Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

τομεας σηματων, ελεγχου και ρομποτικης

Ημι-Επιβλεπόμενη Αποθορυβοποίηση Σήματος Φωνής μέσω Τεχνικών Διαχωρισμού Πηγών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Δημήτριου Μπράλιου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΛΟΓΟΥ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ

Αθήνα, Ιούλιος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Ημι-Επιβλεπόμενη Αποθορυβοποίηση Σήματος Φωνής μέσω Τεχνικών Διαχωρισμού Πηγών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Δημήτριου Μπράλιου

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Ιουλίου, 2022.

.....	.....	.....
Πέτρος Μαραγκός	Αθανάσιος Ροντογιάννης	Γεράσιμος Ποταμιάνος
Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Παν. Θεσσαλίας

Αθήνα, Ιούλιος 2022

.....

Δημητριος Μπραλιος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στην παρούσα διπλωματική εργασία εξετάζουμε το πρόβλημα της Αποθορυβοποίησης Σήματος Φωνής μέσω του Διαχωρισμού Πηγών. Η εξαγωγή σήματος φωνής από θορυβώδες σήμα έχει πληθώρα εφαρμογών και αποτελεί θεμελιώδες κομμάτι άλλων συστημάτων, όπως βοηθήματα ακοής και συστήματα αναγνώρισης ομιλίας. Επομένως, είναι σημαντικό να διαθέτουμε μεθόδους οι οποίες λειτουργούν αξιόπιστα σε μεγάλο εύρος καταστάσεων. Η ραγδαία ανάπτυξη των τεχνικών Βαθιάς Μάθησης έχει οδηγήσει στην ανάπτυξη κυρίως πλήρως επιβλεπόμενων μεθόδων που επιτυγχάνουν εντυπωσιακή απόδοση στο πρόβλημα. Θεωρούμε όμως το πρόβλημα στην ημι-επιβλεπόμενη περίπτωση όπου τα δεδομένα εκπαίδευσης αποτελούνται από "καθαρά" σήματα ομιλίας, ενώ κατά την αξιολόγηση οι θόρυβοι είναι άγνωστοι. Θέτουμε το πρόβλημα σε αυτή τη μορφή ώστε η μέθοδος επίλυσης που θα αναπτύξουμε να μην υποφέρει από προβλήματα γενίκευσης ως προς το είδος και το περιβάλλον θορύβου.

Επικεντρωνόμαστε στις μεθόδους των Μη Αρνητικών Αυτοκωδικοποιητών (Non Negative Autoencoders - NAE) και τις παλαιότερες μεθόδους Μη Αρνητικής Παραγοντοποίησης Πίνακα (Non Negative Matrix Factorization - NMF), τις οποίες μελετάμε διεξοδικά. Με βάση την ημι-επιβλεπόμενη μεθοδολογία με NMF για το πρόβλημα και παλαιότερη έρευνα για τα μοντέλα NAE σχεδιάζουμε και προτείνουμε ημι-επιβλεπόμενη μεθοδολογία για μοντέλα NAE. Συγκεκριμένα, η μεθοδολογία αυτή αποτελείται από δυο στάδια. Στο πρώτο στάδιο εκπαιδεύουμε ένα μοντέλο NAE σε "καθαρά" σήματα ομιλίας με στόχο την ανακατασκευή τους μέσω μιας ενδιάμεσης αναπαράστασης μικρότερης διαστατικότητας. Έπειτα, συνδυάζουμε τον αποκωδικοποιητή ομιλίας του εκπαιδευμένου μοντέλου με έναν τυχαία αρχικοποιημένο αποκωδικοποιητή θορύβου για τον διαχωρισμό, κατά τον οποίον προσαρμόζουμε κατάλληλα, μέσω ενός επαναληπτικού αλγορίθμου, τις παραμέτρους του αποκωδικοποιητή θορύβου καθώς και τις εισόδους των δυο αποκωδικοποιητών.

Στο πειραματικό μέρος της εργασίας, πρώτα εκπαιδεύουμε μοντέλα NMF και μοντέλα NAE με διάφορες μορφές, σε "καθαρά" σήματα ομιλίας και έπειτα τα συγκρίνουμε. Για την αξιολόγηση των μεθόδων χρησιμοποιούμε δυο σύνολα δεδομένων που καλύπτουν ένα μεγάλο εύρος τύπων θορύβου, με μεταβαλλόμενα επίπεδα θορύβου. Αφού αξιολογήσουμε την ημι-επιβλεπόμενη μέθοδο NMF, πραγματοποιούμε τροποποιήσεις σε αυτή που έχουν ως αποτέλεσμα την αύξηση της απόδοσης σε ορισμένες περιπτώσεις, αλλά με αυξημένο υπολογιστικό κόστος. Στη συνέχεια, πραγματοποιούμε πειράματα ώστε να ρυθμίσουμε την ημι-επιβλεπόμενη μέθοδο NAE, καταλήγοντας σε ένα συνδυασμό από ρυθμίσεις οι οποίες μεγιστοποιούν την απόδοση. Καταφέρνουμε έτσι να ρυθμίσουμε τη μέθοδο NAE ώστε να λειτουργεί ικανοποιητικά στο πρόβλημα και να φτάνει την απόδοση της NMF στο πρώτο σύνολο δεδομένων. Όμως, στο δεύτερο σύνολο δεδομένων η απόδοση της προτεινόμενης μεθόδου υστερεί σε σχέση με την NMF.

Λέξεις Κλειδιά — Αποθορυβοποίηση Σήματος Φωνής, Διαχωρισμός Πηγών, Διαχωρισμός Σημάτων Φωνής, Μη-Αρνητικοί Αυτοκωδικοποιητές, Μη-Αρνητική Παραγοντοποίηση Πίνακα, Ημι-Επιβλεπόμενη Μάθηση

Abstract

In this thesis we study the problem of Speech Enhancement via Source Separation. Methods that extract a speech signal from a noisy recording have numerous applications and play a vital role in other systems, such as hearing aids and speech recognition systems. Therefore, it is of paramount importance to have methods that operate reliably in a wide range of conditions. The rise of Deep Learning has resulted in the development of mostly fully supervised methods with impressive performance on the task. However, we study the problem in the semi-supervised case where the training data consist of "clean" speech signals only and the noise is unknown. Formulating the problem this way we ensure that the developed model will be free of generalization problems regarding noisy environments and noise type.

We focus our attention on Non Negative Autoencoders (NAE) and Non Negative Matrix Factorization (NMF) methods, both of which we study thoroughly. Based on the semi-supervised NMF approach and prior research on NAE models, we develop and propose a semi-supervised NAE approach for the task, which consists of two steps. In the first step we train a NAE model on "clean" speech signals with the objective of reconstructing them through an intermediate representation of lower dimension. Then we combine the pre-trained decoder with a randomly initialized noise decoder in order to separate a noisy signal. During separation we iteratively learn the parameters of the noise decoder as well as the inputs of the two decoders.

For the experimental part of the thesis we first train NMF and NAE models on "clean" speech signals and then we compare their performance on the separation task. For the comparison we use two datasets which cover a wide range of noise types and variable noise levels. After evaluating the semi-supervised NMF approach, we also make adjustments which result in increased performance albeit with a higher computational cost. Next we conduct tuning experiments for the semi-supervised NAE approach resulting in a combination of parameters and settings that maximize performance. We manage to tune the NAE method so that its performance on the task is satisfactory. Specifically, its performance on the first dataset is on the same level as NMF. However, on the second dataset we observe a drop in performance compared to NMF.

Keywords — Speech Enhancement, Source Separation, Speech Separation, Non-Negative Autoencoders, Non-Negative Matrix Factorization, Semi-Supervised Learning

Ευχαριστίες

Πρώτα, θέλω να ευχαριστήσω θερμά τον καθηγητή μου κ. Πέτρο Μαραγκό που με εμπιστεύθηκε και ανέλαβε την επίβλεψη αυτής της εργασίας. Ακόμη, θέλω να ευχαριστήσω τον διδακτορικό φοιτητή Χρήστο Γαρούφη για την καθοδήγησή και στήριξή του, που συνέβαλε καθοριστικά στην ολοκλήρωση της εργασίας. Επίσης, θέλω να ευχαριστήσω τον καθηγητή κ. Πάρι Σμαραγδή και τον διδακτορικό φοιτητή Θύμιο Τζίνη. Τέλος, θέλω να ευχαριστήσω τους φίλους και την οικογένειά μου για την αγάπη τους και τη συμπαράστασή τους.

Δημήτριος Μπράλιος
Ιούλιος 2022

Περιεχόμενα

Λίστα Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

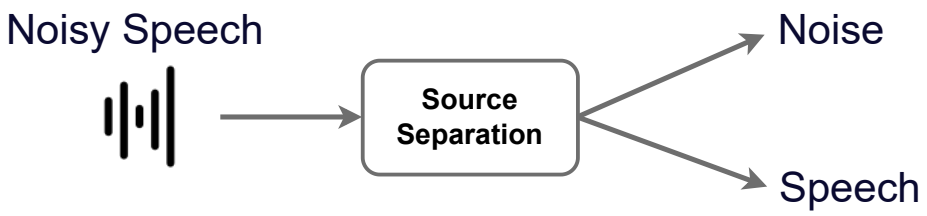
Εισαγωγή

1.1 Ορισμός του Προβλήματος

Δυο θεμελιώδη προβλήματα της Ψηφιακής Επεξεργασίας Σήματος (ΨΕΣ), τα οποία επίσης έχουν αποτελέσει και αντικείμενα έρευνας στον τομέα της Μηχανικής Μάθησης είναι ο Διαχωρισμός Πηγών (Source Separation) και η Αποθορυβοποίηση Σήματος Φωνής (Speech Enhancement).

Στο πρόβλημα της Αποθορυβοποίησης Σήματος Φωνής δεδομένου ενός θορυβώδους σήματος φωνής στόχος είναι να εξάγουμε το "καθαρό" σήμα φωνής καθώς και να βελτιώσουμε την ποιότητά του. Συνήθως, γίνεται η υπόθεση ότι ο θόρυβος είναι προσθετικός. (Loizou 2007)

Το πρόβλημα του Διαχωρισμού Πηγών (Source Separation) ορίζεται ως το πρόβλημα της αποσύνθεσης ενός σήματος που ονομάζεται μείγμα στα επιμέρους πηγαία σήματα που το αποτελούν. Αποτελεί γενικό πρόβλημα καθώς τα σήματα που επιθυμούμε να διαχωρίσουμε μπορεί να είναι ηχητικά, εικόνες, βίντεο ή ακόμα και βιομετρικά σήματα. Για παράδειγμα, το μείγμα μπορεί να είναι το ηχητικό σήμα ενός μουσικού κομματιού και τα επιμέρους πηγαία σήματα που επιθυμούμε να εξάγουμε να είναι τα σήματα των μουσικών οργάνων που συμμετέχουν στο κομμάτι.



Σχήμα 1.1.1: Σχηματική αναπαράσταση ενός προβλήματος Αποθορυβοποίησης Σήματος Φωνής μέσω του Διαχωρισμού Πηγών. Ως είσοδο λαμβάνουμε ένα θορυβώδες σήμα φωνής το οποίο διαχωρίζουμε σε δυο πηγές, ένα εκτιμώμενο σήμα φωνής και το εκτιμώμενο σήμα θορύβου.

Επομένως, μπορούμε να θεωρήσουμε το πρόβλημα της Αποθορυβοποίησης Σήματος Φωνής ως μια ειδική περίπτωση του προβλήματος του Διαχωρισμού Πηγών. Θεωρώντας το θορυβώδες σήμα ως το προσθετικό μείγμα ενός "καθαρού" σήματος φωνής και ενός σήματος θορύβου, στόχος μας είναι να εξάγουμε αυτό το "καθαρό" σήμα φωνής. Αυτή η περίπτωση του προβλήματος μερικές φορές ονομάζεται Διαχωρισμός Φωνής (Speech Separation).

Στον τομέα της επεξεργασίας ηχητικών σημάτων τόσο το πρόβλημα του Διαχωρισμού Πηγών όσο και το πρόβλημα Αποθορυβοποίησης Σήματος Φωνής έχουν ιδιαίτερη σημασία. Η ανάγκη της αφαίρεσης θορύβου από σήματα φωνής εμφανίζεται συχνά, σε εφαρμογές τηλεπικοινωνίας, ακουστικά βαρηκοΐας καθώς και ως βήμα προεπεξεργασίας σε εφαρμογές όπως η Αναγνώριση Ομιλητή και Αναγνώριση Ομιλίας (Loizou 2007). Συνεπώς, είναι σημαντικό να διαθέτουμε μεθόδους οι οποίες λειτουργούν αξιόπιστα σε μεγάλο εύρος καταστάσεων και δεν αποτυγχάνουν καταστροφικά.

Στην εργασία αυτή εμβαθύνουμε κυρίως σε μεθόδους που έχουν αναπτυχθεί για το πρόβλημα του Διαχωρισμού Πηγών, τις εφαρμόζουμε για την επίλυση του προβλήματος Αποθορυβοποίησης Σήματος Φωνής ως ειδική περίπτωση του προβλήματος του Διαχωρισμού Πηγών όπως φαίνεται στο Σχήμα 1.1.1. Επισημαίνουμε ότι, υποθέτουμε ότι ο θόρυβος είναι προσθετικός και ασχολούμαστε με την περίπτωση που τα ηχητικά σήματα αποτελούνται από μόνο ένα κανάλι.

1.2 Κίνητρο της Εργασίας

Μια ερευνητική κατεύθυνση που έχει συγκεντρώσει μεγάλη ώθηση στην ερευνητική κοινότητα είναι η Βαθιά Μάθηση (Deep Learning) που βασίζεται στα Νευρωνικά Δίκτυα (Neural Networks). Τα Νευρωνικά Δίκτυα έχουν χρησιμοποιηθεί σε μεγάλο εύρος ερευνητικών πεδίων και το πεδίο της Ψηφιακής Επεξεργασίας Σήματος δεν αποτελεί εξαίρεση. Συγκεκριμένα, μεγάλη πρόοδος έχει σημειωθεί με τη χρήση αυτών των μεθόδων

στα προβλήματα του Διαχωρισμού Πηγών και Αποθορυβοποίησης Σήματος Φωνής που ορίσαμε. Συνήθως, εκπαιδεύονται με πλήρη επίβλεψη σε μεγάλο όγκο δεδομένων, τα οποία αποτελούνται από τα δεδομένα εισόδου και τις αντίστοιχες επιθυμητές εξόδους. Για παράδειγμα, στο πρόβλημα της Αποθορυβοποίησης Σήματος Φωνής τα δεδομένα εισόδου είναι τα θορυβώδη σήματα και τα επιθυμητά σήματα εξόδου τα καθαρά σήματα φωνής.

Ωστόσο, παρά τα σημαντικά επιτεύγματα των συγκεκριμένων μεθόδων η εκπαίδευσή τους με τη συγκεκριμένη μεθοδολογία έχει ορισμένα μειονεκτήματα.

- Η εκπαίδευση των Νευρωνικών Δικτύων απαιτεί μεγάλο όγκο δεδομένων εκπαίδευσης, ώστε η απόδοση στο πρόβλημα να είναι ικανοποιητική.
- Συνήθως τα εκπαιδευμένα μοντέλα που προκύπτουν είναι λειτουργικά μόνο για το πρόβλημα και τις συνθήκες που έχουν εκπαιδευτεί. Μπορεί δηλαδή ένα μοντέλο να έχει εκπαιδευτεί να εξάγει το καθαρό σήμα φωνής από θορυβώδη σήματα ομιλίας σε εξωτερικούς χώρους, αλλά να αποτυγχάνει στην περίπτωση που το σήμα εισόδου προέρχεται από διαφορετικό περιβάλλον θορύβου, όπως το εσωτερικό ενός κτηρίου.
- Ακόμη, σπάνια μπορούν να επαναχρησιμοποιηθούν σε παρόμοια προβλήματα χωρίς περαιτέρω εκπαίδευση. Έτσι, το παραπάνω εκπαιδευμένο μοντέλο δεν μπορεί να χρησιμοποιηθεί για το πρόβλημα της επέκτασης φάσματος σημάτων φωνής.
- Τέλος, είναι μέθοδοι που δίνουν έξοδο με ένα πέρασμα. Το γεγονός αυτό περιορίζει τυχόν προσαρμογή του μοντέλου στο κάθε δείγμα που λαμβάνει ως είσοδο.

Αντίθετα, δεν αντιμετωπίζει τα συγκεκριμένα προβλήματα η παλαιότερη μέθοδος Non Negative Matrix Factorization (NMF) που έχει χρησιμοποιηθεί εκτενώς σε προβλήματα Ψηφιακής Επεξεργασίας Σήματος συμπεριλαμβανομένων και των προβλημάτων που μας ενδιαφέρουν. Διαθέτει ακόμα το πλεονέκτημα της λειτουργίας με μειωμένη επίβλεψη. Όμως, αυτή η οικογένεια μεθόδων εμφανίζει διαφορετικά μειονεκτήματα.

- Η επέκταση της είναι δύσκολη και απαιτεί την εξαγωγή νέων εξισώσεων εκμάθησης των παραμέτρων.
- Είναι δύσκολη η αλλαγή της συνάρτησης σφάλματος που χρησιμοποιεί η μέθοδος και η χρήση σύγχρονων συναρτήσεων σφάλματος δεν έχει ερευνηθεί.
- Η απόδοση τους υστερεί σε σύγκριση με τα πιο πρόσφατα Νευρωνικά Δίκτυα.

1.3 Στόχοι και Συνεισφορά της Εργασίας

Στόχος της εργασίας είναι η μελέτη των Non Negative Autoencoders (NAE), ενός τύπου Νευρωνικού Δικτύου που έχει προταθεί ως επέκταση της NMF, με σκοπό την κατασκευή μιας μεθόδου που αντιμετωπίζει τις δυο κατηγορίες μειονεκτημάτων που απαριθμήσαμε.

Θεωρούμε το πρόβλημα στην ημι-επιβλεπόμενη περίπτωση όπου ως δεδομένα εκπαίδευσης έχουμε μονάχα "καθαρά" σήματα ομιλίας, ενώ κατά την αξιολόγηση του μοντέλου οι θόρυβοι είναι άγνωστοι. Θεωρώντας το πρόβλημα σε αυτή τη μορφή το καθιστά δυσκολότερο σε σχέση με την περίπτωση της πλήρους επίβλεψης. Όμως, η μέθοδος επίλυσης που θα αναπτύξουμε δεν θα υποφέρει από προβλήματα γενίκευσης ως προς το είδος και το περιβάλλον θορύβου. Συνεπώς, δεύτερος και κυριότερος στόχος αποτελεί ο σχεδιασμός και η δοκιμή ημι-επιβλεπόμενης μεθόδου που βασίζεται σε NAE για την επίλυση προβλημάτων Διαχωρισμού Πηγής. Η δοκιμή της μεθόδου γίνεται στα πλαίσια του προβλήματος Αποθορυβοποίηση Φωνής.

Οι κύριες συνεισφορές της παρούσας εργασίας μπορούν να συνοψιστούν στις παρακάτω τρεις κατηγορίες

- Σχεδιασμός και πρόταση ημι-επιβλεπόμενης μεθόδου βασιζόμενη σε μοντέλα NAE για το πρόβλημα Αποθορυβοποίησης Σήματος Φωνής μέσω Διαχωρισμού Πηγών.
- Προσαρμογή της προτεινόμενης μεθόδου και των υπερ-παραμέτρων της ώστε να μεγιστοποιήσουμε την απόδοση στο πρόβλημα.

- Σύγκριση της μεθόδου NAE με την προϋπάρχουσα μέθοδο NMF σε δυο σύνολα δεδομένων που καλύπτουν ένα μεγάλο εύρος τύπων θορύβου καθώς και σε μεταβαλλόμενα επίπεδα θορύβου.

1.4 Οργάνωση της Εργασίας

Η εργασία είναι οργανωμένη στα ακόλουθα κεφάλαια.

- Στο Κεφάλαιο 2 παρέχουμε το θεωρητικό υπόβαθρο Ψηφιακής Επεξεργασίας Σήματος, Μηχανικής Μάθησης, Νευρωνικών Δικτύων και της μεθόδου NMF, το οποίο είναι απαραίτητο για την ανάλυση και κατανόηση των μεθόδων που παρουσιάζουμε στη συνέχεια,
- Στο Κεφάλαιο 3 πραγματοποιούμε μια επισκόπηση της βιβλιογραφίας για το πρόβλημα του Διαχωρισμού Πηγών στην περίπτωση ηχητικών σημάτων και του προβλήματος Αποθορυβοποίησης Σήματος Φωνής. Επίσης, εξετάζουμε μεθόδους που έχουν προταθεί για το συγκεκριμένο πρόβλημα που επιχειρούμε να επιλύσουμε. Τέλος, παρουσιάζουμε σύνολα δεδομένων και τις μετρικές αξιολόγησης που θα χρησιμοποιήσουμε στο πειραματικό μέρος.
- Στο Κεφάλαιο 4 παρουσιάζουμε τους Non Negative Autoencoders (NAE), την μεθοδολογία εκπαίδευσής τους σε καθαρά σήματα ομιλίας καθώς και την προτεινόμενη ημι-επιβλεπόμενη μεθοδολογία για την επίλυση του προβλήματος διαχωρισμού.
- Στο Κεφάλαιο 5 παρουσιάζουμε το πειραματικό πλαίσιο καθώς και τα αποτελέσματα από τις μεθόδους που χρησιμοποιούμε.
- Στο Κεφάλαιο 6 εξάγουμε ορισμένα συμπεράσματα από τα πειράματα του προηγούμενου κεφαλαίου και συζητάμε πιθανά μελλοντικά βήματα και επεκτάσεις των μεθόδων.

1.5 Συμβολισμός

Στο σημείο αυτό προτού εμβαθύνουμε, αξίζει να παρουσιάσουμε το συμβολισμό που θα ακολουθήσει, ώστε ο αναγνώστης να είναι σε θέση να παρακολουθήσει το κείμενο.

Αναφερόμαστε με τα σύμβολα $\mathbb{N}, \mathbb{Z}, \mathbb{R}$ στα σύνολα των φυσικών, των ακεραίων και των πραγματικών αριθμών αντίστοιχα. Αναλυτικότερα, $\mathbb{N} = \{1, 2, 3, \dots\}$, $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ και με το σύμβολο $\mathbb{R}$ αναφερόμαστε στην ευθεία των πραγματικών αριθμών $(-\infty, \infty)$. Στα πλαίσια των μεθόδων που θα αναλύσουμε, συμβολίζουμε το σύνολο των μη αρνητικών πραγματικών αριθμών ως $\mathbb{R}_{\geq 0}$, δηλαδή $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$.

