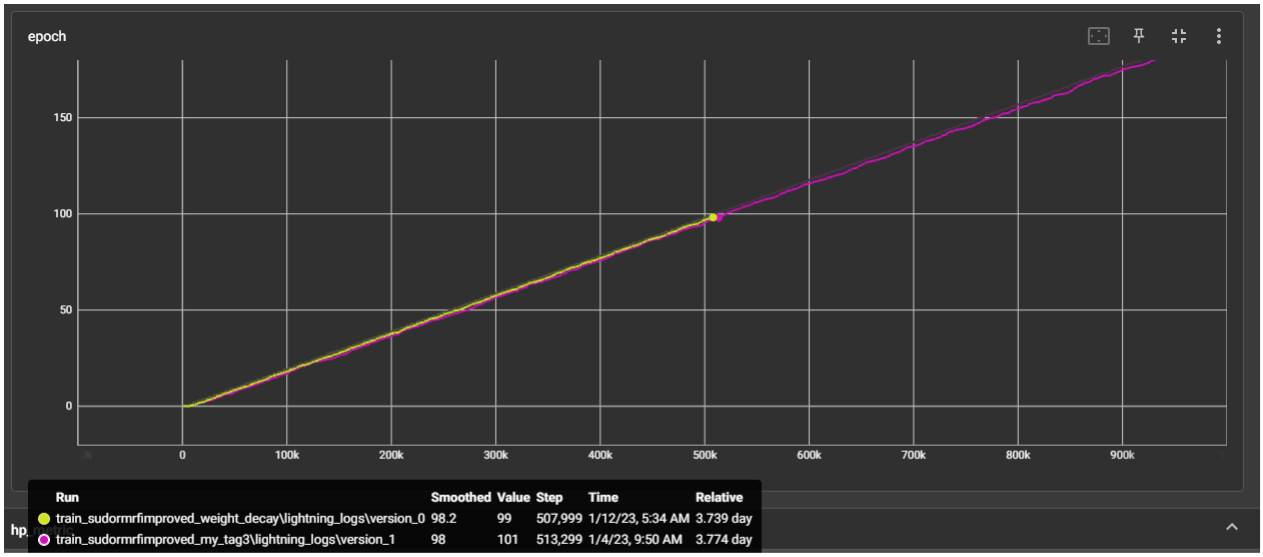
conf.yml | train.log | train.py | eval.log | final_metrics.json | conf.yml

```yaml
data:
  n_src: 2
  sample_rate: 16000
  segment: 3
  task: sep_clean
  train_dir: data/wav16k/max/train-360
  valid_dir: data/wav16k/max/dev
filterbank:
  fb_name: free
  kernel_size: 41
  n_filters: 512
  stride: 20
main_args:
  exp_dir: exp/train_sudormrfimproved_weight_decay
  help: null
masknet:
  bn_chan: 128
  in_chan: 512
  mask_act: relu
  n_src: 2
  num_blocks: 16
  upsampling_depth: 4
optim:
  lr: 0.001
  optimizer: adam
  weight_decay: 1.0e-05
positional arguments: {}
training:
  batch_size: 10
  early_stop: true
  epochs: 100
  gradient_clipping: 5
  half_lr: true
  num_workers: 4
```

Σχήμα 5.9: Ρυθμίσεις εκπαίδευσης SuDORMRFImprovednetweighted

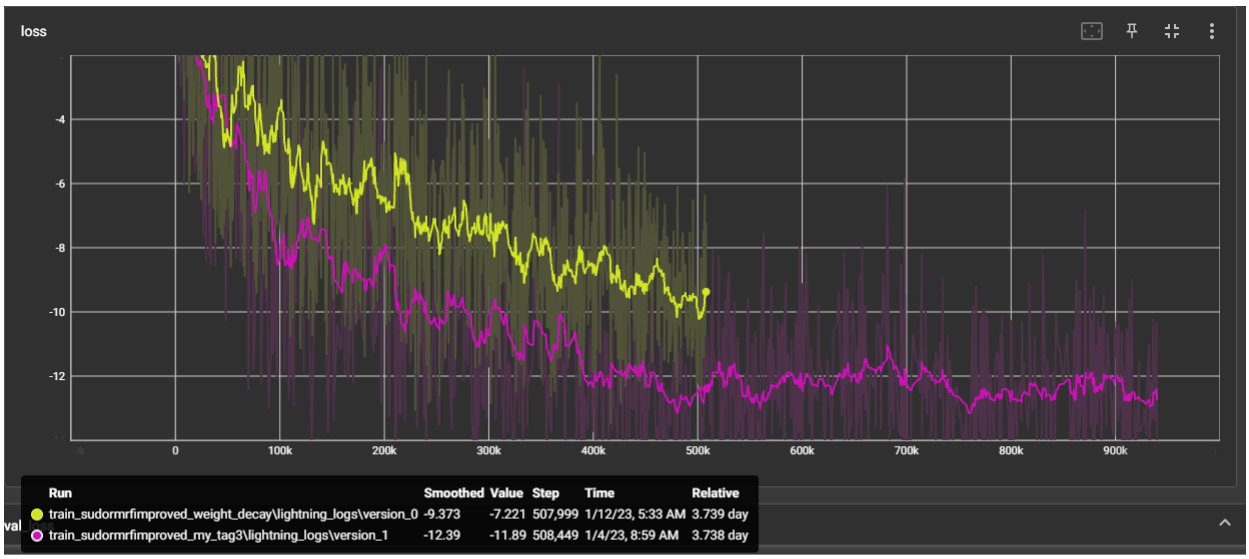
οι μόνες ουσιαστικές αλλαγές είναι ο αριθμός εποχών και η ένταξη του weight_decay.