Επιπλέον συμβολίζουμε:

- τα βαθμωτά μεγέθη με μικρά γράμματα, π.χ. $a \in \mathbb{R}$,
- τα διανύσματα ή τους μονοδιάστατους πίνακες (στήλη) με μικρά γράμματα σε bold γραμματοσειρά, π.χ. $\boldsymbol{a} \in \mathbb{R}^n$,
- τους πίνακες με κεφαλαία γράμματα σε bold γραμματοσειρά, π.χ. $\mathbf{A} \in \mathbb{R}^{M \times N}$,
- πίνακα $\mathbf{A}$ με διαστάσεις $M \times N$ ως $\mathbf{A}_{(M \times N)}$,
- τον ανάστροφο ενός πίνακα $\mathbf{A}$ ως $\mathbf{A}^\intercal$
- την τιμή του πίνακα $\mathbf{A}$ στην i-οστή γραμμή και στην j-οστή στήλη ως $\mathbf{A}_{i,j}$

Για παράδειγμα,

$$a = 1.337 \qquad \boldsymbol{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 4 & 6 \end{bmatrix} \qquad \mathbf{A}_{1,2} = 2$$

Τέλος, για τις νόρμες ℓ_p για $\mathbf{x} \in \mathbb{R}^n$, ισχύει:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

με $\|\mathbf{x}\| = \|\mathbf{x}\|_2$.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Ψηφιακή Επεξεργασία Σήματος

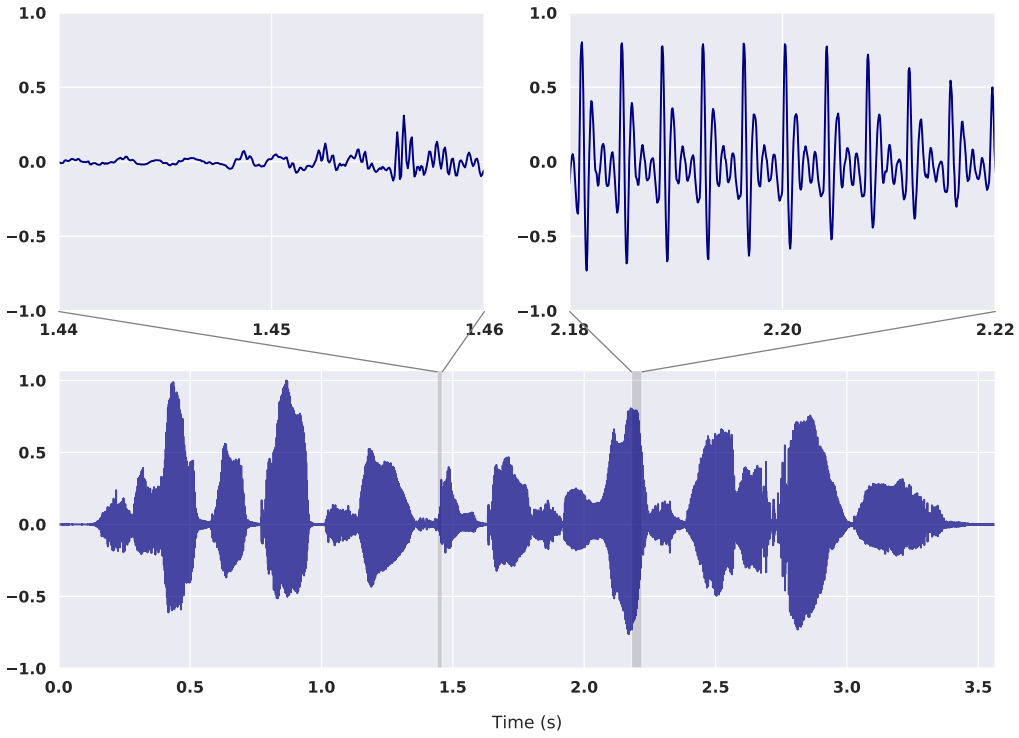
Με τον όρο σήμα (signal) αναφερόμαστε σε μια συνάρτηση μίας ή περισσοτέρων ανεξάρτητων μεταβλητών, η οποία περιέχει πληροφορία. Ένα ψηφιακό σήμα είναι μια ακολουθία διακριτών τιμών. Συνήθως εργαζόμαστε με ψηφιακά σήματα επειδή είναι δυνατή η επεξεργασία τους από ψηφιακούς υπολογιστές. Συνεπώς, η Ψηφιακή Επεξεργασία Σήματος (ΨΕΣ) αναφέρεται στην επεξεργασία ψηφιακών σημάτων με ψηφιακούς υπολογισμούς. (A. Oppenheim and Schafer 2010)

Μια χρήσιμη διάκριση των σημάτων με ανεξάρτητη μεταβλητή τον χρόνο είναι ανάμεσα σε χρονοσυνεχή και χρονοδιακριτά. Τα χρονοσυνεχή σήματα ορίζονται στο συνεχές σύνολο του χρόνου και παριστάνονται από συνεχή ανεξάρτητη μεταβλητή. Τα χρονοδιακριτά σήματα ορίζονται σε διακριτές χρονικές στιγμές, δηλαδή αναπαρίστανται ως ακολουθίες.

2.1.1 Αναπαραστάσεις Σημάτων Ήχου

Ο ήχος μεταδίδεται ως ακουστικό κύμα σε κάποιο μέσο όπως ο αέρας και αναπαρίσταται ως ηχητικό σήμα. Το ηχητικό σήμα συνήθως είναι μονοδιάστατο είτε χρονοσυνεχές είτε χρονοδιακριτό και περιγράφει τη μεταβολή της ατμοσφαιρικής πίεσης προς τον χρόνο.

Η αναπαράσταση αυτή παρόλο που είναι απλή δεν είναι ιδιαίτερα χρήσιμη για εφαρμογές επεξεργασίας ηχητικών σημάτων.



Σχήμα 2.1.1: Στο κάτω γράφημα έχουμε την πλήρη κυματομορφή ηχητικού σήματος ομιλίας, της πρότασης "She had your dark suit in greasy wash water all year" μιας γυναίκας ομιλητή. Στα δύο πάνω γραφήματα εστιάζουμε σε δυο διαφορετικά τμήματα του σήματος. Το πάνω αριστερά αποτελεί την αρχή του φωνήεντος i στη λέξη **i**n και το πάνω δεξιά αποτελεί τμήμα του φωνήεντος a στη λέξη w**a**sh.

Short-Time Fourier Transform

Συχνά τα ηχητικά σήματα μετασχηματίζονται σε αναπαραστάσεις στο πεδίο χρόνου και συχνότητας. Μια τέτοια αναπαράσταση προκύπτει από τον μετασχηματισμό Fourier βραχέος χρόνου (Short-Time Fourier Transform ή

STFT) ενός σήματος.

Ο μετασχηματισμός Fourier είναι θεμελιώδης μαθηματική ιδέα στην επεξεργασία σήματος. Πρακτικά δίνει το φάσμα των συχνοτήτων ενός σήματος, μετασχηματίζοντας το στο πεδίο της συχνότητας από το πεδίο του χρόνου.

Στην περίπτωση των χρονοσυνεχών σημάτων ο Μετασχηματισμός Fourier Συνεχούς Χρόνου (Continuous Time Fourier Transform ή CTFT) ορίζεται ως εξής:

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-j2\pi ft}\, dt$$

ενώ ο αντίστροφος μετασχηματισμός ορίζεται ως εξής:

$$x(t) = \int_{-\infty}^{+\infty} X(f) e^{j2\pi ft}\, df$$

όπου f η συχνότητα γραμμικού χρόνου σε Hertz και t ο χρόνος σε second.

Στην περίπτωση των χρονοδιακριτών σημάτων ο Μετασχηματισμός Fourier Διακριτού Χρόνου (Discrete Time Fourier Transform ή DTFT) ορίζεται ως εξής:

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x[n] e^{-j\omega n}$$

ενώ ο αντίστροφος μετασχηματισμός ορίζεται ως εξής:

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n}\, d\omega$$

όπου ω η κανονικοποιημένη συχνότητα σε ακτίνια. Επισημαίνουμε πως ο DTFT είναι περιοδικός με περίοδο 2π.

Επειδή εργαζόμαστε με πεπερασμένα ψηφιακά σήματα αντί του μετασχηματισμού Fourier συνεχούς χρόνου χρησιμοποιούμε τον Διακριτό Μετασχηματισμό Fourier (Discrete Fourier transform ή DFT). Για ψηφιακό σήμα $x[n]$ με N δείγματα, ορίζουμε τον Διακριτό Μετασχηματισμό Fourier ως εξής:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}, \quad k = 0, 1, \ldots, N-1$$

ενώ ο αντίστροφος μετασχηματισμός ορίζεται ως εξής:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{j\frac{2\pi}{N}kn}, \quad n = 0, 1, \ldots, N-1$$

Συνήθως, τα σήματα που μας ενδιαφέρουν, όπως τα ηχητικά σήματα, έχουν χρονικά μεταβαλλόμενα φασματικά χαρακτηριστικά. Ως αποτέλεσμα η εφαρμογή του μετασχηματισμού Fourier σε ολόκληρο το σήμα δε δίνει μια χρήσιμη αναπαράσταση. Για να αντιμετωπίσουμε το γεγονός αυτό, χωρίζουμε το σήμα σε παράθυρα μέσα στα οποία τα φασματικά χαρακτηριστικά δεν μεταβάλλονται πολύ και εφαρμόζουμε τον μετασχηματισμό Fourier στο καθένα από τα παράθυρα. Ορίζουμε έτσι τον μετασχηματισμό Fourier βραχέος χρόνου για χρονοδιακριτά σήματα ως εξής:

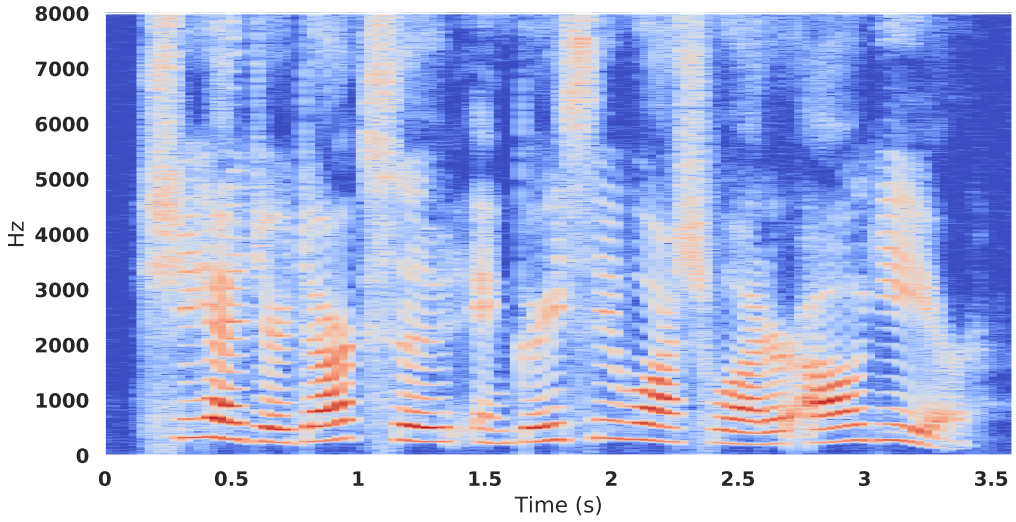
$$X_m(e^{j\omega}) = \sum_{m=-\infty}^{+\infty} x[n] w[n - mR] e^{-j\omega n}$$

όπου $w[n]$ παράθυρο μήκους M, R απόσταση μεταξύ δυο διαδοχικών παραθύρων, ενώ για κάθε δείκτη m έχουμε

τον DTFT του παραθυροποιημένου σήματος $x[n]w[n-mR]$. Πρακτικά επειδή εργαζόμαστε με ψηφιακά σήματα αντί για DTFT χρησιμοποιούμε τον DFT.

Λαμβάνοντας το λογαριθμικό μέτρο του STFT, έχουμε λοιπόν μια δισδιάστατη αναπαράσταση του σήματος που μπορεί να παρασταθεί γραφικά ως εικόνα που ονομάζουμε φασματογράφημα. Η εικόνα αυτή μας δείχνει τα φασματικά χαρακτηριστικά του σήματος και πώς αυτά μεταβάλλονται σε σχέση με τον χρόνο.

Η επιλογή μήκους παραθύρου είναι σημαντική ανάλογα την επεξεργασία που θέλουμε να πραγματοποιήσουμε. Για μικρό παράθυρο έχουμε καλή χρονική ανάλυση αλλά μέτρια συχνοτική ευκρίνεια. Αντίθετα, για μεγάλο παράθυρο έχουμε καλή συχνοτική ευκρίνεια αλλά μέτρια χρονική ανάλυση. (Rabiner and Schafer 2010)



Σχήμα 2.1.2: Φασματογράφημα του ηχητικού σήματος ομιλίας από το Σχήμα 2.1.1. Όπως είδαμε και στο Σχήμα 2.1.1 τα φωνήεντα και γενικά οι έμφωνοι ήχοι έχουν περιοδικό χαρακτήρα. Έτσι, παρατηρούμε εδώ οριζόντιες γραμμώσεις εξαιτίας των αρμονικών της θεμελιώδους συχνότητας στα τμήματα έμφωνων ήχων.

2.1.2 Επεξεργασία Weighted Overlap Add

Σε περίπτωση που θέλουμε να επεξεργαστούμε το σήμα στο πεδίο χρόνου-συχνότητας, συχνά είναι απαραίτητο να μπορούμε να μετατρέψουμε την επεξεργασμένη αυτή αναπαράσταση πίσω στο πεδίο του χρόνου. Για τον σκοπό αυτό στην εργασία αυτή χρησιμοποιούμε τη μεθοδολογία Weighted Overlap Add (WOLA), η οποία αποτελείται από τα παρακάτω βήματα (Smith 2011).

1. Χωρίζουμε το σήμα σε χρονικά παράθυρα, πολλαπλασιάζοντάς το με το μετατοπισμένο παράθυρο w. Προκύπτουν έτσι τα σήματα $x_m[n] = x[n]w[n-mR]$ όπου m ο δείκτης του παραθύρου.

2. Μετασχηματίζουμε κάθε σήμα $x_m[n]$ με τον DFT, στο $X_m[k]$.

3. Επεξεργαζόμαστε το $X_m[k]$ και παίρνουμε την τροποποιημένη αναπαράσταση $Y_m[k]$ στο πεδίο χρόνου-συχνότητας, που επιθυμούμε να τη μετατρέψουμε πίσω στο πεδίο του χρόνου.

4. Μετασχηματίζουμε κάθε στήλη του $Y_m[k]$ με τον αντίστροφο του Διακριτό Μετασχηματισμό Fourier παράγοντας τα $y_m[n]$.

5. Πολλαπλασιάζουμε τα $y_m[n]$ με το παράθυρο w.

6. Μετατοπίζουμε κάθε παραθυρωμένο σήμα $y_m^w[n]$ κατά mR και αθροίζουμε στο σήμα εξόδου $y[n]$.

Σε περίπτωση που δεν πραγματοποιήσουμε κάποια επεξεργασία στο Βήμα 3, για να έχουμε τέλεια ανακατασκευή

θα πρέπει να ισχύει το παρακάτω (Smith 2011).

$$x[n] = \sum_{m=-\infty}^{+\infty} x[n]w[n - mR]w[n - mR]$$

Το οποίο ισχύει αν και μόνο αν

$$\sum_{m=-\infty}^{+\infty} w^2[n - mR] = 1, \quad \forall n \in \mathbb{Z}$$

Συνεπώς, συνήθως επιλέγεται παράθυρο το οποίο ικανοποιεί την παραπάνω συνθήκη. Στην εργασία αυτή επιλέγουμε το παράθυρο root-Hann που ορίζεται ως εξής:

$$w[n] = \sqrt{\frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi n}{M-1}\right)} = \sin\left(\frac{\pi n}{M-1}\right), \quad n = 0, \ldots, M-1$$

2.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι ένα πεδίο έρευνας στο οποίο σχεδιάζονται και κατασκευάζονται μέθοδοι και μοντέλα τα οποία μαθαίνουν από σύνολα δεδομένων με σκοπό να επιλύουν κάποιο πρόβλημα, χωρίς να έχουν προγραμματιστεί άμεσα για την επίλυση του. Σε μια προσέγγιση Μηχανικής Μάθησης βασιζόμαστε σε ένα σύνολο δεδομένων, που ονομάζεται σύνολο εκπαίδευσης, το οποίο χρησιμοποιείται για την εκμάθηση των παραμέτρων ενός προσαρμοστικού μοντέλου (Bishop 2006). Η προσεγγίσεις αυτές έχουν πολύ μεγάλο εύρος εφαρμογών, από αναγνώριση χειρόγραφων ψηφίων από εικόνες και ανίχνευση προσώπου έως αναγνώριση ομιλίας και ομιλητή.

Μαθηματικά ένα τέτοιο μοντέλο μπορεί να εκφραστεί ως μια παραμετρική συνάρτηση $\boldsymbol{y} = f(\boldsymbol{x}; \theta)$ η οποία δέχεται ως είσοδο το διάνυσμα $\boldsymbol{x}$ και δίνει ως έξοδο το διάνυσμα $\boldsymbol{y}$. Κατά το διάστημα που ονομάζουμε φάση εκπαίδευσης χρησιμοποιούμε δεδομένα σε μορφή διανυσμάτων από το σύνολο εκπαίδευσης, ώστε να προσαρμόσουμε τις παραμέτρους θ του μοντέλου. Η προσαρμογή των παραμέτρων γίνεται ώστε η έξοδος του μοντέλου $\boldsymbol{y}$ να προσεγγίζει την επιθυμητή. Απαιτείται λοιπόν μια συνάρτηση σφάλματος η οποία αξιολογεί την έξοδο $\boldsymbol{y}$ και ποσοτικοποιεί την απόκλιση της εξόδου από την επιθυμητή έξοδο. Συνεπώς, με βάση την συνάρτηση σφάλματος μπορούμε να σχεδιάσουμε τον αλγόριθμο εκμάθησης που προσδιορίζει την μεταβολή και κατ' επέκταση εκμάθηση των παραμέτρων θ.

Οι μέθοδοι Μηχανικής Μάθησης μπορούν χονδρικά να χωριστούν στις κατηγορίες της Επιβλεπόμενης Μάθησης (Supervised Learning) και της Μη Επιβλεπόμενης Μάθησης (Unsupervised Learning), με βάση την πρόσβαση στην πληροφορία που έχουν κατά την φάση της εκμάθησης (Goodfellow, Bengio, and Courville 2016).

2.2.1 Επιβλεπόμενη Μάθηση

Στην περίπτωση της Επιβλεπόμενης Μάθησης κάθε δείγμα του συνόλου εκπαίδευσης συνοδεύεται από ένα δείγμα στόχο ή έχει επισημειωθεί με μια ετικέτα. Έτσι η συνάρτηση σφάλματος υπολογίζει το σφάλμα ανάμεσα στην έξοδο του μοντέλου και το διάνυσμα στόχο ή την ετικέτα. Στόχος είναι με βάση την εκπαίδευση το μοντέλο να δίνει την επιθυμητή έξοδο ακόμα και σε εισόδους που δε βρίσκονται στο σύνολο εκπαίδευσης αλλά προέρχονται από την ίδια κατανομή δεδομένων.

Τα προβλήματα Επιβλεπόμενης Μάθησης μπορούν να χωριστούν σε δυο κατηγορίες, τα προβλήματα ταξινόμησης (classification) και τα προβλήματα παλινδρόμησης (regression).

Ταξινόμηση

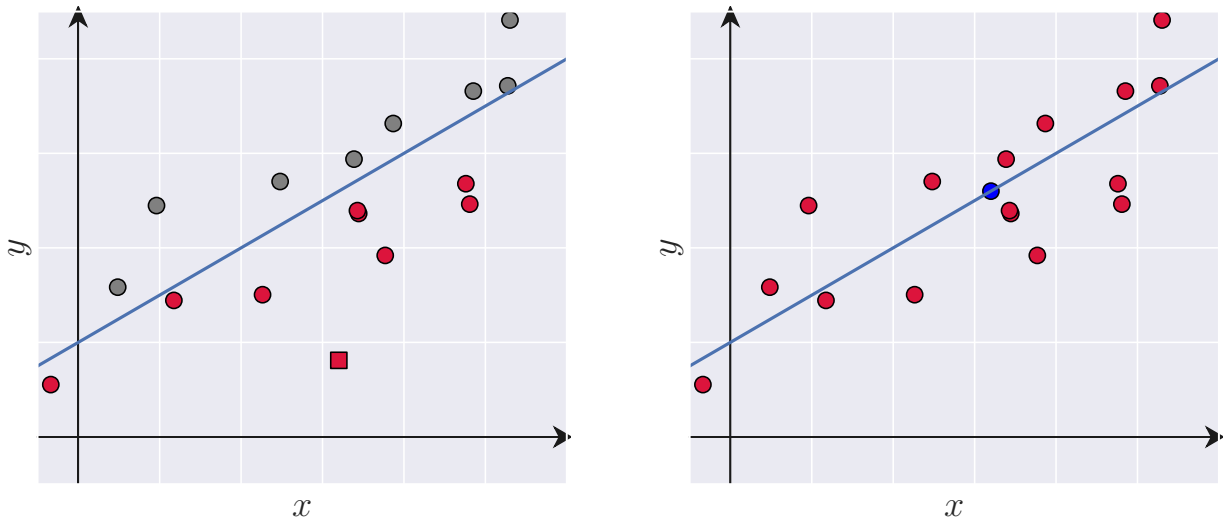
Στόχος στα προβλήματα ταξινόμησης είναι η αντιστοίχιση κάποιας ετικέτας στα δεδομένα εισόδου, δηλαδή η ταξινόμησή τους σε κατηγορίες. Όταν οι κατηγορίες είναι δυο τότε έχουμε την περίπτωση δυαδικής ταξινόμησης

(binary classification). Επιπλέον, στην περίπτωση της ταξινόμησης με πολλαπλές ετικέτες (multi-label classification) καλούμαστε να αντιστοιχίσουμε στο δείγμα εισόδου πάνω από μια ετικέτες. Ένα τυπικό πρόβλημα ταξινόμησης είναι η αναγνώριση ψηφίου από εικόνα, όπου τα ψηφία από το 0 έως το 9 αποτελούν τις κατηγορίες της ταξινόμησης.

Στο Σχήμα 2.2.1a έχουμε μια περίπτωση ταξινόμησης. Τα δεδομένα εισόδου είναι διανύσματα δυο διαστάσεων και αναπαρίστανται ως σημεία στο επίπεδο, ενώ η κατηγορία τους υποδηλώνεται με το χρώμα. Το μοντέλο είναι μια παραμετρική εξίσωση ευθείας, την οποία έχουμε μάθει από τα δεδομένα εκπαίδευσης. Όταν κληθούμε να ταξινομήσουμε κάποιο δείγμα ελέγχουμε σε ποιο από τα ημι-επίπεδα που ορίζει η ευθεία βρίσκεται και το επισημειώνουμε με την κατάλληλη ετικέτα.

Παλινδρόμηση

Στην περίπτωση των προβλημάτων παλινδρόμησης η επιθυμητή έξοδος δεν είναι κάποια ετικέτα αλλά αποτελείται από μια ή περισσότερες συνεχείς μεταβλητές (Bishop 2006). Επομένως, το μοντέλο καλείται με βάση τα χαρακτηριστικά του δείγματος εισόδου να "προβλέψει" το συνεχές διάνυσμα εξόδου. Για παράδειγμα, στο Σχήμα 2.2.1b βλέπουμε τα μονοδιάστατα δεδομένα εισόδου (τιμή του οριζόντιου άξονα) με τις αντίστοιχες τιμές στόχους (κατακόρυφος άξονας). Το μοντέλο πάλι είναι μια παραμετρική ευθεία που έχουμε μάθει από τα δεδομένα εκπαίδευσης. Όταν κληθούμε να προβλέψουμε την τιμή ενός δείγματος, η πρόβλεψη θα είναι η τιμή της ευθείας στην αντίστοιχη τετμημένη.



(a) Πρόβλημα ταξινόμησης. Το δείγμα εισόδου που αναπαρίσταται με τετράγωνο σχήμα έχει ταξινομηθεί στην κατηγορία με το χρώμα κόκκινο.

(b) Πρόβλημα παλινδρόμησης. Το μπλε σημείο υποδηλώνει μια πρόβλεψη του μοντέλου, όπου για την συγκεκριμένη τετμημένη η τιμή της πρόβλεψης ισούται με την αντίστοιχη τεταγμένη.

Σχήμα 2.2.1: Απλά προβλήματα Επιβλεπόμενης Μηχανικής Μάθησης

2.2.2 Μη-Επιβλεπόμενη Μάθηση

Στην κατηγορία της Μη-Επιβλεπόμενης Μάθησης ανήκουν τα μοντέλα τα οποία κατά την φάση εκπαίδευσης έχουν πρόσβαση μόνο στα δεδομένα εισόδου χωρίς να έχουν γνώση της επιθυμητής εξόδου. Το μοντέλο καλείται να μάθει χρήσιμες ιδιότητες και την δομή των δεδομένων εισόδου (Goodfellow, Bengio, and Courville 2016). Συνήθως, μας ενδιαφέρει η εκμάθηση της κατανομής των δεδομένων εισόδου είτε άμεσα μαθαίνοντας τις παραμέτρους μιας παραμετρικής κατανομής είτε έμμεσα κατασκευάζοντας ένα μοντέλο το οποίο μαθαίνει να παράγει δεδομένα της κατανομής των δεδομένων εκπαίδευσης.

Άλλες μέθοδοι αυτής της κατηγορίας επιλύουν προβλήματα όπως η συσταδοποίηση και η μείωση της διάστασης

των δεδομένων. Κατά την συσταδοποίηση επιθυμούμε να βρούμε μια κατηγοριοποίηση από τα δεδομένα εισόδου και με βάση αυτή να κατηγοριοποιήσουμε άλλες τυχόν εισόδους. Κατά την μείωση διάστασης μας ενδιαφέρει η εκμάθηση ενός τρόπου μετασχηματισμού των δεδομένων εισόδου σε ένα χώρο μικρότερης διάστασης διατηρώντας όμως χρήσιμα χαρακτηριστικά των δεδομένων.

Στην κατηγορία αυτή μπορεί να ανήκει τόσο η μέθοδος NMF όσο και τα μοντέλα NAE, που θα αναλύσουμε διεξοδικότερα στη συνέχεια.

2.2.3 Ημι-Επιβλεπόμενη Μάθηση

Ο όρος Ημι-Επιβλεπόμενη Μάθηση (Semi-Supervised Learning) αναφέρεται στις προσεγγίσεις που βρίσκονται ανάμεσα στις δυο αυτές κατηγορίες. Πιο συχνά χρησιμοποιείται όταν μόνο κάποιο μέρος των δεδομένων εκπαίδευσης είναι επισημειωμένα με ετικέτες ή συνοδεύονται με δεδομένα στόχους (Goodfellow, Bengio, and Courville 2016).

Στην περίπτωση μας όπου το πρόβλημα είναι ο διαχωρισμός πηγών σε μείγματα ομιλίας και θορύβου, δεν βρισκόμαστε αμιγώς στην περίπτωση της Μη-Επιβλεπόμενης Μάθησης, καθώς τα μοντέλα που χρησιμοποιούμε είναι εκπαιδευμένα σε δεδομένα ομιλίας. Επομένως, η επίβλεψη κατά τον διαχωρισμό προέρχεται από την μοντελοποίηση των δεδομένων εκπαίδευσης. Συνεπώς, πιθανώς καταχρηστικά, χρησιμοποιούμε τον όρο Ημι-επιβλεπόμενη Μάθηση.

2.2.4 Γενίκευση Μεθόδων Μηχανικής Μάθησης

Όπως είδαμε η Μηχανική Μάθηση βασίζεται σε μεγάλο βαθμό στα δεδομένα του συνόλου εκπαίδευσης, γεγονός που ελλοχεύει τον κίνδυνο της αδυναμίας γενίκευσης σε δεδομένα εκτός του συνόλου αυτού. Συγκεκριμένα, με τον όρο γενίκευση εννοούμε την ικανότητα του μοντέλου που εξετάζουμε να αποδίδει καλά όχι μονό στα δεδομένα εκπαίδευσης αλλά και σε άγνωστα δεδομένα που δεν έχει αντιμετωπίσει κατά την εκπαίδευση. Μια ειδοποιός διαφορά ανάμεσα στη Μηχανική Μάθηση και τα προβλήματα βελτιστοποίησης είναι ότι στην πρώτη μας ενδιαφέρει η ικανότητα της γενίκευσης και όχι απλά η ελαχιστοποίηση της συνάρτησης σφάλματος στο σύνολο εκπαίδευσης (Goodfellow, Bengio, and Courville 2016).

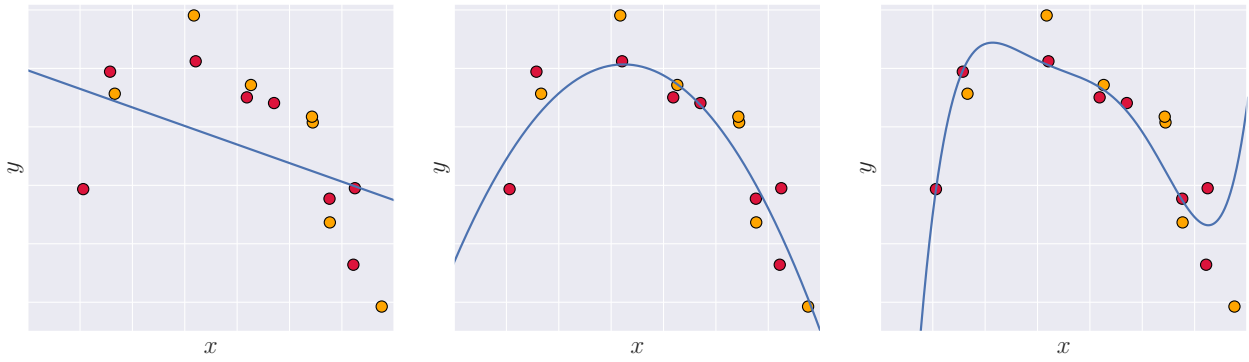
Επομένως, η εκπαίδευση ενός μοντέλου δεν γίνεται σε ολόκληρο το σύνολο των διαθέσιμων δεδομένων. Αντίθετα, το χωρίζουμε στο σύνολο εκπαίδευσης (train set) και το συνήθως μικρότερο σε μέγεθος σύνολο εξέτασης (test set). Συνεπώς, η αξιολόγηση του μοντέλου καθορίζεται από τον βαθμό που επιτυγχάνονται οι δύο παρακάτω στόχοι:

1. Μικρό σφάλμα στο σύνολο εκπαίδευσης.

2. Μικρή απόσταση ανάμεσα στο σφάλμα στο σύνολο εκπαίδευσης και στο σφάλμα στο σύνολο εξέτασης.

Η αποτυχία στον πρώτο στόχο αποδίδεται με τον όρο underfitting ενώ η αποτυχία στον δεύτερο με τον όρο overfitting.

Στο Σχήμα 2.2.2 έχουμε ένα παράδειγμα ενός προβλήματος παλινδρόμησης αντίστοιχο με εκείνο του Σχήματος 2.2.1b. Με κόκκινο χρώμα έχουμε τα δεδομένα εκπαίδευσης ενώ με πορτοκαλί τα δεδομένα εξέτασης τα οποία έχουν προκύψει από πολυώνυμο δευτέρου βαθμού με προσθετικό θόρυβο. Προσαρμόζουμε τρία μοντέλα στα δεδομένα εκπαίδευσης. Πρώτα ένα πολυώνυμο πρώτου βαθμού το οποίο δεν ανταποκρίνεται καλά ούτε στα δεδομένα εκπαίδευσης ούτε στα δεδομένα του συνόλου εξέτασης, δηλαδή έχουμε underfitting. Έπειτα, δοκιμάζουμε ένα πολυώνυμο δευτέρου βαθμού το οποίο επιτυγχάνει ικανοποιητική απόδοση και στα δυο σύνολα. Τέλος, έχουμε ένα πολυώνυμο πέμπτου βαθμού το οποίο πετυχαίνει πολύ χαμηλό σφάλμα στο σύνολο εκπαίδευσης αλλά υψηλό στο σύνολο εξέτασης και άρα έχουμε overfitting.



(a) Underfitting (b) Ικανοποιητική απόδοση (c) Overfitting

Σχήμα 2.2.2: Δοκιμή πολυωνύμων διαφορετικού βαθμού σε πρόβλημα παλινδρόμησης. Αριστερά έχουμε πολυώνυμο πρώτου βαθμού, στο κέντρο έχουμε πολυώνυμο δευτέρου βαθμού και δεξιά έχουμε πολυώνυμο πέμπτου βαθμού. Με κόκκινο χρώμα έχουμε τα δεδομένα εκπαίδευσης ενώ με πορτοκαλί τα δεδομένα εξέτασης.

2.2.5 Υπερ-παράμετροι και Σύνολο Επαλήθευσης

Τα περισσότερα μοντέλα μηχανικής μάθησης διαθέτουν υπερ-παραμέτρους, που καθορίζουν την μορφή και την συμπεριφορά τους. Οι τιμές των υπερ-παραμέτρων δεν μαθαίνονται κατά την εκπαίδευση του μοντέλου αλλά καθορίζονται από πριν. Για παράδειγμα, στο παράδειγμα του Σχήματος 2.2.2 υπερ-παράμετρος είναι ο βαθμός του πολυωνύμου που χρησιμοποιούμε.

Η επιλογή και προσαρμογή των υπερ-παραμέτρων δεν πρέπει να γίνεται με βάση το σύνολο εκπαίδευσης καθώς ελλοχεύει ο κίνδυνος του overfitting. Στο παράδειγμα του Σχήματος 2.2.2 αν επιλέγαμε τον βαθμό του πολυωνύμου με βάση την απόδοση στο σύνολο εκπαίδευσης θα επιλέγαμε το πολυώνυμο τετάρτου βαθμού, όμως θα είχαμε κάνει overfit στα δεδομένα εκπαίδευσης.

Επίσης, προηγουμένως τονίσαμε την σημασία ύπαρξης ενός συνόλου δεδομένων εξέτασης, το οποίο δείχνει την ικανότητα γενίκευσης του μοντέλου αφού έχει εκπαιδευτεί. Είναι εξίσου σημαντικό να μην χρησιμοποιηθεί το σύνολο εξέτασης για τον καθορισμό των υπερ-παραμέτρων. Επομένως, χρειαζόμαστε ένα τρίτο σύνολο δεδομένων το οποίο ονομάζεται σύνολο επαλήθευσης (validation set) και αποτελείται από ένα τμήμα των αρχικών δεδομένων εκπαίδευσης. Συγκεκριμένα, χωρίζουμε τα αρχικά δεδομένα εκπαίδευσης σε δυο ξένα σύνολα, το σύνολο εκπαίδευσης και το σύνολο επαλήθευσης. Συνήθως, το σύνολο επαλήθευσης είναι αρκετά μικρότερο από το σύνολο εκπαίδευσης. Χρησιμοποιούμε το πρώτο για την εκμάθηση των παραμέτρων του μοντέλου. Έπειτα, χρησιμοποιούμε το δεύτερο για να αξιολογήσουμε το μοντέλο και να ενημερώσουμε κατάλληλα τις υπερ-παραμέτρους. Επειδή το σύνολο επαλήθευσης χρησιμοποιείται για την προσαρμογή των υπερ-παραμέτρων, το σφάλμα σε αυτό το σύνολο θα υπερεκτιμά σε ένα βαθμό την ικανότητα γενίκευσης του μοντέλου, αλλά συνήθως λιγότερο από το σφάλμα στο σύνολο εκπαίδευσης. Τέλος, αφού η προσαρμογή των υπερ-παραμέτρων και η εκπαίδευση έχουν ολοκληρωθεί μπορούμε να εκτιμήσουμε την ικανότητα γενίκευσης χρησιμοποιώντας το σύνολο εξέτασης.

Οι υπερ-παράμετροι εκτός από την μορφή του μοντέλου μπορεί να καθορίζουν τον τρόπο εκπαίδευσής του. Όπως θα δούμε στη συνέχεια αρκετά μοντέλα μηχανικής μάθησης μαθαίνουν τις παραμέτρους τους σε βήματα. Στην περίπτωση αυτή ο αριθμός των βημάτων εκπαίδευσης αποτελεί υπερ-παράμετρο και οποιαδήποτε σύγκριση ανάμεσα σε διαφορετικό αριθμό βημάτων θα πρέπει να γίνεται με βάση το συνόλου επαλήθευσης. Συνεπώς, οποιαδήποτε σύγκριση ανάμεσα σε διαφορετικές τεχνικές εκπαίδευσης ή παραμέτρους του αλγορίθμου εκπαίδευσης πρέπει να γίνεται με βάση το σύνολο επαλήθευσης.

2.3 Νευρωνικά Δίκτυα

Ο όρος Νευρωνικό Δίκτυο (Neural Network) προέκυψε από τις προσπάθειες δημιουργίας μαθηματικών αναπαραστάσεων της επεξεργασίας στα βιολογικά συστήματα, όπως στον ανθρώπινο εγκέφαλο (Bishop 2006). Ωστόσο, στα πλαίσια του πεδίου της Μηχανικής Μάθησης και της παρούσας εργασίας ο όρος αυτός αναφέρεται σε μια συγκεκριμένη κατηγορία μοντέλων η οποία δεν σχετίζεται με την μαθηματική μοντελοποίηση βιολογικών συστημάτων. Τα τελευταία χρόνια, η κατηγορία αυτή έχει σημειώσει μεγάλη επιτυχία στη ραγδαία αύξηση της απόδοσης σε μεγάλο εύρος προβλημάτων.

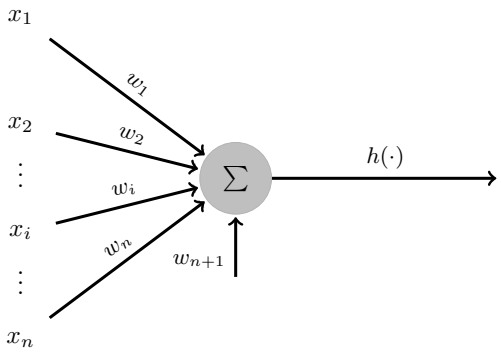
2.3.1 Νευρωνικά Δίκτυα με Τροφοδότηση προς τα Εμπρός

Έχοντας ως στόχο την εκμάθηση μιας συνάρτησης f^*, τα Νευρωνικά Δίκτυα με Τροφοδότηση προς τα Εμπρός (Feedforward Neural Networks) ορίζουν μια αντιστοίχιση $\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{\theta})$. Βάσει αυτής μαθαίνουν τις παραμέτρους $\boldsymbol{\theta}$ ώστε να πετύχουν την καλύτερη δυνατή προσέγγιση της συνάρτησης f^*. (Goodfellow, Bengio, and Courville 2016)

Με τον όρο τροφοδότηση προς τα εμπρός εννοούμε ότι καθώς δίνεται η είσοδος $\boldsymbol{x}$ στην f και υπολογίζεται η έξοδος $\boldsymbol{y}$, οι επιμέρους έξοδοι δεν δίνονται ως είσοδοι πίσω στο μοντέλο.

Perceptron

Βασικό συστατικό ενός νευρωνικού δικτύου όπως υποδηλώνει το όνομα είναι ο νευρώνας (neuron). Επομένως, αξίζει να αναλύσουμε το Perceptron ένα πολύ απλό νευρωνικό δίκτυο με τροφοδότηση προς τα εμπρός που αποτελείται από μόλις έναν νευρώνα, ώστε να κατανοήσουμε καλύτερα τα περισσότερο πολύπλοκα νευρωνικά δίκτυα.



Σχήμα 2.3.1: Σχηματική αναπαράσταση ενός Perceptron

Συγκεκριμένα, ορίζεται ως η συνάρτηση $f : \mathbb{R}^n \to \mathbb{R}$ η οποία λαμβάνει ως είσοδο το διάνυσμα $\boldsymbol{x} \in \mathbb{R}^n$, δίνει ως έξοδο την τιμή $y \in \mathbb{R}$ και οι παράμετροί της είναι το διάνυσμα $\boldsymbol{w} \in \mathbb{R}^{n+1}$. Το διάνυσμα $\boldsymbol{w}$ συνήθως ονομάζεται διάνυσμα βαρών καθώς σταθμίζει τα χαρακτηριστικά του διανύσματος εισόδου.

Η συνάρτηση f είναι ως εξής:

$$y = f(\boldsymbol{x}; \boldsymbol{w}) = h\left(w_{n+1} + \sum_{i=1}^{n} w_i x_i\right)$$

όπου το βάρος w_{n+1} ονομάζεται πόλωση (bias) και $h(x)$ η βηματική συνάρτηση που ορίζεται ως εξής:

$$h(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Θα μπορούσαμε να χρησιμοποιήσουμε το μοντέλο αυτό σε κάποιο πρόβλημα δυαδικής ταξινόμησης, κατηγοριοποιώντας το δείγμα εισόδου $\boldsymbol{x}$ με βάση την έξοδο y.

Προσαρμόζοντας τα βάρη $\boldsymbol{w}$ ορίζουμε ένα υπερ-επίπεδο στο χώρο $\mathbb{R}^n$ το οποίο χωρίζει πιθανές εισόδους σε δυο ημι-επίπεδα και κατ' επέκταση τις ταξινομεί σε δυο κατηγορίες. Συμπεραίνουμε ότι όταν δεν υπάρχει κάποιο επίπεδο που χωρίζει τα δεδομένα σε δυο κατηγορίες, δηλαδή τα δεδομένα αυτά δεν είναι γραμμικά διαχωρίσιμα, το μοντέλο αυτό πάντα θα έχει κάποιο σφάλμα.

2.3.2 Βαθιά Νευρωνικά Δίκτυα με Τροφοδότηση προς τα Εμπρός

Σε αντίθεση με το peceptron, στα βαθιά νευρωνικά δίκτυα, η $f(\boldsymbol{x})$ αποτελείται από την σύνθεση πολλών συναρτήσεων. Επειδή έχουμε τροφοδότηση προς τα εμπρός μπορούμε να κατασκευάσουμε έναν κατευθυνόμενο ακυκλικό γράφο με τον τρόπο που συντίθενται οι συναρτήσεις αυτές. Το μήκος του γράφου αυτό καθορίζει το βάθος του δικτύου. Για παράδειγμα, μπορεί τρεις συναρτήσεις $f^{(1)}, f^{(2)}$ και $f^{(3)}$ να σχηματίζουν την $f(\boldsymbol{x}) = f^{(3)}\left(f^{(2)}\left(f^{(1)}(\boldsymbol{x})\right)\right)$. Ονομάζουμε λοιπόν, την $f^{(1)}$ ως το πρώτο επίπεδο (layer), την $f^{(2)}$ ως το δεύτερο επίπεδο και ούτω καθεξής.

Το πεδίο έρευνας των βαθιών νευρωνικών δικτύων έχει ονομαστεί Βαθιά Μάθηση (Deep Learning).

Multilayer Perceptron

Με βάση τα παραπάνω μπορούμε να χρησιμοποιήσουμε ως βασική μονάδα το perceptron ώστε να σχηματίσουμε ένα απλό βαθύ δίκτυο το οποίο ονομάζεται Multilayer Perceptron.