5.4.2.2 Γραφήματα και αποτελέσματα



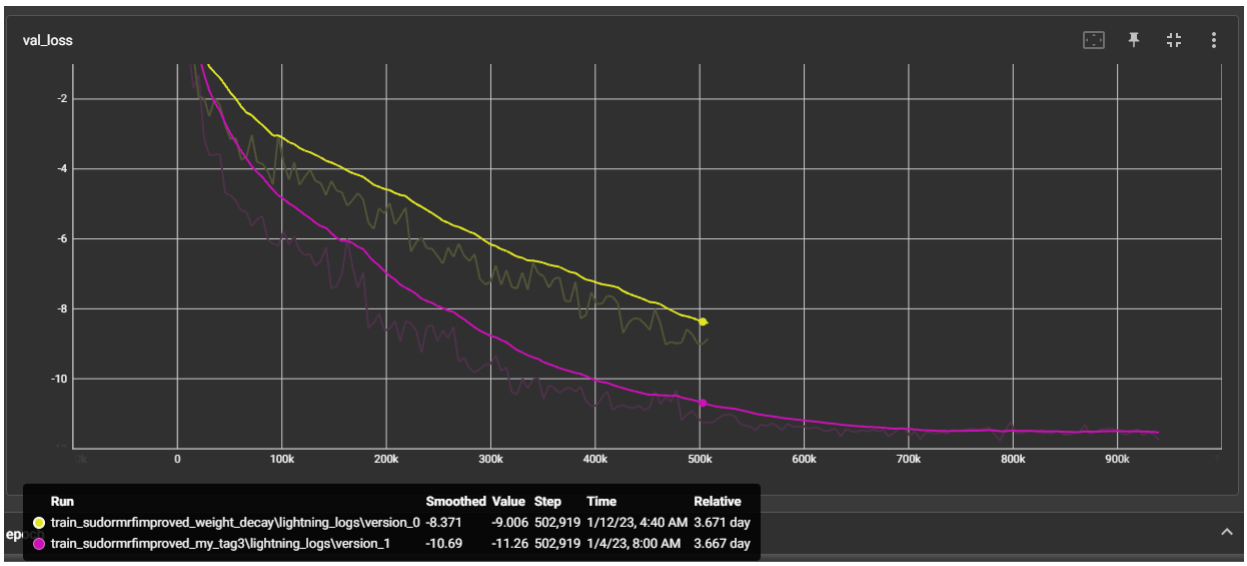
Σχήμα 5.10: Ο χρόνος εκπαίδευσης του SuDORMRFImprovednetweighted

Ο χρόνος εκπαίδευσης του SuDORMRFImprovednetweighted δεν έχει σημαντική διαφορά πέρα από την αλλαγή στις εποχές πράγμα που είναι αναμενόμενο σύμφωνα με το γράφημα 5.10.



Σχήμα 5.11: Σφάλμα εκπαίδευσης του SuDORMRFImprovednetweighted

Συγκριτικά με το προηγούμενο πείραμα φαίνεται ότι η προσθήκη της διαβάθμισης βαρών (weighted_decay) έχει επιδράσει αρνητικά πάνω στη βελτιστοποίηση του σφάλματος εκπαίδευσης όπως φαίνεται στο σχήμα 5.11.



Σχήμα 5.12: Σφάλμα επικύρωσης του SuDORMRFImprovednetweighted

Συγκριτικά με το προηγούμενο πείραμα φαίνεται ότι η προσθήκη της weighted decay (διαβάθμισης βαρών) έχει επιδράσει αρνητικά πάνω στη βελτιστοποίηση του σφάλματος επικύρωσης όπως φαίνεται στο σχήμα 5.12.


```
{
"si_sdr": 11.074471914462745,
"si_sdr_imp": 11.074145640882564,
"sdr": 11.497446380785952,
"sdr_imp": 11.429589784112673,
"sir": 19.477738750568907,
"sir_imp": 19.40988054347176,
"sar": 12.659178137307944,
"sar_imp": -56.23605470760995,
"stoi": 0.8968635380983753,
"stoi_imp": 0.13362754385627004
}
```


Σχήμα 5.13: Αποτελέσματα SuDORMRFImprovednetweighted

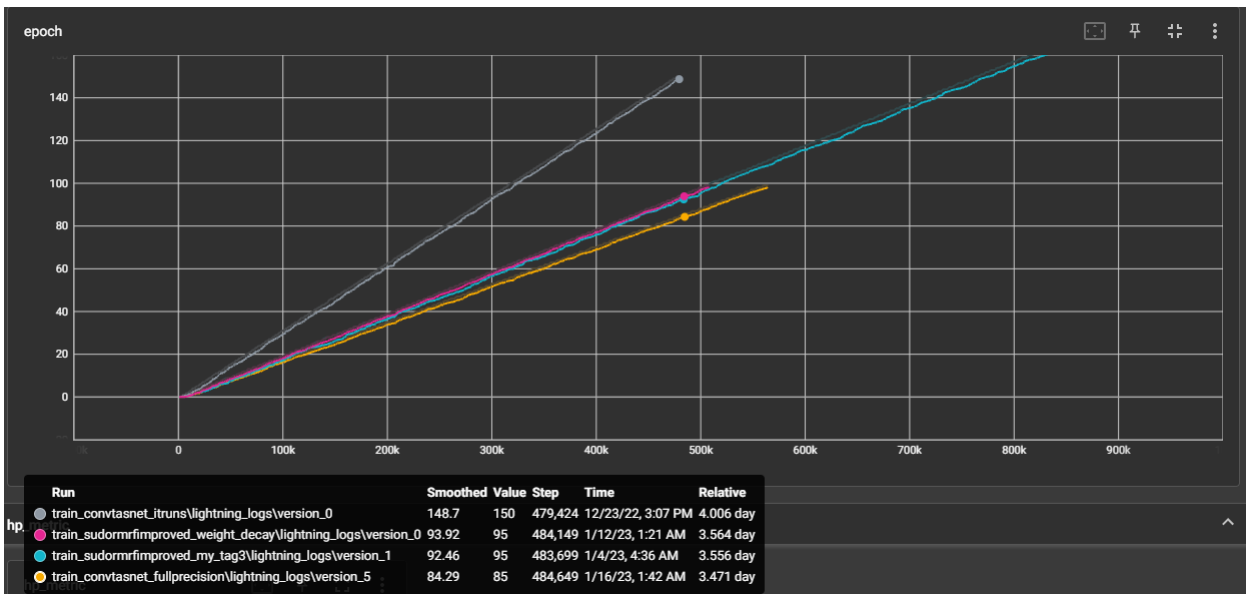
Όπως ήταν αναμενόμενο τα αποτελέσματα είναι σαφώς χειρότερα και όχι μόνο λόγω των λιγότερων εποχών πράγμα που επιβεβαιώνουν τα σφάλματα εκπαίδευσης και επικύρωσης. Η αλλαγή στο weight_decay είναι αυτή που προκαλεί την φαινομενική πτώση στην ποιότητα αποτελέσματος. Αυτό συμβαίνει γιατί πιθανώς η τιμή της παραμέτρου επηρεάζει υπερβολικά τις διορθώσεις και έτσι να δυσκολεύεται να μάθει ο αλγόριθμος.