Συγκεκριμένα, σε κάθε επίπεδο έχουμε έναν αριθμό από νευρώνες, όπως το perceptron ακολουθούμενο όχι απαραίτητα από την βηματική συνάρτηση αλλά πιθανώς με κάποια άλλη συνάρτηση που ονομάζεται συνάρτηση ενεργοποίησης. Κάθε επίπεδο λαμβάνει ως είσοδο την έξοδο του προηγουμένου. Ως πρώτο επίπεδο έχουμε το επίπεδο εισόδου το οποίο δεν πραγματοποιεί κάποια επεξεργασία αλλά τροφοδοτεί την είσοδο στο επόμενο επίπεδο μέχρι το τελικό επίπεδο εξόδου. Το επίπεδο εξόδου μετασχηματίζει τις αναπαραστάσεις που δέχεται ως είσοδο σε μια μορφή που καθορίζεται ανάλογα με το πρόβλημα, όπως για παράδειγμα τιμή στο εύρος $[0, 1]$ που δηλώνει την πιθανότητα η είσοδος να ανήκει σε κάποια κατηγορία. Τα ενδιάμεσα επίπεδα ονομάζονται κρυφά και πραγματοποιούν το κύριο μέρος της επεξεργασίας, ενώ ο αριθμός νευρώνων σε καθένα από αυτά αποτελεί σχεδιαστική επιλογή του μοντέλου. (Bishop 2006)

Για να περιγράψουμε μαθηματικά το μοντέλο αυτό θεωρούμε ότι έχουμε K επίπεδα χωρίς να μετράμε το επίπεδο εισόδου. Το κάθε ένα από αυτά έχει $M^{(k)}$ στον αριθμό νευρώνες, ενώ υπολογίζει την ενεργοποίηση $\boldsymbol{a}^{(k)}$ και έξοδο $\boldsymbol{z}^{(k)}$. Για το πρώτο κρυφό επίπεδο η ενεργοποίηση προκύπτει από τα βάρη του $\mathbf{W}^{(1)}$ και την είσοδο $\boldsymbol{x} \in \mathbb{R}^n$ ως εξής:

$$a_j^{(1)} = \sum_{i=1}^{n} \left(\mathbf{W}_{j,i}^{(1)} x_i\right) + \mathbf{W}_{j,(n+1)}^{(1)}$$

όπου $j = 1, \cdots, M^{(1)}$ και με $\mathbf{W}_{j,(n+1)}$ συμβολίζουμε τις πολώσεις του επιπέδου.

Η έξοδος του υπολογίζεται εφαρμόζοντας την συνάρτηση ενεργοποίησης h στοιχείο προς στοιχείο στο διάνυσμα $\boldsymbol{a}^{(1)}$.

$$\boldsymbol{z}^{(1)} = h\left(\boldsymbol{a}^{(1)}\right)$$

Για τα επόμενα επίπεδα οι ενεργοποιήσεις προκύπτουν αντίστοιχα αφού κάθε επίπεδο λαμβάνει ως είσοδο την έξοδο του προηγούμενου $z_j^{(k-1)}$.

$$a_j^{(k)} = \sum_{i=1}^{M^{(k-1)}} \left(\mathbf{W}_{j,i}^{(k)} z_j^{(k-1)}\right) + \mathbf{W}_{j,\left(M^{(k-1)}+1\right)}^{(k)}$$

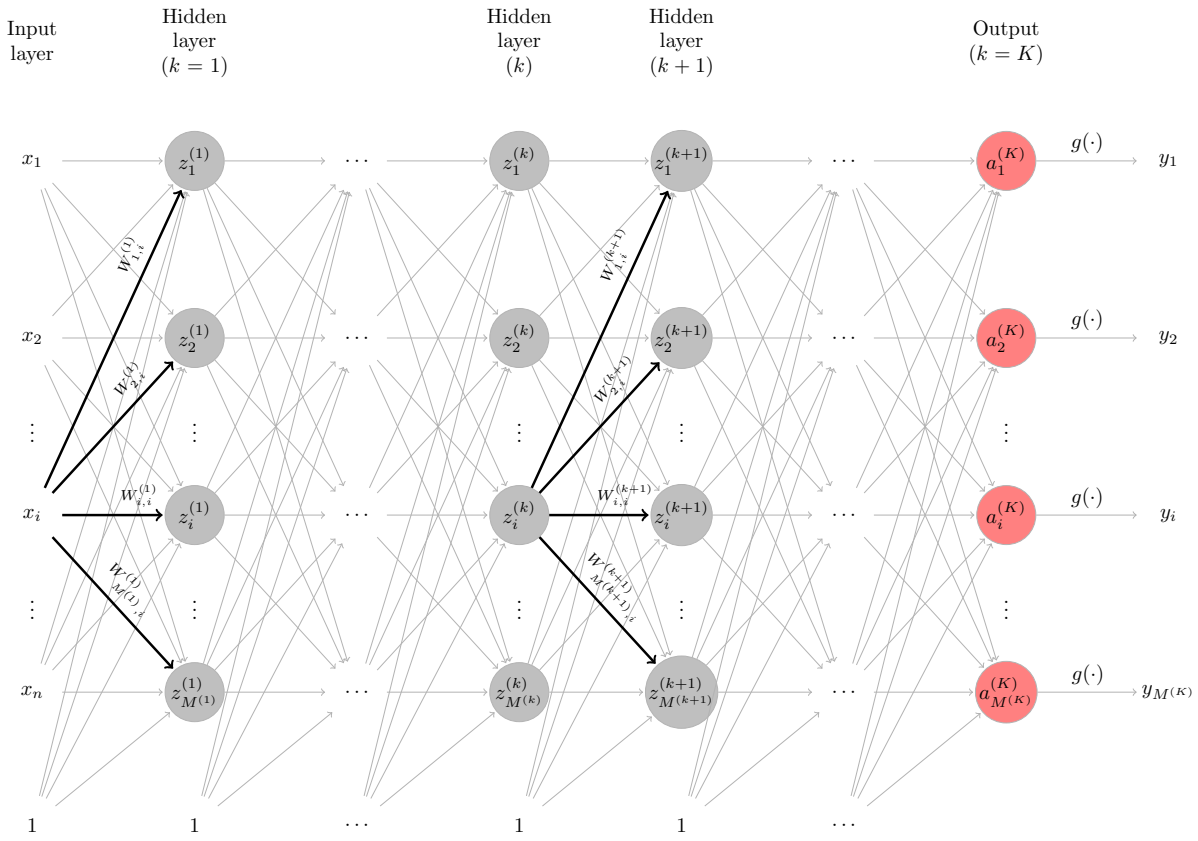
Αντίστοιχα πραγματοποιείται και ο υπολογισμός της εξόδου του επιπέδου με την χρήση της συνάρτησης ενερ-

γοποίησης.

$$\boldsymbol{z}^{(k)} = h\left(\boldsymbol{a}^{(k)}\right)$$

Εξαίρεση αποτελεί το τελικό επίπεδο που χρησιμοποιεί την συνάρτηση g αντί για την h, ώστε να φέρουμε την έξοδο $\boldsymbol{y}$ στην επιθυμητή μορφή.

$$\boldsymbol{y} = g\left(\boldsymbol{a}^{(K)}\right)$$



Σχήμα 2.3.2: Σχηματική αναπαράσταση ενός Multilayer Perceptron

Το δίκτυο που ορίσαμε και απεικονίζουμε στο σχήμα 2.3.2, αποτελεί ένα πλήρως συνδεδεμένο δίκτυο, καθώς κάθε νευρώνας συνδέεται με την έξοδο όλων των νευρώνων του προηγούμενου επιπέδου. Γενικά, σε αντίθεση με το δίκτυο που ορίσαμε, οι συναρτήσεις ενεργοποίησης σε κάθε επίπεδο ή ακόμα και σε κάθε νευρώνα μπορεί να είναι διαφορετικές.

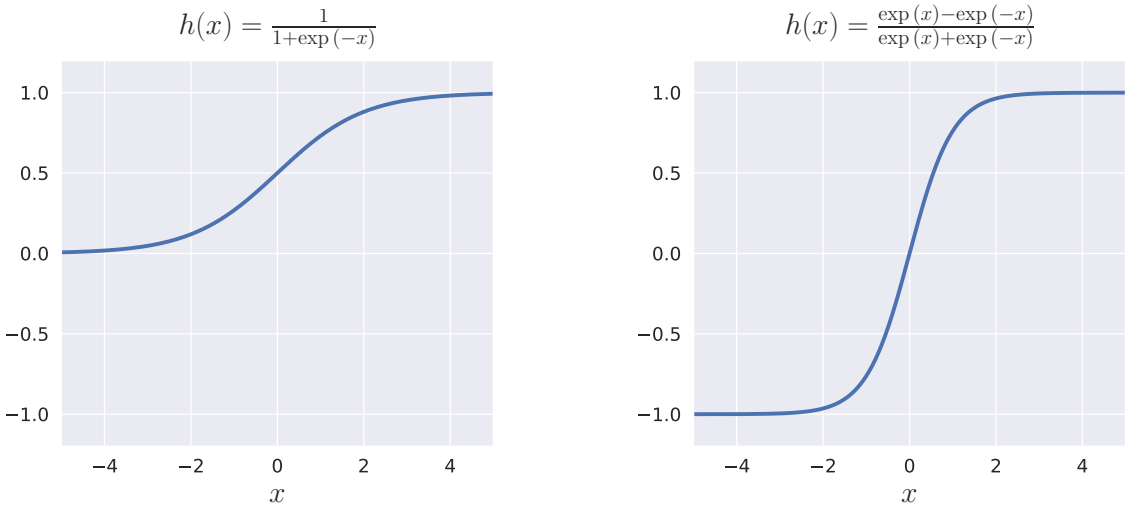
Συναρτήσεις Ενεργοποίησης

Οι ενεργοποιήσεις προκύπτουν ως ένας γραμμικός μετασχηματισμός των διανυσμάτων που λαμβάνουν ως είσοδο. Άρα, χωρίς κάποια συνάρτηση ενεργοποίησης, μια ακολουθία ενός αριθμού από γραμμικά επίπεδα είναι ισοδύναμη με ένα γραμμικό επίπεδο.

$$\boldsymbol{y} = \underbrace{\mathbf{W}^{(3)}\mathbf{W}^{(2)}\mathbf{W}^{(1)}}_{=\mathbf{W}'}\boldsymbol{x}$$

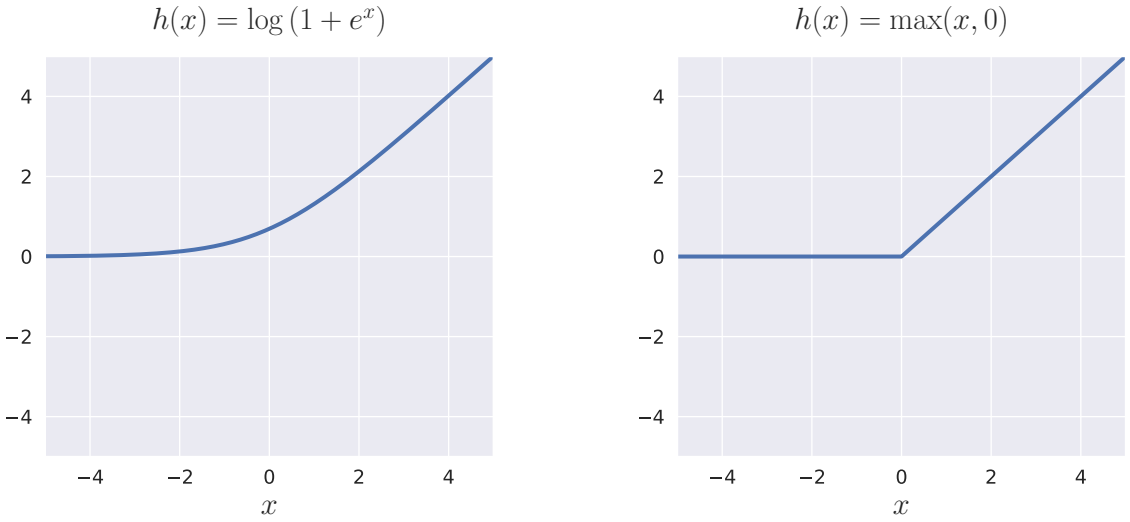
Άρα, ένα βαθύ γραμμικό δίκτυο είναι εξίσου εκφραστικό όσο ένα γραμμικό δίκτυο με ένα επίπεδο, δηλαδή αδυνατεί να προσεγγίσει μια μη γραμμική συνάρτηση. Αποδεικνύεται λοιπόν, ότι χρησιμοποιώντας κάποιες μη γραμμικές συναρτήσεις ενεργοποίησης το νευρωνικό δίκτυο είναι καθολικός προσσεγιστής, ενώ το αποτέλεσμα αυτό ισχύει για μεγάλο εύρος μη γραμμικών συναρτήσεων (Hornik, Stinchcombe, and White 1989; Leshno

et al. 1993). Πρακτικά, υπάρχει δίκτυο το οποίο προσεγγίζει σχεδόν οποιαδήποτε συνάρτηση στόχο f^* όσο καλά επιθυμούμε (Bishop 2006; Goodfellow, Bengio, and Courville 2016).



(a) Sigmoid (b) Tanh

Σχήμα 2.3.3: Μη γραμμικές συναρτήσεις ενεργοποίησης



(a) Softplus (b) ReLU

Σχήμα 2.3.4: Μη γραμμικές συναρτήσεις ενεργοποίησης

Στα Σχήματα 2.3.3 και 2.3.4 έχουμε μερικές μη γραμμικές συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στην βιβλιογραφία.

2.3.3 Εκπαίδευση Νευρωνικών Δικτύων με Τροφοδότηση προς τα Εμπρός

Μέχρι τώρα για τα νευρωνικά δίκτυα με τροφοδότηση προς τα εμπρός έχουμε περιγράψει μόνο την μορφή της αντιστοίχισης $\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{\theta})$ χωρίς να εξετάσουμε πως πραγματοποιείται η εκμάθηση των παραμέτρων $\boldsymbol{\theta}$. Στόχος της εκπαίδευσης είναι η εύρεση των παραμέτρων $\boldsymbol{\theta}$ ώστε να προσεγγίσουμε την συνάρτηση στόχο f^*. Στην περίπτωση των προβλημάτων μηχανικής μάθησης, η γνώση που έχουμε για την συνάρτηση f^* είναι τα δεδομένα εκπαίδευσης. (Goodfellow, Bengio, and Courville 2016)

Βελτιστοποίηση Παραμέτρων

Για την εκπαίδευση, απαραίτητη είναι μια συνάρτηση σφάλματος που λαμβάνει υπ' όψιν την έξοδο του μοντέλου $\boldsymbol{y}$ και στην περίπτωση της επιβλεπόμενης μάθησης και την έξοδο στόχο $\boldsymbol{y}^*$. Συγκεκριμένα, κατά την εκπαίδευση η είσοδος $\boldsymbol{x}$ τροφοδοτείται στο δίκτυο ώστε να προκύψει η έξοδος $\boldsymbol{y}$ και έπειτα η τιμή του σφάλματος, η οποία θα εξαρτάται από τις παραμέτρους $\boldsymbol{\theta}$. Το σφάλμα σε ολόκληρο το σύνολο δεδομένων το συμβολίζουμε ως $J(\boldsymbol{\theta})$. Στόχος μας είναι η εύρεση ενός διανύσματος παραμέτρων $\boldsymbol{\theta}$ το οποίο ελαχιστοποιεί την τιμή του σφάλματος και αποτελεί ένα πρόβλημα βελτιστοποίησης των παραμέτρων.

Συνήθως, η συνάρτηση κόστους $J(\boldsymbol{\theta})$ επιλέγεται ώστε να είναι ομαλή συνεχής συνάρτηση του $\boldsymbol{\theta}$, οπότε σε ελάχιστο αυτή θα πρέπει να ισχύει το εξής:

$$\nabla J(\boldsymbol{\theta}) = 0$$

Είναι σαφές, λόγω της υψηλής μη γραμμικής εξάρτησης του κόστους από τα βάρη, ότι η εύρεση αναλυτικής λύσης είναι πρακτικά αδύνατη. Γι' αυτό καταφεύγουμε σε επαναληπτικές αριθμητικές διαδικασίες. Οι περισσότερες διαδικασίες αυτού του είδους περιλαμβάνουν την επιλογή ενός αρχικού διανύσματος $\boldsymbol{\theta}^{(0)}$ και στη συνέχεια κίνηση στον χώρο των παραμέτρων με βάση κανόνα της μορφής:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \Delta\boldsymbol{\theta}^{(t)}$$

όπου t ο δείκτης του βήματος. Πληθώρα αλγορίθμων αυτής της μορφής έχουν αναπτυχθεί και μεγάλος αριθμός αυτών κάνουν χρήση της κλίσης $\nabla J(\boldsymbol{\theta})$ υπολογισμένης στο $\boldsymbol{\theta}^{(t)}$ για τον υπολογισμό του βήματος $\Delta\boldsymbol{\theta}^{(t)}$. (Bishop 2006)

Η απλούστερη επαναληπτική μέθοδος που χρησιμοποιεί πληροφορία κλίσης είναι η Βελτιστοποίηση Καθοδικής Κλίσης (Gradient Descent), η οποία αποτελείται από τον παρακάτω κανόνα ενημέρωσης των βαρών.

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta\nabla J\left(\boldsymbol{\theta}^{(t)}\right)$$

οπού η είναι η παράμετρος που ονομάζεται ρυθμός εκπαίδευσης (learning rate) και καθορίζει το μέγεθος των βημάτων. Επειδή η συνάρτηση σφάλματος υπολογίζεται από ολόκληρο το σύνολο δεδομένων κάθε υπολογισμός της κλίσης απαιτεί ολόκληρη την επεξεργασία του συνόλου αυτού.

Διαφορετικές εκδόσεις της μεθόδου αυτής δεν δουλεύουν με το ολόκληρο το σύνολο δεδομένων σε κάθε βήμα αλλά με ένα υποσύνολο αυτού, οι οποίες ονομάζονται minibatch ή minibatch stochastic ενώ συχνά καλούνται απλά stochastic (Goodfellow, Bengio, and Courville 2016). Για παράδειγμα, η μέθοδος της Στοχαστικής Καθοδικής Κλίσης (Stochastic Gradient Descent ή SGD) στη μορφή που προτάθηκε υπολογίζει το κόστος, την κλίση και πραγματοποιεί το βήμα ενημέρωσης των βαρών για ένα δείγμα εισόδου κάθε φορά (Bishop 2006). Το δείγμα αυτό μπορεί να είναι τυχαία επιλεγμένο με αντικατάσταση είτε επιλέγοντας ακολουθιακά τα δείγματα του συνόλου εκπαίδευσης.

Οι περισσότεροι αλγόριθμοι που χρησιμοποιούνται χρησιμοποιούν πάνω από ένα δείγματα αλλά όχι ολόκληρο το σύνολο δεδομένων. Αυτό συμβαίνει λόγω δυνατότητας παραλληλοποίησης των υπολογισμών αλλά και επειδή έχει παρατηρηθεί ότι μικρά υποσύνολα (batches) οδηγούν σε καλύτερη γενίκευση του μοντέλου (Goodfellow, Bengio, and Courville 2016). Η πιο δημοφιλής μέθοδος ενημέρωσης βαρών που χρησιμοποιείται είναι η Adam (Kingma and Ba 2014).

Τοπικά Ελάχιστα

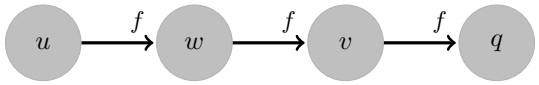
Λόγω της πολυπλοκότητας των νευρωνικών δικτύων η συνάρτηση σφάλματος συνήθως παρουσιάζει μεγάλο αριθμό τοπικών ελαχίστων. Εκ πρώτης όψεως το γεγονός αυτό παρουσιάζεται ως πρόβλημα για την εκπαίδευση με επαναληπτικές μεθόδους που βασίζονται σε παραγώγους. Όμως στην πράξη δεν αποτελεί πρόβλημα συχνά, λόγω της χρήσης μεθόδων όπως SGD και Adam και αφού μεγάλος αριθμός των τοπικών ελαχίστων είναι ισοδύναμα ως προς την τιμή του σφάλματος σε αυτά (Goodfellow, Bengio, and Courville 2016).

Μετάδοση Σφάλματος προς τα Πίσω

Για την χρήση των μεθόδων ενημέρωσης των βαρών που περιγράψαμε είναι αναγκαίο να έχουμε μια αποτελεσματική τεχνική για τον υπολογισμό των κλίσεων $\nabla J(\boldsymbol{\theta})$. Αυτό μπορεί να επιτευχθεί με την τεχνική Μετάδοσης Σφάλματος προς τα Πίσω (Error Backpropagation ή απλά Backprop), η οποία επιτρέπει στην πληροφορία του σφάλματος να μεταδοθεί προς τα πίσω ώστε να υπολογίσουμε τις ζητούμενες κλίσεις. Σε ένα γενικό νευρωνικό δίκτυο τροφοδότησης προς τα εμπρός με αυθαίρετη τοπολογία και αυθαίρετες παραγωγίσιμες μη γραμμικές συναρτήσεις ενεργοποίησης μπορούμε να εφαρμόσουμε την τεχνική αυτή.

Αρχικά, από το νευρωνικό δίκτυο προκύπτει ο Γράφος Υπολογισμού (Computational Graph) του οποίου οι κόμβοι είναι μεταβλητές. Ακμές έχουμε από έναν κόμβο προς έναν άλλο αν η μεταβλητή του πρώτου χρησιμοποιείται στον υπολογισμό της μεταβλητής του δεύτερου. Ο γράφος αυτός θα είναι κατευθυνόμενος και ακυκλικός, θα περιέχει τις μεταβλητές εισόδου και ο τελευταίος κόμβος θα είναι η τιμή του σφάλματος.

Για παράδειγμα, για την σύνθεση συναρτήσεων $z = f(f(f(u)))$ με $w = f(u)$, $v = f(w)$ και $q = f(v)$ όπου $u, w, v, q \in \mathbb{R}$ και $f : \mathbb{R} \to \mathbb{R}$ προκύπτει ο γράφος υπολογισμού του Σχήματος 2.3.5.



Σχήμα 2.3.5: Σχηματική αναπαράσταση ενός Γράφου Υπολογισμού

Στο σημείο αυτό αξίζει να παρουσιάσουμε τον κανόνα της αλυσίδας που χρησιμοποιείται για τον υπολογισμό παραγώγων όταν έχουμε σύνθεση συναρτήσεων. Για την συνάρτηση του Σχήματος 2.3.5 θα είναι

$$\frac{dq}{du} = \frac{dq}{dv}\frac{dv}{du}$$

Εφαρμόζοντας πάλι τον κανόνα θα έχουμε

$$\frac{dq}{du} = \frac{dq}{dv}\frac{dv}{du} = \frac{dq}{dv}\frac{dv}{dw}\frac{dw}{du}$$

Ενώ αν είχαμε τα διανύσματα $\boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{w} \in \mathbb{R}^m$ με $\boldsymbol{w} = g(\boldsymbol{u})$ και $q = f(\boldsymbol{w})$ τότε ο κανόνας ορίζει

$$\frac{\partial q}{\partial u_i} = \sum_j \frac{\partial q}{\partial w_j}\frac{\partial w_j}{\partial u_i}$$

Επομένως, από τον γράφο υπολογισμού με την βοήθεια του κανόνα της αλυσίδας μπορούμε να πάρουμε την παράγωγο ενός βαθμωτού μεγέθους ως προς οποιαδήποτε κόμβο του γράφου.

Στην τεχνική μετάδοσης σφάλματος προς τα πίσω για ένα νευρωνικό δίκτυο, αφού πραγματοποιήσουμε την τροφοδότηση προς τα εμπρός, ξεκινώντας από πίσω προς τα εμπρός στον γράφο υπολογισμού υπολογίζουμε τις παραγώγους του σφάλματος ως προς τις παραμέτρους του εκάστοτε επιπέδου με την βοήθεια του κανόνα της αλυσίδας. Μεταδίδοντας τις τιμές των μερικών παραγώγων προς τα πίσω αποφεύγουμε επανυπολογισμούς και υπολογίζουμε όλες τις παραγώγους αποδοτικά.

2.3.4 Autoencoders

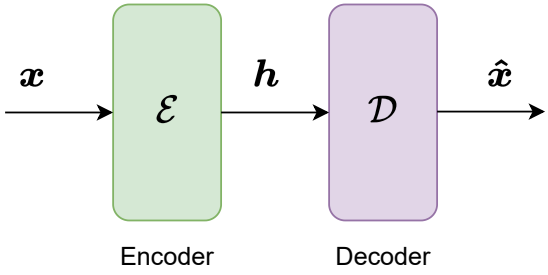
Ένας Autoencoder είναι μια ειδική περίπτωση νευρωνικού δικτύου. Συγκεκριμένα είναι ένα νευρωνικό δίκτυο το οποίο εκπαιδεύεται ώστε να αναπαραγάγει την είσοδο που δέχεται στην έξοδο, συχνά (αλλά όχι απαραίτητα) μέσω μιας ενδιάμεσης αναπαράστασης μικρότερης διάστασης σε σχέση με την είσοδο αλλά όχι απαραίτητα. (Goodfellow, Bengio, and Courville 2016)

Αποτελείται από δύο μέρη: τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder). Ο κωδικοποιητής $\mathcal{E}$

δέχεται την είσοδο του δικτύου $\boldsymbol{x}$ και δίνει ως έξοδο την ενδιάμεση αναπαράσταση $\boldsymbol{h} = \mathcal{E}(\boldsymbol{x})$. Ο αποκωδικοποιητής $\mathcal{D}$ δέχεται ως είσοδο την ενδιάμεση αναπαράσταση $\boldsymbol{h}$ και δίνει ως έξοδο την ανακατασκευή της εισόδου $\hat{\boldsymbol{x}} = \mathcal{D}(\boldsymbol{h})$.

Η εκπαίδευση στην πλειονότητα των περιπτώσεων πραγματοποιείται αντίστοιχα με τα νευρωνικά δίκτυα τροφοδότησης προς τα εμπρός με μετάδοση σφάλματος προς τα πίσω, ελαχιστοποιώντας κάποιο σφάλμα ανακατασκευής.

Η χρησιμότητά τους προκύπτει από τους περιορισμούς που τους επιβάλλονται, με επακόλουθο η ανακατασκευή να είναι προσεγγιστική και να μην μπορεί να είναι τέλεια. Ο συνηθέστερος περιορισμός είναι η διάσταση της ενδιάμεσης αναπαράστασης, η οποία επιλέγεται να είναι μικρότερη από την διάσταση της εισόδου. Σκοπός είναι το δίκτυο να αναγκαστεί να εξάγει χρήσιμα χαρακτηριστικά που περιγράφουν τα δεδομένα εισόδου.



Σχήμα 2.3.6: Σχηματική αναπαράσταση ενός Autoencoder

Η επιθυμητή ανακατασκευή μπορεί να μην ταυτίζεται πάντα με την είσοδο. Για παράδειγμα, οι Autoencoders Αποθορυβοποίησης (Denoising Autoencoders) εκπαιδεύονται στα πλαίσια επιβλεπόμενης μάθησης ώστε να λαμβάνουν μια θορυβώδη είσοδο $\boldsymbol{x}$ και να δίνουν την έξοδο $\hat{\boldsymbol{x}}$ χωρίς θόρυβο.

Μια άλλη κατηγορία autoencoder που θα μας απασχολήσει στην παρούσα εργασία είναι οι Μη Αρνητικοί Autoencoders (Non Negative Autoencoders), στους οποίους η είσοδος, η έξοδος και η ενδιάμεση αναπαράσταση είναι μη αρνητικά μεγέθη. Δηλαδή ισχύει $\boldsymbol{x}, \hat{\boldsymbol{x}}, \boldsymbol{h} \in \mathbb{R}_{\geq 0}$. Στο Κεφάλαιο 4 γίνεται εκτενής παρουσίαση τους.

2.4 Non-Negative Matrix Factorization

Δεδομένου πίνακα $\mathbf{X}$ διαστάσεων $M \times N$ με η αρνητικά στοιχεία, δηλαδή $\mathbf{X} \in \mathbb{R}_{\geq 0}$, το πρόβλημα της Μη αρνητικής Παραγοντοποίησης Πίνακα (Non-Negative Matrix Factorization ή NMF) ορίζεται ως η εύρεση παραγοντοποίησης

$$\mathbf{X} \approx \mathbf{WH}$$

όπου $\mathbf{W} \in \mathbb{R}_{\geq 0}$ και $\mathbf{H} \in \mathbb{R}_{\geq 0}$ είναι μη αρνητικοί πίνακες (δηλαδή όλα τα στοιχεία τους είναι μη αρνητικά), διαστάσεων $M \times K$ και $K \times N$ αντίστοιχα. Συνήθως, η διάσταση K επιλέγεται ώστε $MK + KN \ll MN$ δηλαδή έχουμε μείωση της διάστασης των αρχικών δεδομένων. Ενδιαφερόμαστε δηλαδή για μια προσέγγιση του $\mathbf{X}$ από το γινόμενο των μη αρνητικών πινάκων $\mathbf{W}$ και $\mathbf{H}$.

Με τον όρο μέθοδος NMF αναφερόμαστε σε μεθόδους που θέτουν ένα πρόβλημα ως πρόβλημα εύρεσης μη αρνητικής παραγοντοποίησης πίνακα και έπειτα το επιλύουν. Έτσι, η μέθοδος NMF έχει χρησιμοποιηθεί στην εκμάθηση κομματιών προσώπων σε εικόνες (Lee and Seung 1999), σε προβλήματα μουσικής (Smaragdis and Brown 2003) καθώς και πληθώρα άλλων προβλημάτων (Févotte and Idier 2011).

Συνήθως θέτουμε το πρόβλημα ως ένα πρόβλημα ελαχιστοποίησης μιας συνάρτησης κόστους ανάμεσα στον αρχικό πίνακα $\mathbf{X}$ και τον ανακατασκευασμένο πίνακα $\mathbf{WH}$, την οποία συμβολίζουμε ως $D(\mathbf{X}|\mathbf{WH})$. Έχουμε λοιπόν

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{X}|\mathbf{WH}) \quad \text{υπό τους περιορισμούς } \mathbf{W} \in \mathbb{R}_{\geq 0}, \mathbf{H} \in \mathbb{R}_{\geq 0}$$

με συνάρτηση κόστους για την οποία ισχύει το εξής:

$$D\left(\mathbf{X}|\mathbf{WH}\right) = \sum_{i,j} d\left(\mathbf{X}_{i,j}|(\mathbf{WH})_{i,j}\right)$$

όπου $d(x|y)$ βαθμωτή συνάρτηση κόστους ανάμεσα στο $x \in \mathbb{R}_{\geq 0}$ και στο $y \in \mathbb{R}_{\geq 0}$.

Η επιλογή της συνάρτησης κόστους εξαρτάται από τα δεδομένα, το πρόβλημα καθώς και τους περιορισμούς αυτού (Cichocki et al. 2009). Ίσως η πιο δημοφιλής συνάρτηση κόστους είναι η γενικευμένη απόκλιση Kullback Leibler για την οποία ισχύει το εξής:

$$d(x|y) = x \log \frac{x}{y} - x + y$$

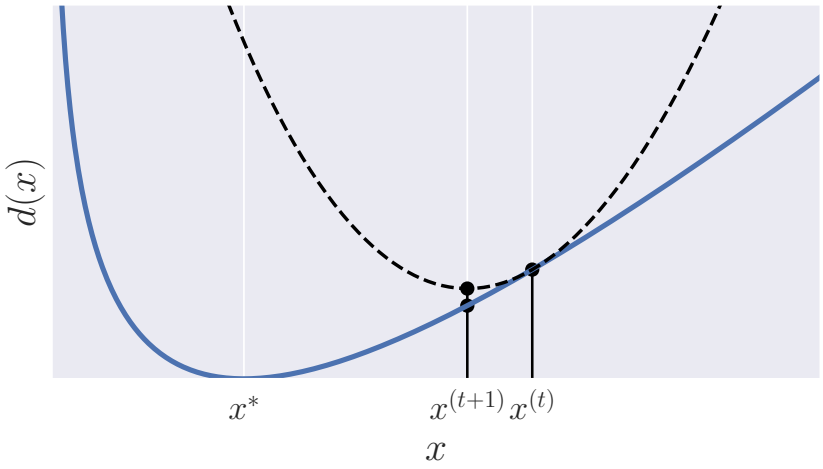
για την οποία έχει βρεθεί επαναληπτικός αλγόριθμος για την εύρεση των πινάκων $\mathbf{W}$ και $\mathbf{H}$ (Lee and Seung 2000) τον οποίο θα παρουσιάσουμε.

Προφανώς, η λύση του προβλήματος δεν είναι μοναδική. Ωστόσο, η συνάρτηση κόστους είναι κυρτή ξεχωριστά ως προς το $\mathbf{W}$ και ως προς το $\mathbf{H}$ αλλά δεν είναι κυρτή ως προς το $\{\mathbf{W}, \mathbf{H}\}$.

Λαμβάνοντας υπ' όψιν τα παραπάνω, η στρατηγική βελτιστοποίησης που ακολουθούμε είναι αυτή της εναλλασσόμενης βελτιστοποίησης. Συγκεκριμένα, σε μια επανάληψη του αλγορίθμου

- Ενημερώνουμε τον πίνακα $\mathbf{W}$ κρατώντας τον πίνακα $\mathbf{H}$ σταθερό
- Ενημερώνουμε τον πίνακα $\mathbf{H}$ κρατώντας τον πίνακα $\mathbf{W}$ σταθερό

Η τεχνική αυτή ονομάζεται Block Coordinate Descent.



Σχήμα 2.4.1: Σχηματική αναπαράσταση της διαδικασίας Majorization-Minimization στη μια διάσταση. Με μπλε έχουμε την συνάρτηση κόστους $d(x)$ την οποία επιθυμούμε να ελαχιστοποιήσουμε, να βρούμε δηλαδή το x^*. Με μαύρο έχουμε την βοηθητική συνάρτηση η οποία αποτελεί άνω φράγμα της συνάρτηση κόστους ενώ έχουν ίδια τιμή στο σημείο $x^{(t)}$. Ελαχιστοποιούμε την βοηθητική συνάρτηση και προκύπτει το νέο σημείο $x^{(t+1)}$.

Για την ενημέρωση του κάθε πίνακα χρησιμοποιούμε την διαδικασία Majorization-Minimization . Έστω ότι επιθυμούμε να ενημερώσουμε τον $\mathbf{W}$ κρατώντας τον $\mathbf{H}$ σταθερό. Βρίσκουμε μια βοηθητική συνάρτηση άνω φράγμα για την συνάρτηση κόστους, για την οποία ισχύει ότι ταυτίζεται με την δεύτερη στο σημείο που βρισκόμαστε. Έπειτα, βρίσκουμε τον $\mathbf{W}$ που ελαχιστοποιεί την συνάρτηση άνω φράγματος. Έτσι, μπορούμε να εγγυηθούμε ότι σε κάθε επανάληψη το κόστος δεν αυξάνεται. Στο Σχήμα 2.4.1 έχουμε την σχηματική αναπαράσταση της διαδικασίας στη μονοδιάστατη περίπτωση.

Έχουμε λοιπόν τους πολλαπλασιαστικούς κανόνες ενημέρωσης που υλοποιούν την παραπάνω διαδικασία για την

γενικευμένη Kullback Leibler απόκλιση:

$$\mathbf{H}_{i,j} \leftarrow \mathbf{H}_{i,j} \frac{\sum_l \left(\frac{\mathbf{W}_{l,i}\mathbf{X}_{l,j}}{(\mathbf{WH})_{l,j}} \right)}{\sum_k \mathbf{W}_{k,i}} \qquad \mathbf{W}_{i,j} \leftarrow \mathbf{W}_{i,j} \frac{\sum_l \left(\frac{\mathbf{H}_{j,l}\mathbf{X}_{i,l}}{(\mathbf{WH})_{i,l}} \right)}{\sum_k \mathbf{H}_{j,k}}$$

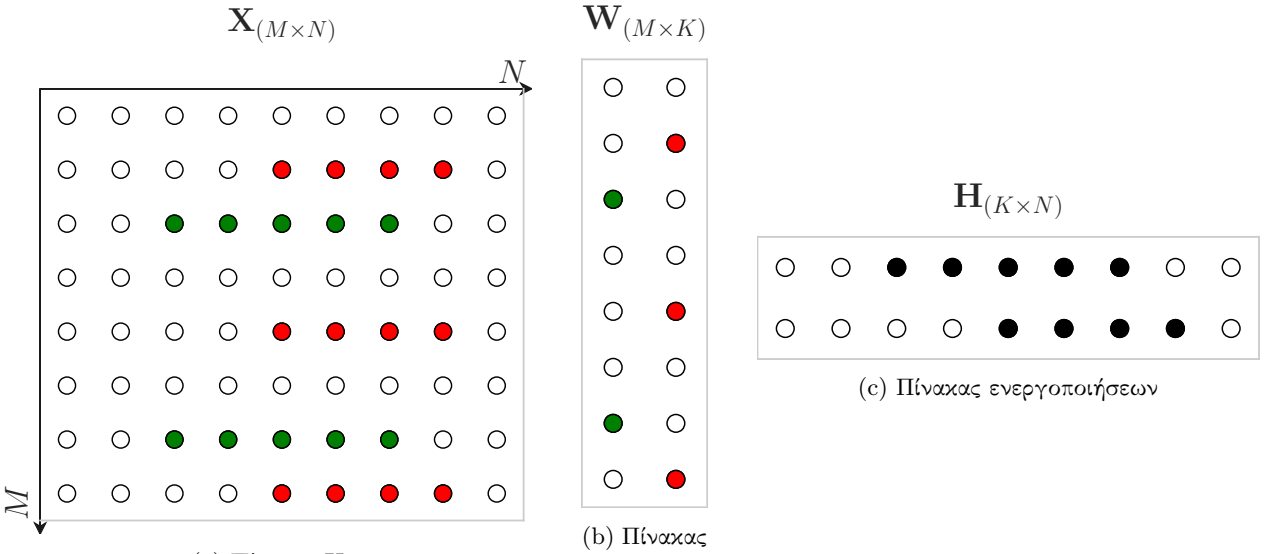
Επισημαίνουμε ότι η συγκεκριμένη μέθοδος δεν διαθέτει θεωρητικές διασφαλίσεις πέρα από την σύγκλιση σε κάποιο τοπικό ελάχιστο (Lee and Seung 2000), αλλά στην πράξη είναι εύκολη στην υλοποίηση και δίνει ικανοποιητικές λύσεις, οι οποίες όμως εξαρτώνται από την αρχικοποίηση των πινάκων.

Αντίστοιχη διαδικασία μπορούμε να ακολουθήσουμε και να εξάγουμε κανόνες ενημέρωσης για άλλες συναρτήσεις κόστους, όπως η β-απόκλιση που αποτελεί γενίκευση της γενικευμένης Kullback Leibler απόκλισης καθώς και της ευκλείδειας απόστασης (Févotte and Idier 2011).

Στη συνάρτηση κόστους μπορούμε να προσθέσουμε όρους που ελέγχουν την ομαλότητα μεταξύ των διαφορετικών στηλών του πίνακα $\mathbf{H}$ ή και την αραιότητα του πίνακα $\mathbf{H}$ (Virtanen 2007). Μπορούμε να ακολουθήσουμε την άνωθεν διαδικασία ώστε να εξάγουμε τους κανόνες ενημέρωσης των πινάκων.

2.4.1 Ερμηνεία NMF

Ένα χρήσιμο χαρακτηριστικό της μεθόδου NMF είναι ότι παραγοντοποιεί την αρχική αναπαράσταση $\mathbf{X}$ ως μια σύνθεση θεμελιωδών κομματιών, χαρακτηριστικό που προκύπτει από την ιδιότητα της μη αρνητικότητας. Τα κομμάτια αυτά είναι μη αρνητικά και η σύνθεσή τους γίνεται προσθετικά. Έτσι σε αντίθεση με άλλες παραγοντοποιήσεις δεν μπορεί το ένα κομμάτι να ακυρώνει το άλλο, με αποτέλεσμα να μπορούμε να ερμηνεύσουμε την σύνθεση τους.



(a) Πίνακας $\mathbf{X}$

(b) Πίνακας βάσεων

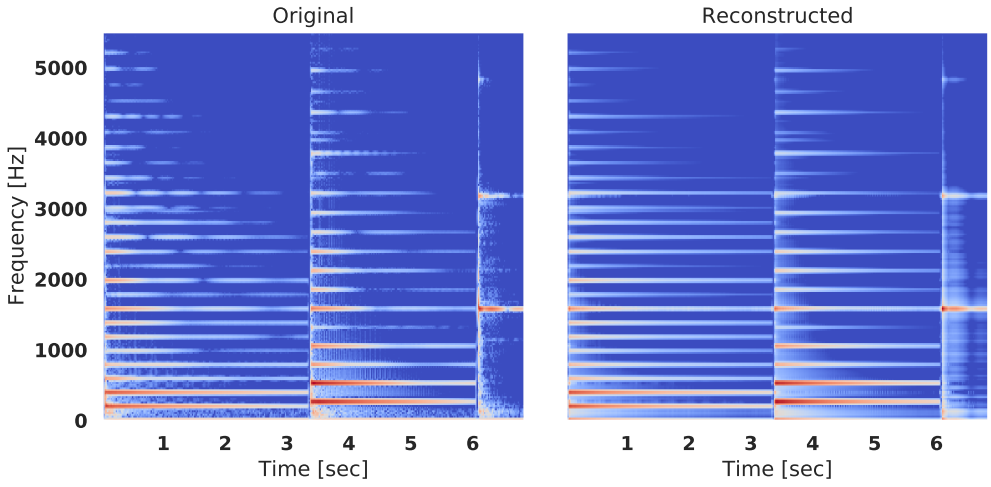
(c) Πίνακας ενεργοποιήσεων

Σχήμα 2.4.2: Απλό παράδειγμα ερμηνείας της μεθόδου NMF. Αριστερά έχουμε τον πίνακα που έχει παραγοντοποιηθεί στους πίνακες που βρίσκονται δυο δεξιά. Στο κέντρο έχουμε τον πίνακα βάσεων με τις δυο βάσεις. Στα δεξιά έχουμε τον πίνακα ενεργοποιήσεων όπου με μαύρη κουκίδα σημαίνει ότι η αντίστοιχη βάση είναι ενεργή.

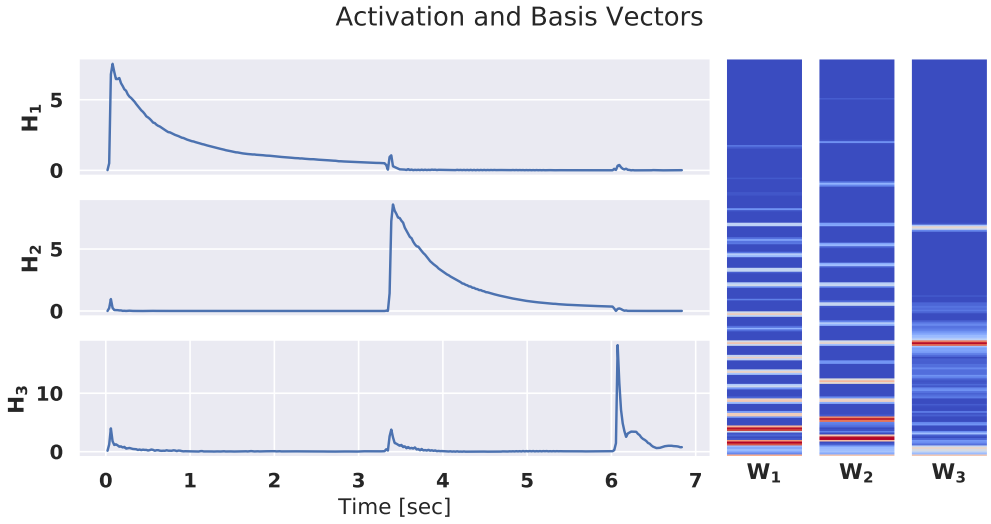
Αν θεωρήσουμε τις στήλες του πίνακα $\mathbf{X}$ ως ένα σύνολο δεδομένων στον χώρο $\mathbb{R}^M_{\geq 0}$ τότε και οι στήλες του πίνακα $\mathbf{W}$ ανήκουν στον χώρο αυτό. Επομένως, μπορούμε να ερμηνεύσουμε τις K στήλες του $\mathbf{W}$ ως πρότυπα τα οποία ονομάζουμε βάσεις. Έπειτα, ο πίνακας $\mathbf{H}$ έχει K γραμμές τις οποίες μπορούμε να ερμηνεύσουμε ως ενεργοποιήσεις για τις αντίστοιχες βάσεις του $\mathbf{W}$. Έτσι ορίζουμε τον $\mathbf{H}$ ως πίνακα ενεργοποιήσεων και τον $\mathbf{W}$ ως πίνακα βάσεων. Στο Σχήμα 2.4.2 οπτικοποιούμε την ερμηνεία αυτή σε ένα απλό παράδειγμα.

Στην περίπτωση που τα δεδομένα $\mathbf{X}$ παράχθηκαν ως ένα μείγμα μη αρνητικών πηγών, συχνά η μέθοδος NMF θα ανακαλύψει τις πηγές αυτές. Συνεπώς, η NMF μπορεί να λειτουργήσει ως μια μέθοδος χωρίς επίβλεψη για προβλήματα διαχωρισμού πηγών.