5.4.2.3 Ολική εικόνα

```yaml
data:
  n_src: 2
  sample_rate: 8000
  segment: 3
  task: sep_clean
  train_dir: data/wav8k/min/train-360
  valid_dir: data/wav8k/min/dev
filterbank:
  kernel_size: 16
  n_filters: 512
  stride: 8
main_args:
  exp_dir: exp/train_convtasnet_fullprecision
  help: null
masknet:
  bn_chan: 128
  hid_chan: 512
  mask_act: relu
  n_blocks: 8
  n_repeats: 3
  n_src: 2
  skip_chan: 128
optim:
  lr: 0.001
  optimizer: adam
  weight_decay: 0.0
positional arguments: {}
training:
  batch_size: 9
  early_stop: true
  epochs: 100
  half_lr: true
  num_workers: 4
```

Σχήμα 5.14: Ρυθμίσεις πειραμάτων convtasnet

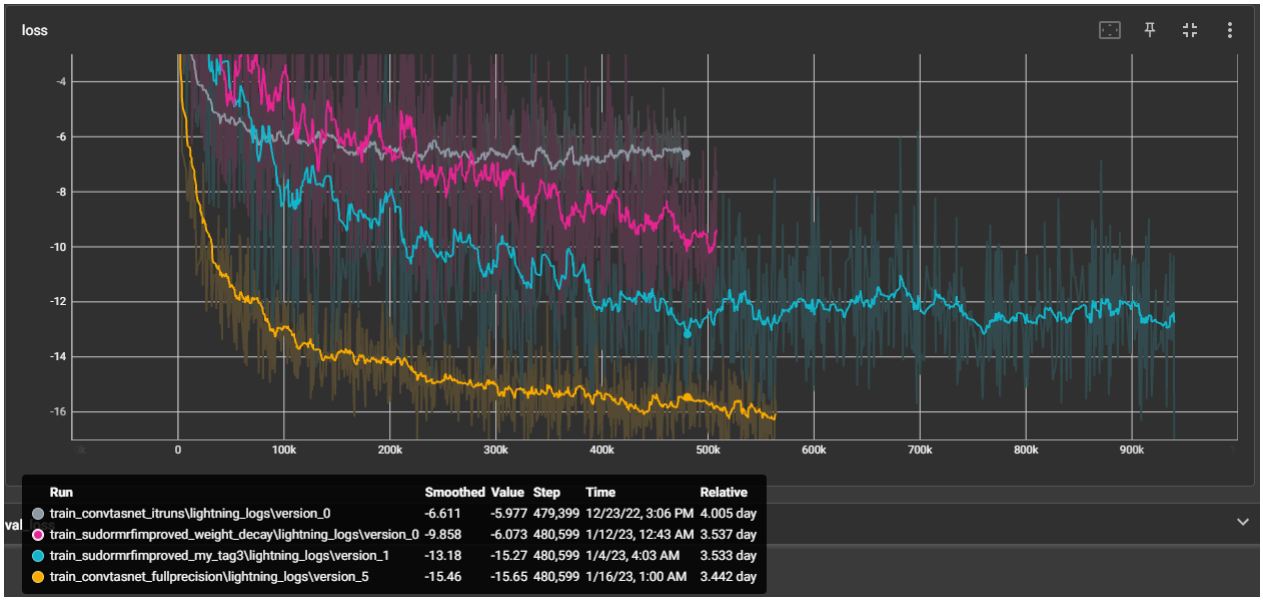
Ισχύει ότι αναφέραμε παραπάνω για τις παραμέτρους το σετ δεδομένων τον στόχο βελτιώσης και το σφάλμα. Ωστόσο υπάρχει διαφορά στο μοντέλο. Πραγματοποιήθηκαν 2 πειράματα για το convtasnet ένα με precesion=16 (είναι ο αριθμός των bit που συνήθως είναι 32 που είναι διαθέσιμος για αποθήκευσή και προφανώς επηρεάζει τις πράξεις) και ένα με κανονική precesion.



Σχήμα 5.15: Χρόνος εκπαίδευσης

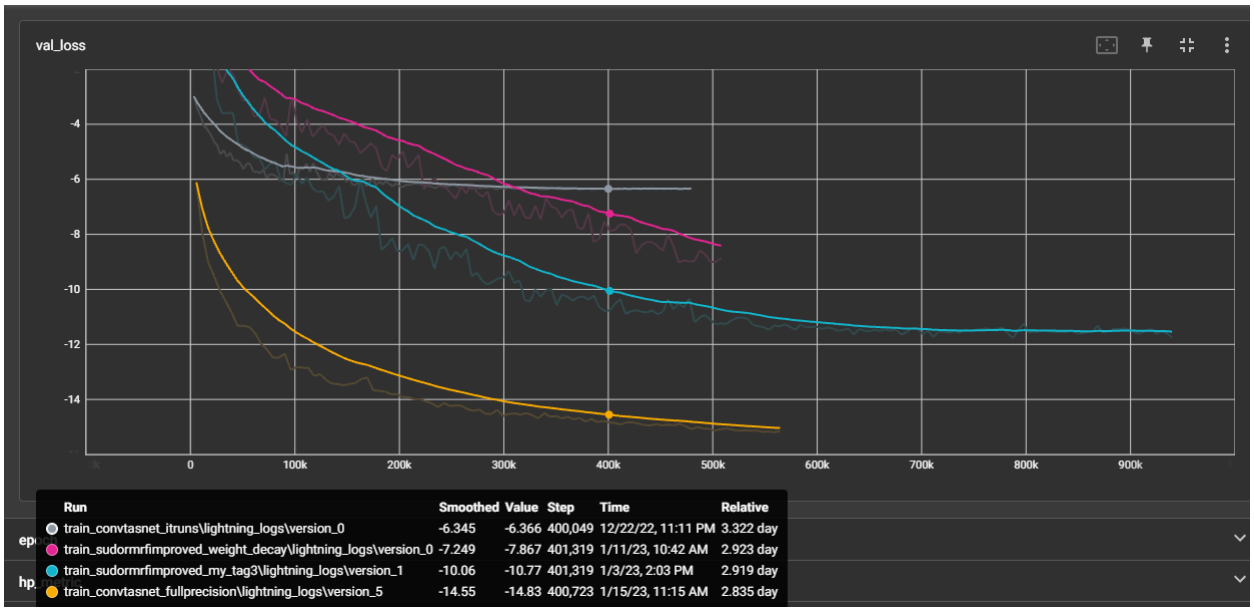
Βλέπουμε στο 5.15 το μοντέλο Convtasnet με διπλάσιες παραμέτρους από το μοντέλο SudoImprovedNet στην περίπτωση του χαμηλότερου precision είναι σαφώς γρηγορότερο ωστόσο όπως θα δούμε και στα α-

ποτελέσματα αυτό δεν σημαίνει και καλύτερο. Εδώ να αναφέρω ότι ο λόγος που γίνεται γρηγορότερο το μοντέλο είναι πως με την μείωση του χώρου αποθήκευσης είναι πιο εύκολη η τροφοδότηση δεδομένων με μεγαλύτερο batch_size.



Σχήμα 5.16: Πορεία σφαλμάτων εκπαίδευσης

Σε αυτό το σχήμα 5.16 βλέπουμε την πορεία των σφαλμάτων να αποδεικνύει ότι γρηγορότερος δεν σημαίνει καλύτερος άλλα το ανάστροφο (πράγμα το οποίο δεν είναι απαραίτητο) η ανάλυση μεταξύ των σφαλμάτων SudoImprovedNet έχει γίνει προηγουμένως και σε αυτή την περίπτωση βλέπουμε το μοντέλο Convtasnet να έχει την καλύτερη απόδοση.



Σχήμα 5.17: Πορεία σφαλμάτων επικύρωσης

Το ίδιο ακριβώς φαινόμενο αντιλαμβανόμαστε και στο σφάλμα επικύρωσης πράγμα που σημαίνει ότι κατά την γενίκευση της εκπαίδευσης δεν είχαμε μεγάλη εξάρτηση σε κάποιο μοντέλο από τα δεδομένα.

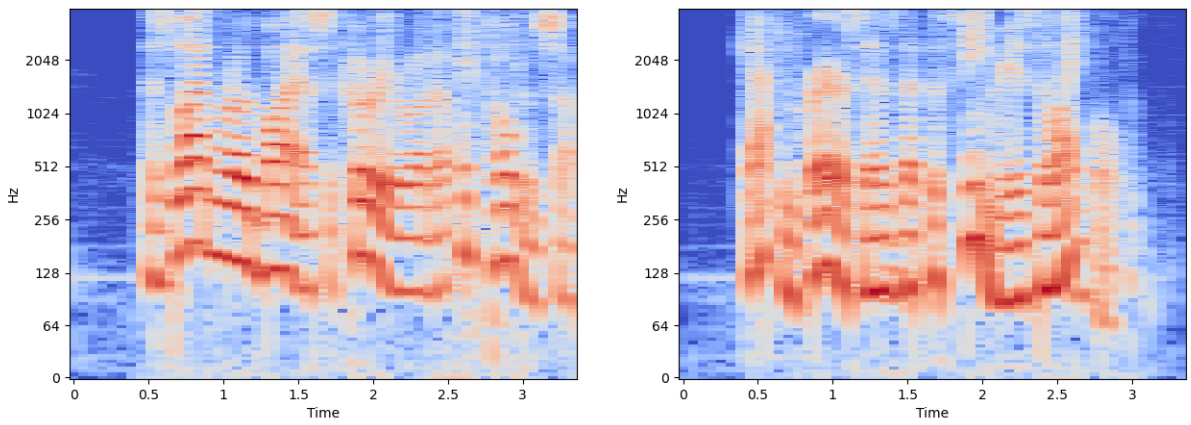
EXPIRIMENT	SISDR
Convtasnet_full_precision	15.532253124789024
SudoImprovedNet	13.668311493335292
SudoImprovedNetWeighted	11.074471914462745
Convtasnet_16_precision	6.862133430547275

Πίνακας 5.1: Πίνακας αποτελεσμάτων SISDR

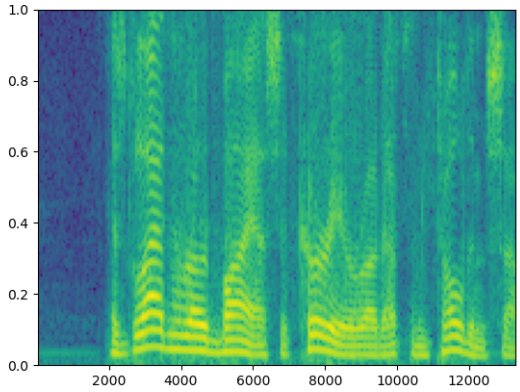
Σύμφωνα με τα αποτελέσματα των πειραμάτων κατάφερα να αναπαράγω τις αποδόσεις του μοντέλου Convtasnet καθώς και να αποδείξω ότι το SuDORMRFImprovednet είναι ένα ικανό μοντέλο για το σετ δεδομένων Libri2mix όπως βλέπουμε στον πίνακα κατάταξης 5.3. Περαιτέρω πιθανές βελτιώσεις μπορείτε να βρείτε στο τέλος του υποκεφαλαίου 5.5.