Στην περίπτωση των ηχητικών σημάτων, μπορούμε να έχουμε μια μη αρνητική αναπαράστασή τους μέσω του μέτρου του STFT τους. Εξετάζουμε λοιπόν την παραπάνω ιδιότητα στο μέτρο του STFT ενός ηχητικού μείγματος που αποτελείται από τρεις διαφορετικές νότες πιάνου διαδοχικά στον χρόνο. Εφαρμόζουμε την μέθοδο NMF με $K = 3$ χρησιμοποιώντας τον αλγόριθμο που περιγράψαμε παραπάνω με την γενικευμένη Kullback Leibler απόκλιση. Στο σχήμα 2.4.3 έχουμε το φασματογράφημα $\mathbf{X}$ του αρχικού σήματος στα αριστερά και στα δεξιά το ανακατασκευασμένο φασματογράφημα $\mathbf{WH}$. Παρατηρούμε ότι η ανακατασκευή είναι αρκετά πιστή στο πρωτότυπο.



Σχήμα 2.4.3: Προσέγγιση ηχητικού σήματος τριών νότων πιάνου με NMF με $K = 3$, στα αριστερά έχουμε το αρχικό φασματογράφημα που επιθυμούμε να προσεγγίσουμε και στα δεξιά έχουμε την προσέγγιση.



Σχήμα 2.4.4: Στα αριστερά έχουμε τις γραμμές του πίνακα ενεργοποιήσεων $\mathbf{H}$ και στα δεξιά τις στήλες του πίνακα βάσεων $\mathbf{W}$ κατόπιν εφαρμογή της μεθόδου NMF στο σήμα του σχήματος 2.4.3.

Έπειτα, στο Σχήμα 2.4.4 αναπαριστούμε γραφικά τις γραμμές του πίνακα $\mathbf{H}$ και τις στήλες του πίνακα $\mathbf{W}$. Πράγματι, οι στήλες του πίνακα $\mathbf{W}$ φαίνονται σαν βάσεις κάθε μια από τις οποίες αντιστοιχεί σε μια μουσική νότα. Αντίστοιχα, οι γραμμές του πίνακα $\mathbf{H}$ μοιάζουν σαν ενεργοποιήσεις που καθορίζουν πότε έχουμε την

αντίστοιχη βάση. Είναι εμφανές ότι οι κορυφές των ενεργοποιήσεων συμπίπτουν με την παρουσία κάποιας νότας στο μείγμα.

Όμως, δεν έχουμε κάποια εγγύηση ότι σε κάθε περίπτωση θα έχουμε εκμάθηση των θεμελιωδών κομματιών που αποτελούν την είσοδο δηλαδή των πηγών. Αυτό οφείλεται και στο γεγονός ότι η NMF αποτελεί ένα απλό μοντέλο το οποίο για παράδειγμα δεν λαμβάνει υπ' όψιν συσχετίσεις ανάμεσα στις στήλες του πίνακα εισόδου. Επομένως, σε ένα πιο δύσκολο παράδειγμα με ένα πιο πολύπλοκο μείγμα, πιθανώς η παραγοντοποίηση που θα λαμβάναμε δεν θα ήταν τόσο χρήσιμη. Ωστόσο, λόγω των πλεονεκτημάτων που διαθέτει όπως η ερμηνευσιμότητα και η δυνατότητα λειτουργίας χωρίς επίβλεψη, έχει αποτελέσει αφετηρία έρευνας με σκοπό την δημιουργία επεκτάσεων της.

Κεφάλαιο 3

Βιβλιογραφική Επισκόπηση

3.1 Μέθοδοι για το Πρόβλημα Αποθορυβοποίησης Σήματος Φωνής

Πρώτα, πραγματοποιούμε μια γενική επισκόπηση των μεθόδων για το πρόβλημα της Αποθορυβοποίησης Σήματος Φωνής. Εξετάζουμε πληθώρα μεθόδων που έχουν προταθεί, είτε με πλήρη επίβλεψη είτε με μερική επίβλεψη στις οποίες έχουμε πρόσβαση σε θορυβώδη σήματα κατά την εκπαίδευση. Μεθόδους με μερική επίβλεψη που δεν έχουν γνώση του θορύβου κατά την εκπαίδευση, τις παρουσιάζουμε στην ενότητα 3.3.

3.1.1 Μέθοδοι ΨΕΣ

Αρχικά, αξίζει να αναφερθούμε σε κλασσικές τεχνικές ψηφιακής επεξεργασίας σήματος, παρά το γεγονός ότι πιο πρόσφατες προσεγγίσεις βασισμένες σε νευρωνικά δίκτυα έχουν επικρατήσει στο ερευνητικό πεδίο και έχουν δώσει εντυπωσιακά αποτελέσματα.

Μια τέτοια τεχνική είναι η Φασματική Αφαίρεση (Spectral Subtraction) (Boll 1979) η οποία είναι ιδιαίτερα απλή. Εργαζόμαστε με το μέτρο του STFT αμελώντας την φάση του. Υπολογίζουμε μια εκτίμηση θορύβου, λαμβάνοντας τον χρονικό μέσο όρο του φάσματος για κάθε κελί συχνότητας, από ένα τμήμα του σήματος χωρίς ομιλία και την αφαιρούμε από το σήμα στο πεδίο της συχνότητας. Έπειτα, επιχειρούμε να αποσβέσουμε τον υπολειπόμενο θόρυβο στο σήμα.

Μια άλλη τεχνική είναι το Φιλτράρισμα Wiener (Lim and A. V. Oppenheim 1979). Κάνουμε την υπόθεση ότι ο θόρυβος είναι προσθετικός και ασυσχέτιστος με το σήμα φωνής. Έτσι, αφού πάρουμε μια εκτίμηση για την φασματική πυκνότητα ισχύος του θορύβου $P_{nn}(f)$ και άλλη μια για την ομιλία $P_{ss}(f)$, υπολογίζουμε το φίλτρο Wiener ως εξής:

$$\frac{P_{ss}(f)}{P_{ss}(f) + P_{nn}(f)}$$

το οποίο εφαρμόζεται στο θορυβώδες σήμα στο πεδίο της συχνότητας, πολλαπλασιάζοντας με μια τιμή ανάμεσα στο 0 και το 1 κάθε κελί χρόνου-συχνότητας. Επειδή χρειαζόμαστε εκτιμήσεις και για τον θόρυβο αλλά και για την καθαρή ομιλία, μπορούμε να χρησιμοποιήσουμε αφαίρεση φάσματος πρώτα ώστε να πάρουμε μια εκτίμηση για την φασματική πυκνότητα ισχύος της ομιλίας.

Τέλος, μια άλλη δημοφιλής μέθοδος είναι ο Εκτιμητής Φασματικού Πλάτους Βραχέος Χρόνου (Short-Time Spectral Amplitude Estimator) (Ephraim and Malah 1984) που εκτιμά τα χαρακτηριστικά του θορύβου από μέρος του φάσματος χωρίς ομιλία και στην συνέχεια με βάση αυτά υπολογίζει ένα κέρδος το οποίο εφαρμόζεται στο πεδίο της συχνότητας. Άλλος τρόπος αποθορυβοποίησης περιγράφεται στο (Ephraim and Van Trees 1995) όπου βρίσκουμε τον υποχώρο των θορυβωδών σημάτων καθώς και τον υποχώρο των σημάτων θορύβου, βάσει των οποίων αφαιρούμε τον θόρυβο από το σήμα.

3.1.2 Μέθοδοι Βασιζόμενες σε Νευρωνικά Δίκτυα

Σήματα στο Πεδίο της Συχνότητας

Η ραγδαία πρόοδος στο πεδίο της βαθιάς μάθησης οδήγησε στην ανάπτυξη νέων μεθόδων για το πρόβλημα. Αρχικά, έχει μελετηθεί εκτενώς κυρίως ως πρόβλημα παλινδρόμησης στο πεδίο της συχνότητας. Συγκεκριμένα, ένα βαθύ νευρωνικό δίκτυο εκπαιδεύεται ώστε να αντιστοιχεί την θορυβώδη είσοδο σε έξοδο καθαρής ομιλίας. Εναλλακτικά, η έξοδος μπορεί να είναι μια μάσκα $\mathbf{M}$ που εφαρμόζεται πολλαπλασιαστικά στο φάσμα του θορυβώδους σήματος $\mathbf{X}$. Δηλαδή, η έξοδος $\hat{\mathbf{X}}$ υπολογίζεται ως εξής:

$$\hat{\mathbf{X}} = \mathbf{M} \odot \mathbf{X}$$

όπου με $\odot$ συμβολίζουμε τον πολλαπλασιασμό στοιχείο προς στοιχείο. Συνήθως, η εκπαίδευση του μοντέλου γίνεται με πλήρως επιβλεπόμενο τρόπο, με θορυβώδη σήματα και τα αντίστοιχα σήματα "καθαρής" φωνής. Έτσι, εκπαιδεύοντας το μοντέλο σε ένα μεγάλο αριθμό δεδομένων, μαθαίνει την αντιστοίχιση επιλύοντας το πρόβλημα.

Ορισμένες φορές το πρόβλημα τίθεται ως πρόβλημα ταξινόμησης όπου υπολογίζεται μια δυαδική μάσκα στο πεδίο της συχνότητας, ταξινομώντας κάθε κελί χρόνου-συχνότητας σε φωνή ή θόρυβο. Η μάσκα στόχος ονομάζεται Ιδανική Δυαδική Μάσκα (Ideal Binary Mask ή IBM) (D. Wang and J. Chen 2018) και ορίζεται ως εξής:

$$\mathrm{IBM}(f,t) = \begin{cases} 1 & \text{αν } \mathrm{SNR}(f,t) > LC \\ 0 & \text{αλλιώς} \end{cases}$$

όπου SNR είναι ο σηματοθορυβικός λόγος (Signal-to-Noise Rate) σε dB και LC ένα κατώφλι για το SNR το οποίο συχνά επιλέγεται αυθαίρετα στα 0 dB (D. Wang 2005). Αν χαλαρώσουμε τον περιορισμό να είναι η μάσκα δυαδική, προκύπτει η Ιδανική Μάσκα Λόγου (Ideal Ratio Mask ή IRM) (D. Wang and J. Chen 2018) που ορίζεται ως εξής:

$$\mathrm{IRM}(f,t) = \left(\frac{S^2(f,t)}{S^2(f,t) + N^2(f,t)^2}\right)^{\beta} = \left(\frac{\mathrm{SNR}(f,t)}{\mathrm{SNR}(f,t)+1}\right)^{\beta}$$

όπου $S^2(f,t)$, $N^2(f,t)$ η ενέργεια των σημάτων φωνής και θορύβου στο αντίστοιχο κελί χρόνου-συχνότητας και β παράμετρος συνήθως ίση με 0.5 (D. Wang and J. Chen 2018).

Μια από τις πρώτες μεθόδους βασισμένη σε νευρωνικά δίκτυα (Lu et al. 2013), χρησιμοποιεί έναν autoencoder αποθορυβοποίησης (denoising autoencoder) με ένα κρυφό επίπεδο. Η είσοδος είναι το θορυβώδες φάσμα ισχύος Mel ενώ η έξοδος είναι το εκτιμώμενο "καθαρό" φάσμα ισχύος Mel. Στο (Maas et al. 2012) χρησιμοποιείται βαθύς autoencoder αποθορυβοποίησης με επίπεδα τύπου Αναδρομικού Νευρωνικού Δικτύου (Recurrent Neural Network ή RNN). Το RNN επεξεργάζεται το σήμα ως ακολουθία συνεπώς μπορεί να μοντελοποιήσει χρονικές εξαρτήσεις. Η είσοδος και έξοδος αυτή την φορά είναι οι Cepstrum συντελεστές συχνότητας Mel (Mel Frequency Cepstral Coefficients ή MFCCs). Έπειτα, στο (Weninger, Erdogan, et al. 2015), προτάθηκε η χρήση RNN τύπου Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) που αντιμετωπίζει μειονεκτήματα του RNN. Ακόμα, μια αλλαγή της μεθόδου αυτής είναι η πρόβλεψη μάσκας που εφαρμόζεται στον STFT κατευθείαν και όχι στο μέτρο αυτού. Ως αποτέλεσμα μπορεί να λάβει υπ' όψιν την φάση του σήματος κατά την εκπαίδευση.

Σήματα στο Πεδίο του Χρόνου

Πιο πρόσφατα, έχουν προταθεί μοντέλα τα οποία δέχονται την είσοδο και δίνουν την έξοδο στο πεδίο του χρόνου αντί για το πεδίο της συχνότητας, κάνοντας έτσι χρήση και της πληροφορίας της φάσης. Στο (Fu, Tsao, et al. 2017) οι συγγραφείς προτείνουν ένα Πλήρως Συνελικτικό Δίκτυο (Fully Convolutional Network ή FCN) που δρα κατευθείαν στην κυματομορφή. Αντίστοιχα, στο (Pandey and D. Wang 2019) προτείνεται η αρχιτεκτονική του Χρονικού Συνελικτικού Νευρωνικού Δικτύου (Temporal Convolutional Neural Network ή TCNN) το οποίο ξεπερνάει σε απόδοση προηγούμενες μεθόδους βασιζόμενες σε LSTM, μειώνοντας παράλληλα τον αριθμό των παραμέτρων του μοντέλου. Στο (Rethage, Pons, and Serra 2018) προσαρμόζεται το μοντέλο WaveNet (Oord et al. 2016) για το πρόβλημα της αποθορυβοποίησης σήματος φωνής.

Άλλες προσεγγίσεις χρησιμοποιούν Generative Adversarial Networks (GANs) (Goodfellow, Pouget-Abadie, et al. 2014) για την επίλυση του προβλήματος, όπως το SEGAN (Pascual, Bonafonte, and Serrà 2017) που εκπαιδεύεται με πλήρη επίβλεψη με σήματα στο πεδίο του χρόνου. Ο generator του GAN δέχεται ως είσοδο το θορυβώδες σήμα και δίνει ως έξοδο το εκτιμώμενο καθαρό. Ο discriminator του GAN δέχεται ως είσοδο ζεύγος με το θορυβώδες σήμα και το καθαρό ή ζεύγος με το θορυβώδες και την εκτίμηση του generator. Ο discriminator εκπαιδεύεται ώστε να ξεχωρίζει αν στο ζεύγος έχουμε εκτίμηση ή το πραγματικό καθαρό σήμα. Αντίθετα, ο generator εκπαιδεύεται ώστε να "ξεγελάει" τον discriminator σε συνδυασμό με ένα σφάλμα ανακατασκευής.

Μερικά από τα καλύτερα ως προς την απόδοση μέχρι στιγμής μοντέλα για το πρόβλημα της αποθορυβοποίησης σήματος φωνής είναι τα παρακάτω. Το MetricGAN (Fu, Liao, et al. 2019) το οποίο χρησιμοποιεί GANs και ξεπερνάει σε απόδοση το προηγούμενό SEGAN. Το DeepMMSE (Zhang et al. 2020) το οποίο χρησιμοποιεί βαθιά νευρωνικά δίκτυα και βασίζεται σε παλαιότερες τεχνικές που εκτιμούν τα φασματικά χαρακτηριστικά του

θορύβου, όπως η (Ephraim and Malah 1984). Τέλος, το Demucs (Defossez, Synnaeve, and Adi 2020) που προτάθηκε για αποθορυβοποίηση σε πραγματικό χρόνο και φτάνει σε απόδοση το DeepMMSE.

Μειωμένη Επίβλεψη

Όμως, οι πλήρως επιβλεπόμενες μέθοδοι παρουσιάζουν δυσκολίες γενίκευσης ως προς τον τύπο θορύβου και ως προς ακουστικές καταστάσεις που δεν αντιμετώπισαν κατά την εκπαίδευση. Έτσι, έχουν ερευνηθεί ημι-επιβλεπόμενες μέθοδοι οι οποίες δεν απαιτούν ζευγάρια θορυβωδών και καθαρών σημάτων.

- Στο (Fujimura et al. 2021) προτείνεται η μέθοδος εκπαίδευσης Noisy-target (NyTT) όπου παράγονται μείγματα αποτελούμενα από θορυβώδη σήματα μαζί με επιπλέον θόρυβο, ενώ στόχος του μοντέλου είναι η ανάκτηση του θορυβώδους σήματος.
- Στο πρόσφατο MetricGAN-U (Fu, C. Yu, et al. 2022) η εκπαίδευση πραγματοποιείται μόνο με θορυβώδη σήματα και την χρήση μετρικής που εκτιμά την ποιότητα σημάτων φωνής.
- Στο (Xiang and Bao 2020) έχουμε χρήση GAN που χρησιμοποιεί στην εκπαίδευση καθαρά σήματα φωνής και σήματα θορύβου τα οποία όμως δεν αποτελούν ζευγάρια.

Τέλος, οι περιπτώσεις αυτές διαφέρουν από την περίπτωση του εξετάζουμε στην ενότητα 3.3 αφού στο πρόβλημά μας δεν έχουμε πρόσβαση στο θόρυβο κατά την εκπαίδευση.

3.2 Μέθοδοι για το Πρόβλημα Διαχωρισμού Πηγών

Όπως εξηγήσαμε και στην εισαγωγή της εργασίας, το πρόβλημα διαχωρισμού πηγών αποτελεί ένα πολύ γενικό πρόβλημα, ακόμα και στην περίπτωση που τα σήματα ενδιαφέροντος είναι ηχητικά. Μπορούμε να χωρίσουμε το πρόβλημα σε υποπεριπτώσεις ανάλογα με τον αριθμό των πηγών που επιθυμούμε να εξάγουμε και τον αριθμό των καναλιών, τα οποία προκύπτουν από διαφορετικά μικρόφωνα που ηχογραφούν το μείγμα.

- Υπερ-καθορισμένο (over-determined) όταν έχουμε περισσότερα κανάλια σε σχέση με πηγές.
- Καθορισμένο (determined) όταν έχουμε τον ίδιο αριθμό καναλιών και πηγών.
- Υπο-καθορισμένο (under-determined) όταν έχουμε λιγότερα κανάλια σε σχέση με πηγές.

Το πρόβλημα που εξετάζουμε εμπίπτει στην τρίτη περίπτωση καθώς έχουμε μείγματα ενός καναλιού. Επομένως, θα παρουσιάσουμε μεθόδους κυρίως για την περίπτωση αυτή.

Στις δύο πρώτες περιπτώσεις το πρόβλημα είναι σημαντικά ευκολότερο και απλές μέθοδοι βασιζόμενες στην Ανάλυση Ανεξάρτητων Συνιστωσών (Independent Component Analysis ή ICA) (Hyvärinen and Oja 2000) έχουν δώσει ικανοποιητικά αποτελέσματα (Smaragdis 1998).

3.2.1 Μέθοδοι ΨΕΣ

Στη μέθοδο ICA γίνεται η υπόθεση ότι οι συνιστώσες είναι μεταξύ τους στατιστικά ανεξάρτητες. Πρακτικά, για τον υπολογισμό των συνιστωσών χρησιμοποιούμε επαναληπτικούς αλγόριθμους ώστε να ελαχιστοποιήσουμε ή να μεγιστοποιήσουμε ένα κριτήριο κόστους που σχετίζεται με την ιδιότητα του non-Gaussianity. Γνωρίζουμε από το Κεντρικό Οριακό Θεώρημα ότι η κατανομή πιθανότητας ενός αθροίσματος από ανεξάρτητες τυχαίες μεταβλητές τείνει σε κατανομή Gauss. Δηλαδή, το άθροισμα δύο ανεξάρτητων τυχαίων μεταβλητών συνήθως έχει κατανομή κοντινότερη στην κατανομή Gauss σε σχέση με τις κατανομές των τυχαίων μεταβλητών που προστίθενται. Έτσι, ταυτίζουμε την ανεξαρτησία με την ιδιότητα του non-Gaussianity. Συνεπώς, για να βρούμε ανεξάρτητες πηγές επιθυμούμε να μεγιστοποιήσουμε το non-Gaussianity τους. Τα μειονεκτήματα αυτής της μεθόδου είναι ότι δεν μπορεί να χρησιμοποιηθεί στην υπο-καθορισμένη περίπτωση και η υπόθεση του γραμμικού και στατικού μοντέλου μείξης, με βάση το οποίο η μείξη γίνεται με τρόπο ανεξάρτητο του χρόνου. (Greenberg 2007)

Μαθηματικά, υποθέτουμε ότι το μείγμα $\mathbf{X}^{(N\times T)}$ προκύπτει από τις πηγές $\mathbf{S}^{(P\times T)}$ μέσω του πίνακα μείξης $\mathbf{A}^{(N\times P)}$

$$\mathbf{X} = \mathbf{AS}$$

Ενώ για τον διαχωρισμό υποθέτουμε ότι οι γραμμές του $\hat{\mathbf{S}}$ είναι στατιστικά ανεξάρτητες και υπολογίζουμε τον πίνακα $\mathbf{V} \approx \mathbf{A}^{+}$ που προσεγγίζει τον ψευδοαντίστροφο του $\mathbf{A}$. Οπότε έχουμε:

$$\hat{\mathbf{S}} = \mathbf{VX}$$

Η Ανάλυση Ανεξάρτητων Υποχώρων (Independent Subspace Analysis ή ISA) βασίζεται στην ICA αλλά μπορεί να εφαρμοστεί και στην υπο-καθορισμένη περίπτωση (Casey and Westner 2000). Εργάζεται στον χώρο $\mathbb{R}^F$ που βρίσκονται τα διανύσματα, που προκύπτουν από κάθε χρονικό παράθυρο του φάσματος, από το μέτρο του STFT ($\mathbb{R}^{F\times T}$) και προσπαθεί να βρει στατιστικά ανεξάρτητους υποχώρους για κάθε πηγή. Δηλαδή, δυο τυχαία διανύσματα μεταξύ διαφορετικών υποχώρων είναι στατιστικά ανεξάρτητα αλλά τυχαία διανύσματα του ίδιου υποχώρου μπορεί να είναι εξαρτημένα. Προβάλλοντας από τον κάθε υποχώρο στον αρχικό χώρο παίρνουμε τα διαχωρισμένα σήματα.

Άλλες μέθοδοι που ανήκουν στην οικογένεια της Υπολογιστικής Ανάλυσης Ακουστικής Σκηνής (Computational Auditory Scene Analysis ή CASA) προσπαθούν να μιμηθούν τον τρόπο που το ανθρώπινο αυτί ερμηνεύει και διαχωρίζει τους ήχους. Για παράδειγμα, η μέθοδος από το (Hu and D. Wang 2010) υπολογίζει την θεμελιώδη συχνότητα του σήματος φωνής (pitch) που θέλουμε να εξάγουμε και με βάση αυτή την πληροφορία υπολογίζει μια δυαδική μάσκα στο πεδίο της συχνότητας. Η διαδικασία αυτή πραγματοποιείται επαναληπτικά με σκοπό την βελτίωση των εκτιμήσεων.