5.4.3 Σύγκριση ζωντανής ροής

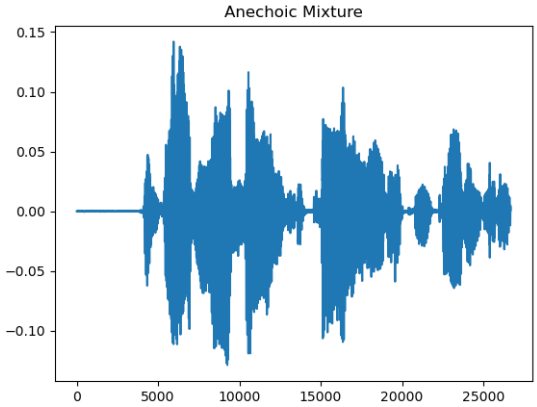
Σε αυτόν τον υπερσύνδεσμο `https://github.com/miaris98/SourceSeparation` μπορείτε να βρείτε τα βάρη εκπαίδευσης των μοντέλων (Sepformer [44],ConvTasNet [35],DualpathRnn [36],SuDORMRFImprovedne [96]) καθώς και να τα χρησιμοποιήσετε για διαχωρισμό μοντέλων. Επιπλέον παρέχονται τα παρακάτω διαγράμματα για καλύτερη εικόνα (προτείνεται η χρήση του MiniLibrix [115] για δοκιμές). Επιπλέον μπορείτε να υπολογίσετε την μετρική Si-SDR σε περιπτώσεις που παρέχονται οι πηγές.



Σχήμα 5.18: Spectogram από τις δύο πηγές (πρόβλεψη δικτύου)



Σχήμα 5.19: Spectogram πρώτης πηγής (πρόβλεψη δικτύου)



Σχήμα 5.20: Ένταση ήχου

Καθώς και τα αποτελέσματα για την εκάστοτε μείξη

```
metrics for this instance
si_sdr for sepformer:tensor(16.2615)
si_sdr for convtasnet:tensor(17.6684)
si_sdr for dualpathrnn:tensor(13.6082)
si_sdr for sudo:tensor(16.4533)
```

Σχήμα 5.21: Πορεία σφαλμάτων επικύρωσης

Τα αποτελέσματα εδώ δεν συμπίπτουν με τα ολικά αποτελέσματά εκπαίδευσης γιατί μερικά μοντέλα δεν έχουν εκπαιδευτεί στο librimix.

5.5 Συμπεράσματα και πιθανές βελτιώσεις

Το πεδίο του διαχωρισμού των πηγών είναι ένα συνεχώς εξελισσόμενο πεδίο μελέτης. Η ικανότητα διάκρισης μεμονωμένων πηγών από ένα μείγμα σημάτων έχει πολλές πιθανές χρήσεις σε τομείς

όπως η επεξεργασία ομιλίας, η παραγωγή μουσικής και η βελτίωση του σήματος ήχου. Οι τρέχουσες τεχνικές διαχωρισμού πηγών τελευταίας τεχνολογίας αποδίδουν πολλά υποσχόμενα αποτελέσματα, αλλά υπάρχει ακόμη περιθώριο βελτίωσης σε ορισμένα σενάρια. Οι προσεγγίσεις βαθιάς μάθησης έχουν βελτιώσει σημαντικά την απόδοση των αλγορίθμων διαχωρισμού πηγών τα τελευταία χρόνια. Ωστόσο, λόγω της δυσκολίας και της διακύμανσης των σημάτων του πραγματικού κόσμου, ο διαχωρισμός της πηγής παραμένει ένα δύσκολο έργο. Ο διαχωρισμός πηγής μπορεί επίσης να χρησιμοποιηθεί ως βήμα προεπεξεργασίας για άλλες εργασίες επεξεργασίας σήματος όπως η αναγνώριση ομιλίας, η μεταγραφή μουσικής και η συμπίεση ήχου.

Άλλο ένα πράγμα που θα πρέπει να προβληματίζει είναι ότι η ενσωμάτωση συνθετικών δεδομένων σε μοντέλα διαχωρισμού πηγών έχει αυξηθεί σε δημοτικότητα τα τελευταία χρόνια λόγω πλεονεκτημάτων όπως η ευκολία πρόσβασης, η ευελιξία και η δυνατότητα ελέγχου των ιδιοτήτων των δεδομένων. Ωστόσο, όταν χρησιμοποιούνται συνθετικά δεδομένα, είναι σημαντικό να λαμβάνεται υπόψη το εύρος συχνοτήτων που καλύπτουν τα δεδομένα. Όταν το εύρος συχνοτήτων περιορίζεται σε χαμηλότερο εύρος, όπως 8 kHz έως 16 kHz, η εγκυρότητα των μοντέλων διαχωρισμού πηγών μπορεί να επηρεαστεί. Μερικά από τα προβλήματα τα οποία μας δίνουν τα συνθετικά δεδομένα είναι. Πρώτον το εύρος συχνοτήτων ενδέχεται να μην αντιπροσωπεύει με ακρίβεια τα σήματα του πραγματικού κόσμου στα οποία θα εφαρμοστεί το μοντέλο. Για παράδειγμα, πολλοί φυσικοί ήχοι, όπως η ομιλία, η μουσική και οι περιβαλλοντικοί ήχοι περιέχουν ένα ευρύ φάσμα συχνοτήτων, που συχνά εκτείνονται πέρα από τα 16 kHz. Επομένως, ένα μοντέλο που εκπαιδεύεται σε συνθετικά δεδομένα με περιορισμένο εύρος συχνοτήτων μπορεί να μην έχει καλή απόδοση σε σήματα πραγματικού κόσμου που περιέχουν υψηλότερες συχνότητες. Δεύτερον, τα συνθετικά δεδομένα με περιορισμένο εύρος συχνοτήτων ενδέχεται να μην καταγράφουν τις παραλλαγές των σημάτων του πραγματικού κόσμου, με αποτέλεσμα το μοντέλο να υπερπροσαρμόζεται στα συνθετικά δεδομένα και να μην γενικεύεται καλά σε άλλα σήματα. Για παράδειγμα, τα φυσικά σήματα όπως η ομιλία και η μουσική έχουν διαφορετικά φασματικά χαρακτηριστικά ανάλογα με τα ηχεία ή τους ομιλητές και τα συνθετικά δεδομένα με περιορισμένο εύρος συχνοτήτων ενδέχεται να μην καταγράφουν αυτές τις παραλλαγές. Τρίτον, τα συνθετικά δεδομένα με περιορισμένο εύρος συχνοτήτων ενδέχεται να μην καταγράφουν τις χρονικές διακυμάνσεις των σημάτων του πραγματικού κόσμου. Για να μετριαστούν αυτά τα ζητήματα, είναι σημαντικό να χρησιμοποιηθούν συνθετικά δεδομένα που καλύπτουν ένα ευρύ φάσμα συχνοτήτων, ιδανικά καλύπτοντας το ίδιο εύρος με τα σήματα του πραγματικού κόσμου στα οποία θα εφαρμοστεί το μοντέλο. Επιπλέον, είναι επίσης σημαντικό να χρησιμοποιείτε συνθετικά δεδομένα που καταγράφουν τις παραλλαγές των σημάτων του πραγματικού κόσμου. Αυτό μπορεί να γίνει χρησιμοποιώντας ένα ποικίλο σύνολο συνθετικών δεδομένων και ενσωματώνοντας σήματα πραγματικού κόσμου στα δεδομένα εκπαίδευσης.

Άλλη μια σημαντική πτυχή του διαχωρισμού πηγών είναι ο τρόπος που θα γίνει η αξιολόγηση της απόδοσης των μεθόδων. Δεδομένου ότι ο ποιο διαδεδομένος τρόπος μέτρησης ποιότητάς Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) δεν είναι κατάλληλος για σήματα με υψηλό δυναμικό εύρος, καθώς μπορεί να οδηγήσει σε υπερεκτίμηση της παραμόρφωσης. Ως εκ τούτου, είναι σημαντικό να δημιουργηθούν αντικειμενικές μετρήσεις που μπορούν να αξιολογήσουν με ακρίβεια την αποτελεσματικότητα των μεθόδων διαχωρισμού πηγής. Ένας καλύτερος υπολογισμός ποιότητας ήχου συνεπάγεται με ένα καλύτερο σφάλμα το οποίο μπορεί να χρησιμοποιηθεί για καλύτερη μάθηση πράγμα που θα συνεπάγεται με αύξηση της ανθεκτικότητας των μοντέλων. Άλλες τεχνικές που μπορούν να χρησιμοποιηθούν για την αύξηση της ανθεκτικότητας των μοντέλων διαχωρισμού πηγών, είναι:

Αύξηση δεδομένων και πολυπλοκότητας αλγορίθμου: Ένας τρόπος για να αυξήσετε την ανθεκτικότητα των μοντέλων διαχωρισμού πηγών είναι να αυξήσετε τα δεδομένα εκπαίδευσης προσθέτοντας θόρυβο ή άλλες παραμορφώσεις στα σήματα. Αυτό μπορεί να βοηθήσει το μοντέλο να μάθει να διαχωρίζει τις πηγές ακόμη και με την παρουσία τέτοιων παραμορφώσεων. Με περισσότερα δεδομένα μπορούν να εκπαιδευτούν αλγόριθμοι μεγαλύτερης πολυπλοκότητας και βάθους χωρίς να εμπίπτουν στο πρόβλημά της υπερβολικής εξάρτησης στα δεδομένα.

Εκπαίδευση με αντίπαλο: Η εκπαίδευση με αντιπάλους είναι μια τεχνική όπου το μοντέλο εκπαιδεύεται ώστε να είναι ανθεκτικό σε αντίθετα παραδείγματα, τα οποία είναι παραδείγματα ειδικά σχεδιασμένα για να ξεγελάσουν το μοντέλο. Αυτά τα παραδείγματα πολλές φορές παράγονται από ένα άλλον αλγόριθμο που θα εκπαιδευτεί με στόχο να βελτιώσει την πτώση της απόδοσής του πρώτου αλγορίθμου. Το πρόβλημα με αυτή την προσέγγιση είναι ότι δεν είναι πολύ οικονομική και δεν μπορεί να εγγυηθεί ότι ο αντίπαλος θα φτιάχνει δεδομένα που θα μοιάζουν με τα αληθινά.

Μέθοδοι συνόλου: Οι μέθοδοι συνόλου συνδυάζουν τις προβλέψεις πολλαπλών μοντέλων για να βελτιώσουν την ανθεκτικότητα του διαχωρισμού της πηγής. Αυτό μπορεί να βοηθήσει το μοντέλο να είναι πιο ανθεκτικό σε θόρυβο και άλλες παραμορφώσεις και να γενικεύεται καλύτερα σε άλλα σήματα.

Προσαρμοστικές μέθοδοι: Οι προσαρμοστικές μέθοδοι είναι εκείνες που μπορούν να προσαρμοστούν σε διαφορετικούς τύπους παραμορφώσεων και θορύβου. Αυτά μπορούν να επιτευχθούν με την ενσωμάτωση προηγούμενων γνώσεων ή πληροφοριών για συγκεκριμένο τομέα στη διαδικασία διαχωρισμού ή με την ανάπτυξη μεθόδων που μπορούν να προσαρμοστούν σε διαφορετικούς τύπους παραμορφώσεων.

Συγκεκριμένα για βελτίωση των πειραμάτων που έγιναν σε αυτή την πτυχιακή εργασία θα μπορούσαν να γίνουν οι εξής περαιτέρω έρευνες. Χρήση επιπλέον σετ δεδομένων όπως Wsj2mix , WHAM! [116]. Ανάπτυξη και δοκιμή επιπλέον αλγορίθμων καθώς και συνδυασμός μεθόδων με την χρήση ψηφοφορίας. Επιπλέον μπορεί να εφαρμοστεί τεχνική pruning που στην ουσία βλέπει ποιοι νευρώνες δεν επιδρούν πάνω στην έξοδο και βάσει κάποιου κατωφλίου, τους καταστρέφει. Επίσης μπορεί να χρησιμοποιηθεί κάποια τεχνική για hyperparameter tuning δηλαδή τεχνική εύρεσης ιδανικών υπερπαραμέτρων. Αυτό μπορεί να γίνει με Grid search, Random search, Bayesian optimization. Επιπρόσθετα τα μοντέλα που είδαμε στα πειράματα έχουν το πολύ 5 εκατομμύρια παραμέτρους ενώ το καλύτερο μοντέλο στο Wsj2mix έχει περίπου 25 εκατομμύρια. Οπότε ενδεχομένως ακόμη και με την διατήρηση ίδιας αρχιτεκτονικής η απόδοσή μπορεί να αυξηθεί μόνο με την αύξησή των περαιτέρω μπλοκ μέσα στην αρχιτεκτονική. Επειδή τα περισσότερα δίκτυα εφαρμόζουν μόνο ενός τύπου κωδικοποιημένων διανυσμάτων (embeddings) θα μπορούσαν επιπλέον να προσθέσουν πολλών τύπων κωδικοποιημένα διανύσματα (embeddings) που εξάγονται με διαφορετικό τρόπο και να γίνονται συνένωση τους πριν την τροφοδοσία στο σύστημα εκτίμησης μάσκας. Τέλος υπάρχει πολύ πρόσφορο έδαφος για έρευνα στον τομέα του διαχωρισμού πηγών και πολλές πιθανές βελτιώσεις που μπορούν να πραγματοποιηθούν.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] J.-L. Lacoume and P. Ruiz, "Sources indentification: A solution based on the cumulants," pp. 199 – 203, 09 1988.