Μια άλλη οικογένεια μεθόδων μοντελοποιεί κάθε κατηγορία πηγών με ένα πιθανοτικό μοντέλο, όπως το Μοντέλο Μείγματος Γκαουσιανών (Gaussian Mixture Model ή GMM) στο (Kristjansson, Attias, and Hershey 2004) και το Κρυφό Μαρκοβιανό Μοντέλο (Hidden Markov Model ή HMM) στο (Roweis 2000). Τα μοντέλα αυτά εκπαιδεύονται μέσω "καθαρών" δεδομένων κάθε κατηγορίας πηγής. Όμως, κατά τον διαχωρισμό αποτυγχάνουν σε περίπτωση που τα επίπεδα ενέργειας των ήχων ή και οι συνθήκες ηχογράφησης είναι διαφορετικές σε σχέση με την εκπαίδευση.

Μια μέθοδος παραγοντοποίησης πίνακα η οποία είναι ιδιαίτερα δημοφιλής για προβλήματα διαχωρισμού πηγών είναι η μέθοδος NMF (Smaragdis, Fevotte, et al. 2014). Όπως εξηγήσαμε και στην ενότητα 2.4 ερμηνεύει τις στήλες ενός πίνακα με μη αρνητικά στοιχεία $\mathbf{X}$ ως ένα σταθμισμένο άθροισμα (με μη αρνητικούς συντελεστές) μη αρνητικών διανυσμάτων βάσης. Έτσι παραγοντοποιεί την αρχική αναπαράσταση ως μια σύνθεση θεμελιωδών κομματιών. Ως αποτέλεσμα, είναι ιδιαίτερα χρήσιμη για προβλήματα διαχωρισμού ηχητικών σημάτων καθώς μπορεί να χρησιμοποιηθεί σε οποιαδήποτε μη αρνητική αναπαράσταση χρόνου-συχνότητας. Για παράδειγμα, η μη αρνητική αναπαράσταση μπορεί να είναι το μέτρο ή η ισχύς του STFT.

Υπενθυμίζουμε ότι ο πίνακας με μη αρνητικά στοιχεία $\mathbf{X}$ προσεγγίζεται από τους $\mathbf{W}$ και $\mathbf{H}$ επίσης με μη αρνητικά στοιχεία.

$$\mathbf{X} \approx \mathbf{WH}$$

όπου ο πίνακας $\mathbf{W}$ ονομάζεται πίνακας βάσεων και ο πίνακας $\mathbf{H}$ ονομάζεται πίνακας ενεργοποιήσεων.

Μπορεί να εφαρμοστεί στην περίπτωση χωρίς επίβλεψη όπου καλούμαστε να διαχωρίσουμε ένα μείγμα χωρίς να έχουμε από πριν κάποια δεδομένα εκπαίδευσης, όπως πραγματοποιήσαμε στο απλό πείραμα στην ενότητα 2.4. Συγκεκριμένα, στο (Virtanen 2007) χρησιμοποιείται η μέθοδος NMF με συνάρτηση σφάλματος που περιέχει σφάλμα αραιότητας και ομαλότητας για τον πίνακα ενεργοποιήσεων. Έτσι, πραγματοποιείται ο διαχωρισμός μειγμάτων μουσικών κομματιών σε πηγές μουσικών οργάνων. Όμως, η ίδια μεθοδολογία αποτυγχάνει στην περίπτωση που έχουμε πιο περίπλοκες πηγές όπως σήματα φωνής. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί μέθοδοι όπως η (Duong, Ozerov, and Chevallier 2014). Η συγκεκριμένη μέθοδος στηρίζεται σε πληροφορία που δίνει ο χρήστης, ώστε να δώσει στο μοντέλο μια αρχική πληροφορία για την θέση των διαφόρων πηγών στο μείγμα, μέσω του πίνακα ενεργοποιήσεων.

Στην περίπτωση με επίβλεψη, διαθέτουμε "καθαρά" δεδομένα για κάθε κατηγορία πηγής. Πρώτα, εκπαιδεύουμε ένα NMF μοντέλο για κάθε πηγή και κρατάμε τους πίνακες βάσης που μάθαμε. Δηλαδή, στην περίπτωση που έχουμε δυο πηγές, έχουμε τους προ-εκπαιδευμένους πίνακες βάσεων $\mathbf{W}_1$ και $\mathbf{W}_2$. Στην συνέχεια, έχοντας ένα μείγμα $\mathbf{X}$ εφαρμόζουμε πάλι την μέθοδο NMF ώστε να βρούμε τον πίνακα ενεργοποιήσεων $\mathbf{H}$ ενώ κρατάμε τον πίνακα βάσεων σταθερό και ίσο με $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]$ (Smaragdis, Fevotte, et al. 2014; Mohammadiha, Smaragdis, and Leijon 2013). Δηλαδή έχουμε:

$$\mathbf{X} \approx [\mathbf{W}_1, \mathbf{W}_2]\,\mathbf{H}$$

Όπου $\mathbf{H} = \begin{bmatrix}\mathbf{H}_1 \\ \mathbf{H}_2\end{bmatrix}$ οπότε έχουμε:

$$\mathbf{X} \approx [\mathbf{W}_1, \mathbf{W}_2] \begin{bmatrix}\mathbf{H}_1 \\ \mathbf{H}_2\end{bmatrix} \implies \mathbf{X} \approx \mathbf{W}_1\mathbf{H}_1 + \mathbf{W}_2\mathbf{H}_2$$

Αφού υπολογιστούν τα $\mathbf{H}_1$ και $\mathbf{H}_2$ μπορούμε να υπολογίσουμε τις εκτιμώμενες πηγές, είτε απευθείας ως

$$\hat{\mathbf{S}}_1 = \mathbf{H}_1\mathbf{W}_1, \quad \hat{\mathbf{S}}_2 = \mathbf{H}_2\mathbf{W}_2$$

είτε με μάσκα

$$\hat{\mathbf{S}}_1 = \frac{\mathbf{H}_1\mathbf{W}_1}{\mathbf{H}_1\mathbf{W}_1 + \mathbf{H}_2\mathbf{W}_2} \odot \mathbf{X}, \quad \hat{\mathbf{S}}_2 = \frac{\mathbf{H}_2\mathbf{W}_2}{\mathbf{H}_1\mathbf{W}_1 + \mathbf{H}_2\mathbf{W}_2} \odot \mathbf{X}$$

Στην περίπτωση πού η μη αρνητική αναπαράσταση είναι η ισχύς του STFT η μάσκα πραγματοποιεί φιλτράρισμα Wiener. Συνήθως χρησιμοποιείται και σφάλμα αραιότητας για τον πίνακα ενεργοποιήσεων ώστε να αποφύγουμε πολύ γενικά διανύσματα βάσεων κατά την εκπαίδευση (Le Roux, Weninger, and Hershey 2015).

Στην ενότητα 3.3 περιγράφουμε την ημι-επιβλεπόμενη περίπτωση.

Τέλος, παρόλο που η NMF έχει το πλεονέκτημα ότι διαθέτει μια διαισθητική ερμηνεία, η απλότητά της δεν της επιτρέπει να λάβει υπ' όψιν όλα τα πολύπλοκα χαρακτηριστικά των ηχητικών σημάτων. Ως αποτέλεσμα, έχουν ερευνηθεί επεκτάσεις της, όπως η χρήση συνελικτικών βάσεων (Smaragdis 2006), καθώς και συνδυασμοί με άλλες τεχνικές, όπως HMMs (Mysore, Smaragdis, and Raj 2010).

3.2.2 Μέθοδοι Βασιζόμενες σε Νευρωνικά Δίκτυα

Όπως και στην περίπτωση του προβλήματος αποθορυβοποίησης σήματος φωνής, η πρόοδος των τεχνικών βαθιάς μάθησης οδήγησε στην ανάπτυξη νέων μεθόδων για το πρόβλημα διαχωρισμού πηγών.

Πρώτα, αξίζει να αναφερθούμε σε μεθόδους που επιχειρούν να συνδυάσουν νευρωνικά δίκτυα με NMF. Μια τέτοια μέθοδος είναι η Deep NMF (Le Roux, Hershey, and Weninger 2015) όπου η μέθοδος NMF με K επαναλήψεις ενημέρωσης των παραμέτρων, αναπαρίσταται ως ένα νευρωνικό δίκτυο βάθους $K+1$ χωρίς διαφορά ανάμεσα στα επίπεδά του. Το νευρωνικό δίκτυο αυτό λαμβάνει ως είσοδο τον αρχικό πίνακα ενεργοποίησης ενώ το κάθε επίπεδο έχει ως παραμέτρους ένα πίνακα βάσεων. Η ενημέρωση των παραμέτρων γίνεται με ένα πολλαπλασιαστικό κανόνα που βασίζεται στις παραγώγους του σφάλματος ως προς τις παραμέτρους του δικτύου. Ο υπολογισμός των παραμέτρων πραγματοποιείται με την τεχνική μετάδοσης σφάλματος προς τα πίσω. Επίσης, η NMF έχει αποτελέσει έμπνευση για τους NAE (Smaragdis and Venkataramani 2017) τους οποίους αναλύουμε στο Κεφάλαιο 4.

Τα πρώτα νευρωνικά δίκτυα που προτάθηκαν για το πρόβλημα ήταν σχετικά απλά και χωρίς πολύ βάθος. Για παράδειγμα, στο (Grais, Sen, and Erdogan 2014) προτείνεται ένα νευρωνικό δίκτυο το οποίο λαμβάνοντας ως είσοδο "καθαρό" καρέ του μέτρου του STFT, το ταξινομεί δυαδικά στην αντίστοιχη πηγή. Κατά τον διαχωρισμό, χρησιμοποιείται το δίκτυο αυτό ώστε να γίνει εκμάθηση δυο διανυσμάτων καρέ που αντιστοιχούν σε κάθε πηγή. Στο (Y. Wang and D. Wang 2013) έχουμε συνδυασμό μεθόδων CASA με νευρωνικό δίκτυο προβλέποντας έτσι

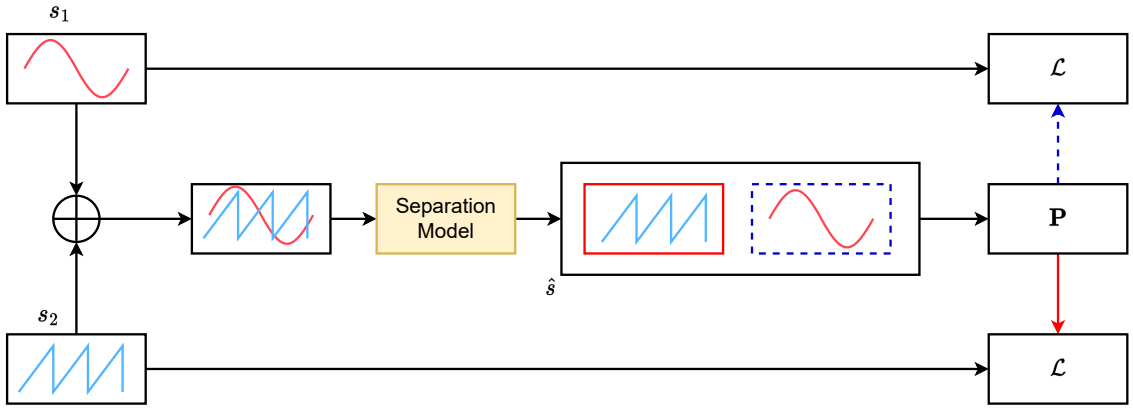
μια μάσκα IBM. Έπειτα, προτάθηκαν μοντέλα βασισμένα σε RNNs εκμεταλλευόμενα χρονικές εξαρτήσεις και βελτιώνοντας έτσι αισθητά την απόδοση (Weninger, Hershey, et al. 2014). Για παράδειγμα, στο (Huang et al. 2015) προτείνεται ένα RNN που λαμβάνει ως είσοδο το διάνυσμα καρέ του μείγματος και δίνει ως έξοδο δυο διανύσματα ως εκτιμήσεις, μια για κάθε πηγή. Η έξοδος παράγεται μέσω μιας μάσκα παρόμοια με Wiener που εφαρμόζεται στο μείγμα.

Όμως, οι παραπάνω μέθοδοι συνήθως έχουν σταθερό αριθμό εξόδων, κάθε μια από τις οποίες είναι αφιερωμένη σε μια κλάση ήχων. Επομένως στην περίπτωση που οι δυο κλάσεις είναι παρόμοιες όπως για παράδειγμα δυο διαφορετικοί ομιλητές, οι μέθοδοι αυτές αποτυγχάνουν όταν καλούνται να διαχωρίσουν μείγματα με ομιλητές που δεν έχουν συναντήσει κατά την εκπαίδευση. Στο (Hershey et al. 2016) προτάθηκε η μέθοδος Βαθιάς Συσταδοποίησης (Deep Clustering) ως λύση σε αυτό το πρόβλημα. Εμπνευσμένη από τεχνικές CASA χρησιμοποιεί ένα νευρωνικό δίκτυο ώστε να αντιστοιχίσει κάθε κελί χρόνου-συχνότητας του μείγματος με ένα διάνυσμα χαρακτηριστικών, με βάση το οποίο μετά τα ομαδοποιεί σε πηγές.

Για την αντιμετώπιση του προβλήματος των εξόδων των μοντέλων που ταυτίζονται με μια συγκεκριμένη κλάση, προτάθηκε η τεχνική εκπαίδευσης με το όνομα Εκπαίδευση Ανεξάρτητη Μετάθεσης (Permutation Invariant Training ή PIT) (D. Yu et al. 2017). Συγκεκριμένα, κατά την εκπαίδευση δοκιμάζονται όλες οι μεταθέσεις των εκτιμώμενων πηγών με τις πηγές στόχους και επιλέγεται αυτή με το μικρότερο σφάλμα. Μαθηματικά ορίζεται, για την περίπτωση που το σφάλμα λαμβάνεται στο πεδίο του χρόνου, ως εξής:

$$\mathcal{L}_{\text{PIT}}(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \min_{\mathbf{P}} \sum_{i=1}^{N} \mathcal{L}\left(s_i, [\mathbf{P}\hat{\boldsymbol{s}}]_i\right)$$

όπου $\boldsymbol{s} = \begin{bmatrix} s_1(t) & \dots & s_N(t) \end{bmatrix}^\intercal$ οι πηγές στόχοι και $\hat{\boldsymbol{s}} = \begin{bmatrix} \hat{s}_1(t) & \dots & \hat{s}_N(t) \end{bmatrix}^\intercal$ οι εκτιμήσεις των πηγών και $\mathbf{P}$ πίνακας μετάθεσης μεγέθους $N \times N$. Με $\mathcal{L}$ συμβολίζουμε την συνάρτηση σφάλματος που χρησιμοποιούμε, η οποία εφαρμόζεται στο επίπεδο σήματος. Αφότου προτάθηκε αποτελεί μέρος των περισσότερων μοντέλων νευρωνικών δικτύων που αναπτύσσονται, ειδικά για την επίλυση του προβλήματος διαχωρισμού σημάτων ομιλίας καθώς στην περίπτωση αυτή δεν υπάρχουν σταθερά χαρακτηριστικά ανά πηγή.



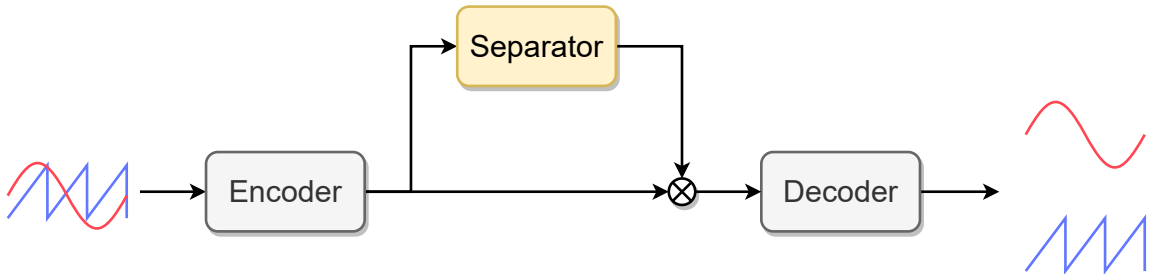
Σχήμα 3.2.1: Σχηματική αναπαράσταση της τεχνικής PIT.

Τα τελευταία χρόνια, αρκετά μοντέλα που αναπτύσσονται, ειδικά στην περίπτωση διαχωρισμού ομιλίας, λαμβάνουν είσοδο και δίνουν έξοδο στο πεδίο του χρόνου. Μια αρχιτεκτονική που λειτουργεί στο πεδίο του χρόνου και έχει αποτελέσει βάση για αρκετά μοντέλα που ακολούθησαν είναι το TasNet (Luo and Mesgarani 2018). Αποτελείται από τρία μέρη τον κωδικοποιητή (encoder), τον διαχωριστή (separator) και τον αποκωδικοποιητή, όπως φαίνεται στο Σχήμα 3.2.2.

- Ο κωδικοποιητής λαμβάνει ως είσοδο την κυματομορφή και δίνει ως έξοδο μια δισδιάστατη αναπαράσταση που μπορεί να ερμηνευθεί κατ' αντιστοιχία με μια χρονοσυχνοτική αναπαράσταση.

- Με βάση αυτή την αναπαράσταση ο διαχωριστής παράγει μάσκες για τις πηγές οι οποίες εφαρμόζονται πολλαπλασιαστικά σε αυτή.
- Ο αποκωδικοποιητής μετατρέπει τις αναπαραστάσεις των πηγών σε κυματομορφές.

Στην μορφή που προτάθηκε ο διαχωριστής είναι ένα βαθύ δίκτυο LSTM ακολουθούμενο από ένα πλήρως συνδεδεμένο επίπεδο, ενώ ο κωδικοποιητής και ο αποκωδικοποιητής αποτελούνται από ένα συνελικτικό επίπεδο ο καθένας. Η εκπαίδευση του μοντέλου πραγματοποιείται με PIT.



Σχήμα 3.2.2: Σχηματική αναπαράσταση του μοντέλου TasNet (Luo and Mesgarani 2018) όπου φαίνονται τα δομικά του μέρη. Αριστερά έχουμε το μείγμα κυματομορφών και στα δεξιά τις διαχωρισμένες κυματομορφές.

Η αρχιτεκτονική αυτή αποτέλεσε πρότυπο για μοντέλα που ακολούθησαν τα οποία ανέβασαν τον πήχη ως προς την απόδοση σε επιβλεπόμενα προβλήματα διαχωρισμού. Συγκεκριμένα, στο Dual-Path RNN (Luo, Z. Chen, and Yoshioka 2020) βελτιώνεται ο διαχωριστής, αυξάνοντας την απόδοση αλλά και την υπολογιστική πολυπλοκότητα παράλληλα. Το Conv-TasNet (Luo and Mesgarani 2019) χρησιμοποιώντας το Χρονικό Συνελικτικό Δίκτυο (Temporal Convolutional Network ή TCN) (Lea et al. 2016) καταφέρνει να επιταχύνει την εκπαίδευση. Ακόμη, καταφέρνει να ξεπεράσει την απόδοση των ιδανικών μασκών του μέτρου του STFT. Αντλώντας έμπνευση από το TasNet, στο (Venkataramani, Casebeer, and Smaragdis 2018) προτείνεται η χρήση ενός κωδικοποιητή πριν και ενός αποκωδικοποιητή μετά το κύριο μοντέλο, επιτρέποντας έτσι την προσαρμογή μοντέλων σχεδιασμένα για το πεδίο της συχνότητας στο πεδίο του χρόνου. Ο κωδικοποιητής και αποκωδικοποιητής εκπαιδεύονται μαζί με το μοντέλο.

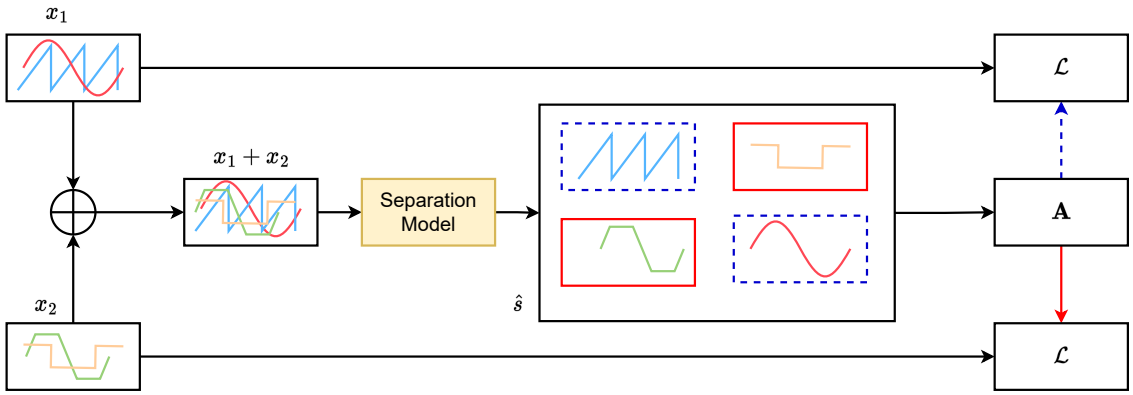
Υπογραμμίζουμε ότι, τα μοντέλα TasNet, Conv-TasNet και Dual-Path RNN προτάθηκαν για το πρόβλημα του διαχωρισμού σημάτων ομιλίας με ομιλητές που μιλούν ταυτόχρονα. Δηλαδή, επιλύουν ένα δυσκολότερο πρόβλημα από τον διαχωρισμό σήματος φωνής από θορυβώδες σήμα. Ως αποτέλεσμα, αυτά τα μοντέλα και οι επεκτάσεις τους αποδίδουν σε πολύ υψηλό επίπεδο για το πρόβλημα αυτό (D. Wang and J. Chen 2018).

Ένα άλλο πρόσφατο και επιτυχημένο μοντέλο διαχωρισμού πηγών είναι το Wave-U-Net (Stoller, Ewert, and Dixon 2018) που πρόκειται για προσαρμογή του U-Net (Ronneberger, Fischer, and Brox 2015). Το μοντέλο αυτό επεξεργάζεται τα δεδομένα σε διαφορετικές κλίμακες ενώ υπολογίζει κατευθείαν τις πηγαίες κυματομορφές χωρίς εφαρμογή μάσκας.