[2] J.-F. Cardoso, "Source separation using higher order moments," in *International Conference on Acoustics, Speech, and Signal Processing,*, pp. 2109–2112 vol.4, 1989.

[3] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *Trans. Sig. Proc.*, vol. 45, p. 434–444, feb 1997.

[4] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994. Higher Order Statistics.

[5] A. J. Bell and T. J. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 11 1995.

[6] A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 43, no. 11, pp. 894–906, 1996.

[7] S.-i. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation.," vol. 8, pp. 757–763, 01 1995.

[8] A. Hyvärinen and E. Oja, "A Fast Fixed-Point Algorithm for Independent Component Analysis," *Neural Computation*, vol. 9, pp. 1483–1492, 07 1997.

[9] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, no. 4, pp. 112–114, 1997.

[10] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Transactions on Signal Processing*, vol. 45, no. 7, pp. 1712–1725, 1997.

[11] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[12] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 4, pp. 1941–1944 vol.4, 1998.

[13] L. Molgedey and H. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical review letters*, vol. 72, pp. 3634–3637, 07 1994.

[14] S. Ikeda and N. Murata, "An approach to blind source separation of speech signals," 09 1998.

[15] A. Belouchrani and M. Amin, "Blind source separation using time-frequency distributions: Algorithm and asymptotic performance.," *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 05 1997.

[16] D. Schobben, K. Torkkola, and P. Smaragdis, “Evaluation of blind signal separation methods,” pp. 261–266, 03 2000.

[17] A. Taleb and C. Jutten, “Source separation in post-nonlinear mixtures,” *Signal Processing, IEEE Transactions on*, vol. 47, pp. 2807 – 2820, 11 1999.

[18] L. Parra and C. Spence, “Convolutive blind separation of non-stationary sources,” *Speech and Audio Processing, IEEE Transactions on*, vol. 8, pp. 320 – 327, 06 2000.

[19] S. Roweis, “One microphone source separation,” in *Advances in Neural Information Processing Systems* (T. Leen, T. Dietterich, and V. Tresp, eds.), vol. 13, MIT Press, 2000.

[20] P. Bofill and M. Zibulevsky, “Underdetermined blind source separation using sparse representations,” *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[21] D.-T. Pham and J.-F. Cardoso, “Blind separation of instantaneous mixtures of nonstationary sources,” *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[22] M. Plumbley, “Algorithms for nonnegative independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 14, no. 3, pp. 534–543, 2003.

[23] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[24] H. Sawada, R. Mukai, S. Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 530–538, sep 2004.

[25] M. Schmidt and O. Rasmus, “Single-channel speech separation using sparse non-negative matrix factorization,” 01 2006.

[26] P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[27] K. Taesu, H. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 70 – 79, 02 2007.

[28] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, may 2014.

[29] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE, dec 2014.

[30] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” in *Interspeech 2016*, ISCA, sep 2016.

[31] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” 2016.

[32] M. Kolbaek, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," vol. 25, p. 1901–1913, oct 2017.

[33] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," 11 2017.

[34] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, mar 2017.

[35] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256–1266, aug 2019.

[36] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," pp. 46–50, 05 2020.

[37] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," pp. 2642–2646, 10 2020.

[38] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, JMLR.org, 2020.

[39] Z. Shi, R. Liu, and J. Han, "La furca: Iterative context-aware end-to-end monaural speech separation based on dual-path deep parallel inter-intra bi-lstm with attention," 01 2020.

[40] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixture invariant training," NIPS'20, (Red Hook, NY, USA), Curran Associates Inc., 2020.

[41] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-to-end training of time domain audio separation and recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, may 2020.

[42] W. Zhang, X. Chang, Y. Qian, and S. Watanabe, "Improving end-to-end single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1385–1394, 2020.

[43] Y. Liu, M. Delfarah, and D. Wang, "Deep casa for talker-independent monaural speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6354–6358, 2020.

[44] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, 2021.

[45] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[46] S. Dovrat, E. Nachmani, and L. Wolf, "Many-speakers single channel speech separation with optimal permutation training," in *Interspeech*, 2021.

[47] M. Lam, J. Wang, D. Su, and D. Yu, "Effective low-cost time-domain audio separation using globally attentive locally recurrent networks," pp. 801–808, 01 2021.

[48] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5759–5763, 2021.

[49] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang, N. Kamo, M. Hira, T. Hayashi, C. Boeddeker, Z. Chen, and S. Watanabe, "ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, jan 2021.

[50] J. Mun, "Understanding the forecast statistics and four moments (4p)," 05 2015.

[51] D. R. Brillinger, *Time series: Data analysis and theory*. Society for Industrial and Applied Mathematics, 2001.

[52] B. Rajoub, "Chapter 2 - characterization of biomedical signals: Feature engineering and extraction," in *Biomedical Signal Processing and Artificial Intelligence in Healthcare* (W. Zgallai, ed.), Developments in Biomedical Engineering and Bioelectronics, pp. 29–50, Academic Press, 2020.

[53] "Chapter 11 - time-frequency synthesis and filtering," in *Time-Frequency Signal Analysis and Processing (Second Edition)* (B. Boashash, ed.), pp. 637–691, Oxford: Academic Press, second edition ed., 2016.

[54] W. Heisenberg, "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik," *Zeitschrift fur Physik*, vol. 43, pp. 172–198, Mar. 1927.

[55] S. Krishnan, "5 - advanced analysis of biomedical signals," in *Biomedical Signal Analysis for Connected Healthcare* (S. Krishnan, ed.), pp. 157–222, Academic Press, 2021.

[56] Z. Shi, R. Liu, and J. Han, "Lafurca: Iterative refined speech separation based on context-aware dual-path parallel bi-lstm," 2020.

[57] S. Sra and I. Dhillon, "Generalized nonnegative matrix approximations with bregman divergences," in *Advances in Neural Information Processing Systems* (Y. Weiss, B. Schölkopf, and J. Platt, eds.), vol. 18, MIT Press, 2005.

[58] D. Wang and G. Brown, "Computational auditory scene analysis: Principles, algorithms and applications," *IEEE Transactions on Neural Networks*, vol. 19, p. 199, 07 2008.

[59] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with ℓ0-constraints," *Neurocomputing*, vol. 80, pp. 38–46, Mar. 2012.

[60] J. Eggert and E. Körner, “Sparse coding and nmf,” vol. 4, pp. 2529 – 2533 vol.4, 08 2004.

[61] K. P. F.R.S., “Liii. on lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[62] H. Hotelling, “Analysis of a complex of statistical variables into principal components.,” *Journal of Educational Psychology*, vol. 24, pp. 498–520, 1933.

[63] I. T. Jolliffe, *Principal component analysis. 2nd ed.* Springer-Verlag, 2002.

[64] K. Diamantaras, “Artificial neural networks,” 2007.

[65] A. F. A. Fernandes, J. R. R. Dórea, and G. J. d. M. Rosa, “Image analysis and computer vision applications in animal sciences: An overview,” *Frontiers in Veterinary Science*, vol. 7, 2020.

[66] S. Haykin, *Neural networks and learning machines*. Pearson India Education Services Pvt. Ltd, 3rd ed., 2021.

[67] O. Calin, *Deep Learning Architectures: A Mathematical Approach*. Springer Publishing Company, Incorporated, 1st ed., 2020.

[68] G. V. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, 1989.

[69] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, “Multilayer feedforward networks with a nonpolynomial activation function can approximate any function,” *Neural Networks*, vol. 6, no. 6, pp. 861–867, 1993.

[70] A. Pinkus, “Approximation theory of the mlp model in neural networks,” *Acta Numerica*, vol. 8, p. 143–195, 1999.

[71] M. Gupta, N. Sinha, and K. Naresh, “series in engineering,” in *Soft Computing and Intelligent Systems: Theory and Applications*, Academic Press, 2000.

[72] P. J. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974.

[73] D. E. Rumelhart and J. L. McClelland, *Learning Internal Representations by Error Propagation*, pp. 318–362. 1987.

[74] A. LeNail, “Nn-svg: Publication-ready neural network architecture schematics,” *Journal of Open Source Software*, vol. 4, no. 33, p. 747, 2019.

[75] R. Rojas, *Neural Networks: A Systematic Introduction*. Berlin, Germany: Springer, 1996.

[76] R. Kneusel, *Math for deep learning: What you need to know to understand neural networks*. San Francisco, CA: No Starch Press, 2021.

[77] P. Murugan, “Feed forward and backward run in deep convolution neural network,” *ArXiv*, vol. abs/1711.03278, 2017.

[78] R. Hahnloser and H. S. Seung, "Permitted and forbidden sets in symmetric threshold-linear networks," in *Advances in Neural Information Processing Systems* (T. Leen, T. Dietterich, and V. Tresp, eds.), vol. 13, MIT Press, 2000.

[79] R. Hahnloser, R. Sarpeshkar, M. Mahowald, R. Douglas, and H. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, pp. 947–51, 07 2000.

[80] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," 2017.

[81] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[82] A. K. Bhoi, P. K. Mallick, C.-M. Liu, and V. E. Balas, eds., *Bio-inspired Neurocomputing*. Springer Singapore, 2021.

[83] E. Zvornicanin, "Differences between bidirectional and unidirectional lstm," Nov 2022.

[84] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[85] T. Toharudin, R. Pontoh, R. Caraka, S. Zahroh, Y. Lee, and R.-C. Chen, "Employing long short-term memory and facebook prophet model in air temperature forecasting," *Communication in Statistics- Simulation and Computation*, 01 2021.

[86] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions on*, vol. 45, pp. 2673 – 2681, 12 1997.

[87] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005. IJCNN 2005.

[88] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[89] D. Obukhov, "Breakthroughs in speech recognition achieved with the use of transformers," Mar 2021.

[90] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," 2020.

[91] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[92] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr - half-baked or well done?," 2018.

[93] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," vol. 15, p. 1929–1958, jan 2014.

[94] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[95] H.-S. Choi, H. Heo, J. H. Lee, and K. Lee, "Phase-aware single-stage speech denoising and dereverberation with u-net," 2020.

[96] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, sep 2020.

[97] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," 2015.

[98] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," 2018.

[99] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013.

[100] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[101] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, may 2020.

[102] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *Advances in Neural Information Processing Systems* (D. Touretzky, ed.), vol. 2, Morgan-Kaufmann, 1989.

[103] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[104] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.

[105] Y. Luo, C. Han, and N. Mesgarani, "Ultra-lightweight speech separation via group communication," 2020.

[106] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," 2015.

[107] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[108] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020.

[109] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech 2015*, pp. 3586–3589, 2015.

[110] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[111] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," 3 2019.

[112] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[113] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021. arXiv:2106.04624.

[114] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[115] C. Joris and P. Manuel, "Minilibrimix dataset," June 2020.

[116] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. Interspeech*, Sept. 2019.

[117] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank design for end-to-end speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, may 2020.

[118] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Chapter 3 - features for content-based audio retrieval," in *Advances in Computers: Improving the Web*, vol. 78 of *Advances in Computers*, pp. 71–150, Elsevier, 2010.

[119] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. Boston: McGraw-Hill, 4th ed ed., 2002.

[120] K. Fukushima, "Cognitron: A self-organizing multilayered neural network," *Biol. Cybern.*, vol. 20, p. 121–136, sep 1975.