Τέλος, ως επέκταση της εκπαίδευσης PIT προτάθηκε η τεχνική εκπαίδευσης MixIT (Wisdom et al. 2020). Χαρακτηρίζεται ως μη επιβλεπόμενη εκπαίδευση αφού απαιτεί μόνο μείγματα και όχι ζεύγη μειγμάτων και των αντίστοιχων πηγών. Έχοντας τα μείγματα $x_1(t)$ και $x_2(t)$ το μοντέλο που εκπαιδεύουμε λαμβάνει ως είσοδο το $x(t) = x_1(t) + x_2(t)$ και δίνει ως έξοδο τις πηγές $\hat{\boldsymbol{s}}$. Το πλήθος των πηγών που προβλέπει το μοντέλο είναι N. Το σφάλμα MixIT υπολογίζεται ως εξής:

$$\mathcal{L}_{\text{MixIT}}\left(x_1, x_2, \hat{\boldsymbol{s}}\right) = \min_{\mathbf{A}} \sum_{i=1}^{2} \mathcal{L}\left(x_i, [\mathbf{A}\boldsymbol{s}]_i\right)$$

όπου $\mathcal{L}$ η συνάρτηση σφάλματος που χρησιμοποιούμε. Με $\mathbf{A}$ έχουμε τον πίνακα μείξης με τιμές 0 ή 1 και μέγεθος $2 \times N$ ενώ κάθε στήλη του πρέπει να αθροίζει σε 1. Αναζητάμε εξαντλητικά τον πίνακα $\mathbf{A}$ ώστε να συνδυάσουμε τις εξόδους του μοντέλου σε δυο μείγματα εκτιμήσεις, $\hat{x}_1(t)$ και $\hat{x}_2(t)$, τα οποία συγκρίνουμε με



Σχήμα 3.2.3: Σχηματική αναπαράσταση της τεχνικής MixIT.

τα μείγματα εισόδου $x_1(t)$ και $x_2(t)$. Η παραπάνω διαδικασία φαίνεται στο Σχήμα 3.2.3.

3.3 Μέθοδοι για το Ημι-Επιβλεπόμενο Πρόβλημα

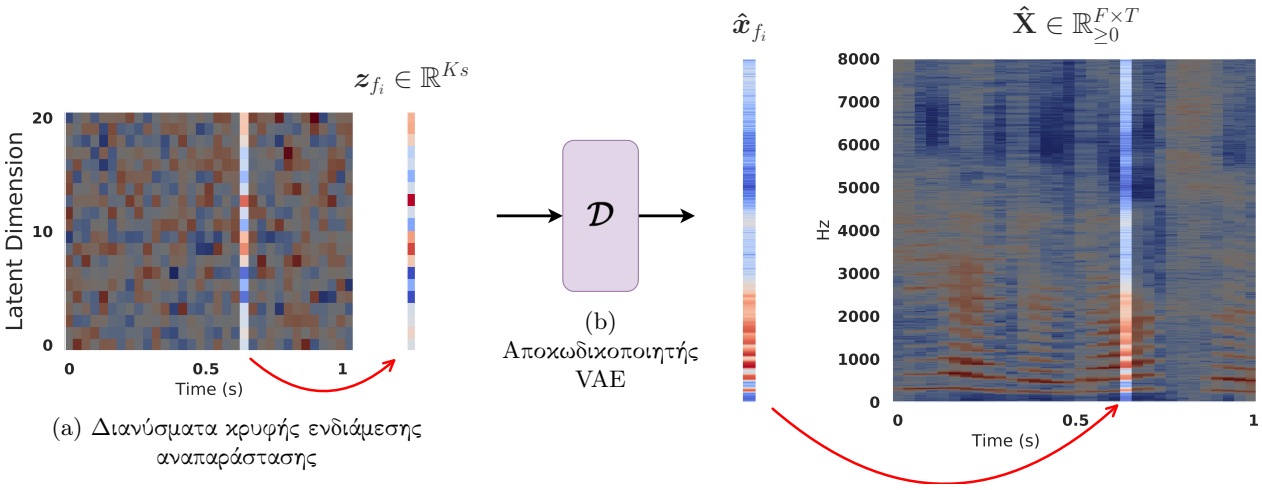
Στην παρούσα ενότητα εξετάζουμε μεθόδους που έχουν προταθεί για το ημι-επιβλεπόμενο πρόβλημα διαχωρισμού φωνής καθώς και γενικότερα για το ημι-επιβλεπόμενο πρόβλημα διαχωρισμού πηγών ηχητικών σημάτων, στην οποία περίπτωση δεν έχουμε γνώση για το θόρυβο ή την δεύτερη κλάση ήχων αντίστοιχα.

Από τις πρώτες μεθόδους που προτάθηκαν για το πρόβλημα είναι η Probabilistic Latent Component Analysis (PLCA) (Smaragdis, Raj, and Shashanka 2007) η οποία είναι ουσιαστικά μια αναδιατύπωση της μεθόδου NMF (Smaragdis, Fevotte, et al. 2014). Πρώτα, χρησιμοποιώντας την NMF μαθαίνουμε ένα πίνακα βάσεων στο σύνολο εκπαίδευσης που αποτελείται από την κλάση ήχων που μας ενδιαφέρει, όπως σήματα φωνής στην περίπτωσή μας. Στη συνέχεια, εφαρμόζουμε την NMF στο μείγμα που επιθυμούμε να διαχωρίσουμε με μεγαλύτερο πίνακα βάσεων αυτή την φορά αφιερώνοντας κάποιες στήλες για τις υπόλοιπες κλάσεις ήχων. Όμως, σε κάθε επανάληψη κρατάμε σταθερές τις στήλες του πίνακα βάσεων που είναι αφιερωμένες στη φωνή και είχαμε μάθει στο προηγούμενο βήμα. Επιπλέον, η συγκεκριμένη μέθοδος προσθέτει περιορισμό αραιότητας στον πίνακα ενεργοποιήσεων. Στην ενότητα 4.6.1 γίνεται εκτεταμένη παρουσίαση της μεθόδου NMF.

Όπως και στην περίπτωση με πλήρη επίβλεψη έχουν προταθεί επεκτάσεις της NMF για την ημι-επιβλεπόμενη περίπτωση. Συγκεκριμένα, στο (Mysore and Smaragdis 2011) οι συγγραφείς πρότειναν μια μέθοδο που συνδυάζει HMMs με την NMF ώστε να ληφθούν υπ' όψιν τα χρονικά χαρακτηριστικά και οι χρονικές εξαρτήσεις των σημάτων φωνής.

Πιο πρόσφατα, η ραγδαία ανάπτυξη των τεχνικών βαθιάς μάθησης οδήγησε στην ανάπτυξη νέων μεθόδων για το πρόβλημα. Πρόσφατες ερευνητικές προσεγγίσεις, οι οποίες ξεπερνάνε την απόδοση της NMF, συνδυάζουν την ισχύ μοντελοποίησης των βαθιών νευρωνικών δικτύων με παλαιότερες μεθόδους όπως η NMF (Bando et al. 2018; Leglaive, Girin, and Horaud 2018; Pariente, Deleforge, and Vincent 2019). Η κύρια ιδέα στην οποία βασίζονται είναι ο συνδυασμός του αποκωδικοποιητή ενός προ-εκπαιδευμένου Variational Autoencoder (VAE) (Kingma and Welling 2013) ως μοντέλο παραγωγής σημάτων φωνής (Σχήμα 3.3.1), μαζί με ένα μη επιβλεπόμενο μοντέλο NMF για τον θόρυβο. Το πρώτο βήμα που είναι κοινό, είναι η εκπαίδευση του μοντέλου VAE σε καθαρά σήματα ομιλίας στο πεδίο της συχνότητας. Έπειτα, κατά τον διαχωρισμό ακολουθείται μια επαναληπτική διαδικασία. Η διαφορά ανάμεσα σε αυτές τις προσεγγίσεις έγκειται στον αλγόριθμο που ενημερώνει τις παραμέτρους του NMF μοντέλου και την κρυφή μεταβλητή που αποτελεί είσοδο στον αποκωδικοποιητή του VAE. Επειδή ο αλγόριθμος διαχωρισμού είναι επαναληπτικός αυτές οι μέθοδοι είναι σχετικά αργές.

Με αφετηρία τις παραπάνω προσεγγίσεις που χρησιμοποιούν νευρωνικά δίκτυα έχουν προταθεί επεκτάσεις τους. Στο άρθρο (Leglaive, Alameda-Pineda, et al. 2020) οι συγγραφείς αντικατέστησαν το VAE με ένα Αναδρομικό



(a) Διανύσματα κρυφής ενδιάμεσης αναπαράστασης

(b) Αποκωδικοποιητής VAE

(c) Ανακατασκευασμένο μέτρο STFT

Σχήμα 3.3.1: Χρήση αποκωδικοποιητή VAE για την παραγωγή σήματος φωνής.

VAE (Recurrent VAE), το οποίο ανήκει στην γενική κατηγορία των Δυναμικών VAEs (Dynamical VAEs) (Girin et al. 2021), ώστε να μοντελοποιηθούν τα χρονικά χαρακτηριστικά του σήματος φωνής. Επίσης, κατά τον διαχωρισμό αντί να εκτιμούν κατευθείαν την κρυφή μεταβλητή $\boldsymbol{z}$, προσαρμόζουν τα βάρη του κωδικοποιητή του VAE. Ακόμη, στο (Nugraha, Sekiguchi, and Yoshii 2020) προτείνεται ο συνδυασμός VAE με Normalizing Flow (Taal et al. 2011; Papamakarios et al. 2021) για το μοντέλο ομιλίας. Τέλος, επεκτάσεις έχουν προταθεί για την περίπτωση που τα σήματα αποτελούνται από περισσότερα από ένα κανάλια (Leglaive, Girin, and Horaud 2019; Sekiguchi et al. 2018).

Όλες οι μέθοδοι που αναφέραμε εργάζονται με το μέτρο ή την ισχύ του STFT και δεν χρησιμοποιούν πληροφορία φάσης, βασίζοντας την απόφαση αυτή σε ισχυρισμούς ότι η φάση έχει δευτερεύουσα σημασία στο πρόβλημα (D. Wang and Lim 1982). Ο μετασχηματισμός του εκτιμώμενου σήματος φωνής στο πεδίο του χρόνου γίνεται με χρήση της φάσης του μείγματος.

3.4 Μετρικές Αξιολόγησης

Η αξιολόγηση της επίδοσης μιας μεθόδου προϋποθέτει την ύπαρξη και χρήση μιας μετρικής κατάλληλη για το πρόβλημα που προσπαθεί να επιλύσει η μέθοδος. Στην περίπτωση των προβλημάτων του Διαχωρισμού Πηγών σε ηχητικά σήματα και της Αποθορυβοποίησης Σήματος Φωνής η έρευνα πάνω στην ανάπτυξη μετρικών αξιολόγησης είναι ιδιαίτερα πλούσια (Vincent, Gribonval, and Févotte 2006; Rix et al. 2001; Taal et al. 2011; Le Roux, Wisdom, et al. 2019).

Ιδανικά θα επιθυμούσαμε η μετρική αξιολόγησης να ανταποκρίνεται στην ανθρώπινη ακουστική αντίληψη και ταυτόχρονα να είναι απλή στην υλοποίηση. Παλαιότερα, λόγω απλής υλοποίησης, χρησιμοποιούνταν το μέσο τετραγωνικό σφάλμα (MSE) ανάμεσα σε κανονικοποιημένα σήματα (Vincent, Gribonval, and Févotte 2006), χωρίς όμως να έχουμε καλή αντιστοίχιση με την ανθρώπινη αντίληψη. Αντίθετα για σήματα ομιλίας, έχουν αναπτυχθεί μετρικές όπως η PESQ (Rix et al. 2001) που εκτιμά την ποιότητα της ομιλίας με βάση μοντέλο αντίληψής της, και η STOI (Taal et al. 2011) που εκτιμά την ευκολία κατανόησης της ομιλίας. Συνήθως όμως η υλοποίηση αυτών των μετρικών δεν είναι απλή.

Μια μετρική αξιολόγησης που χρησιμοποιείται ευρέως στη βιβλιογραφία είναι η Signal-to-Distortion Ratio (SDR) που προτάθηκε ως μέρος του BSS Eval toolkit (Vincent, Gribonval, and Févotte 2006) και μετριέται σε decibel (dB). Εκτιμά την ποιότητα του σήματος με βάση ένα λόγο ενεργειών. Στην περίπτωση μας έχοντας ένα μείγμα

$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}$ στο πεδίο του χρόνου, ορίζεται ανάμεσα σε στο σήμα στόχο $\boldsymbol{s}$ και το εκτιμώμενο σήμα $\hat{\boldsymbol{s}}$ ως εξής:

$$\text{SDR} = 10 \log_{10} \left(\frac{\|\boldsymbol{s}\|^2}{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2} \right)$$

Όμως, η παραπάνω μετρική αξιολόγησης παρά την ευρεία χρήση, έχει ένα σημαντικό μειονέκτημα. Συγκεκριμένα, εξαρτάται από πλάτος του εκτιμώμενου σήματος $\hat{\boldsymbol{s}}$. Δηλαδή, πολλαπλασιάζοντας το σήμα $\hat{\boldsymbol{s}}$ με μια σταθερά μπορεί η μετρική αυτή να δώσει καλύτερο ή χειρότερο αποτέλεσμα. Το γεγονός αυτό στην περίπτωσή μας είναι ανεπιθύμητο επειδή το αποτέλεσμα της μετρικής μεταβάλλεται χωρίς όμως να διαφέρει η ανθρώπινη αντίληψη του ηχητικού σήματος.

Επομένως, για την επίλυση του προβλήματος που παρουσιάζεται έχει προταθεί η μετρική Scale Invariant Signal-to-Distortion Ratio (SI-SDR) που επίσης μετριέται σε decibel (dB) (Le Roux, Wisdom, et al. 2019).

$$\text{SI-SDR} = 10 \log_{10} \left(\frac{\|\alpha \boldsymbol{s}\|^2}{\|\alpha \boldsymbol{s} - \hat{\boldsymbol{s}}\|^2} \right), \quad \alpha = \arg\max_{\alpha} \|\alpha \boldsymbol{s} - \hat{\boldsymbol{s}}\|^2 = \frac{\hat{\boldsymbol{s}}^\intercal \boldsymbol{s}}{\|\boldsymbol{s}\|^2}$$

Πρακτικά, πολλαπλασιάζουμε το σήμα στόχο s με τον παράγοντα α ώστε τα σήματα $\alpha \boldsymbol{s}$ και $\alpha \boldsymbol{s} - \hat{\boldsymbol{s}}$ να είναι ορθογώνια μεταξύ τους. Συνεπώς, ο παράγοντας α διασφαλίζει ότι ο λόγος και κατ' επέκταση η μετρική θα παραμένει αναλλοίωτη όταν είτε το σήμα $\hat{\boldsymbol{s}}$ είτε το $\boldsymbol{s}$ πολλαπλασιάζονται με σταθερό παράγοντα.

3.5 Σύνολα Δεδομένων

Στα πλαίσια της εργασίας χρησιμοποιούμε τα σύνολα δεδομένων TIMIT (Garofolo et al. 1993), DEMAND (Thiemann, Ito, and Vincent 2013) και MUSDB18 (Rafii et al. 2017). Το TIMIT χρησιμοποιείται ως σύνολο δεδομένων "καθαρής" ομιλίας, ενώ τα άλλα δύο ως σύνολα προσθετικού θορύβου. Συγκεκριμένα:

TIMIT

Το σύνολο δεδομένων TIMIT (Garofolo et al. 1993) αποτελείται από 6300 καθαρά σήματα ομιλίας συνολικής διάρκειας 4 ωρών από 630 ομιλητές και 2342 διακριτές προτάσεις. Επίσης, τα δεδομένα προέρχονται από 8 διαφορετικές διαλέκτους των Αγγλικών Αμερικής. Τα σήματα του είναι μονοφωνικά με ρυθμό δειγματοληψίας 16 kHz.

Είναι χωρισμένο σε σύνολο train μεγέθους 4620 δειγμάτων από 462 ομιλητές και σε σύνολο test μεγέθους 1344 δειγμάτων με 168 ομιλητές. Ανάμεσα στα δύο αυτά σύνολα καμία πρόταση και κανένας ομιλητής δεν εμφανίζεται και στα δύο. Στο σύνολο test μπορούν να προστεθούν 2 προτάσεις ανά ομιλητή δηλαδή επιπλέον 336 δείγματα, οι οποίες όμως προτάσεις υπάρχουν και στο σύνολο train.

DEMAND

Το σύνολο δεδομένων DEMAND: Diverse Environments Multichannel Acoustic Noise Database (Thiemann, Ito, and Vincent 2013) περιλαμβάνει σήματα θορύβων από διάφορα περιβάλλοντα. Οι ήχοι είναι χωρισμένοι στις κατηγορίες: Οικία, Φύση, Γραφείο, Δημόσιος Χώρος, Δρόμος και Μέσα Μαζικής Μεταφοράς. Κάθε κατηγορία αποτελείται από 3 σήματα διάρκειας 5 λεπτών, ενώ κάθε σήμα ηχογραφείται από μια διάταξη 16 μικροφώνων και άρα προκύπτουν 16 κανάλια. Ο ρυθμός δειγματοληψίας των ήχων είναι 48 kHz.

MUSDB18

Το σύνολο δεδομένων MUSDB18 (Rafii et al. 2017) αποτελείται από 150 μουσικά κομμάτια διαφόρων μουσικών ειδών. Συνολικά, τα κομμάτια έχουν διάρκεια περίπου 10 ωρών, ενώ χωρίζονται σε σύνολο train και σύνολο test με 100 και 50 κομμάτια αντίστοιχα. Τα σήματα είναι στερεοφωνικά με ρυθμό δειγματοληψίας 44.1 kHz. Για

κάθε κομμάτι έχουμε τα σήματα από τα φωνητικά, τα τύμπανα, το μπάσο και ένα σήμα με τα υπόλοιπα μουσικά όργανα, αθροίζοντας τα προκύπτει το μείγμα που είναι το σήμα του κομματιού.

Κεφάλαιο 4

Non Negative Autoencoders

4.1 Περιγραφή Μοντέλου

Όπως είδαμε στην ενότητα 2.4 η μέθοδος NMF προσεγγίζει ένα πίνακα με μη αρνητικές τιμές $\mathbf{X} \in \mathbb{R}_{\geq 0}$ διαστάσεων $M \times N$ ως γινόμενο δυο πινάκων μικρότερης τάξης, $\mathbf{W} \in \mathbb{R}_{\geq 0}^{M \times K}$ και $\mathbf{H} \in \mathbb{R}_{\geq 0}^{K \times N}$.

$$\mathbf{X} \approx \mathbf{W} \cdot \mathbf{H}$$

Η μέθοδος Non Negative Autoencoder (NAE) προτάθηκε ως γενίκευση της NMF (Smaragdis and Venkataramani 2017). Στην βασική του μορφή αποτελείται από έναν Autoencoder με δυο γραμμικά επίπεδα.

$$\begin{aligned} 1^{\text{st}} \text{ layer}: &\quad \mathbf{H} = g\left(\mathbf{W}^{\dagger} \cdot \mathbf{X}\right) \\ 2^{\text{nd}} \text{ layer}: &\quad \hat{\mathbf{X}} = g\left(\mathbf{W} \cdot \mathbf{H}\right) \end{aligned}$$

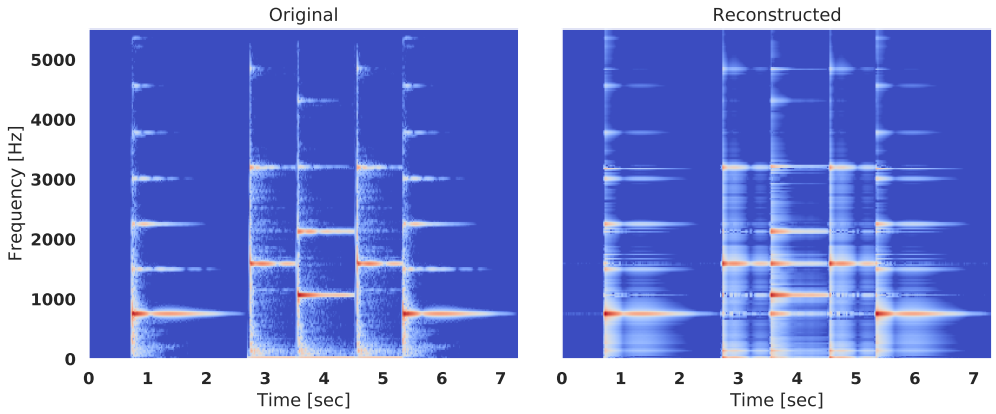
όπου $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ συνάρτηση που δίνει έξοδο μη αρνητικές τιμές. Συνήθως, χρησιμοποιούνται συναρτήσεις όπως ReLU $g(x) = \max(x, 0)$, softplus $g(x) = \log(1 + e^x)$ ή ακόμη και η απόλυτη τιμή $g(x) = |x|$. Συνεπώς, η χρήση της συνάρτησης g ως μη γραμμικότητα διασφαλίζει ότι η ενδιάμεση αναπαράσταση $\mathbf{H}$ θα είναι μη αρνητική, όπως και η ανακατασκευασμένη έξοδος $\hat{\mathbf{X}}$. Σε αντίθεση με την μέθοδο NMF ο πίνακας βάσεων $\mathbf{W}$ ενδέχεται να περιέχει και αρνητικές τιμές. Οι πίνακες $\mathbf{W}$ και $\mathbf{H}$ έχουν μέγεθος $M \times K$ και $K \times N$ αντίστοιχα, όπου ο θετικός ακέραιος αριθμός K ορίζει την τάξη του μοντέλου.

Για δεδομένο πίνακα $\mathbf{X}$ το μοντέλο εκπαιδεύεται σαν autoencoder, ώστε να ελαχιστοποιεί κάποια συνάρτηση σφάλματος ανάμεσα στον πίνακα $\mathbf{X}$ και την ανακατασκευή $\hat{\mathbf{X}}$. Το πρώτο επίπεδο αποτελεί τον κωδικοποιητή και το δεύτερο τον αποκωδικοποιητή του autoencoder. Ερμηνεύουμε τις παραμέτρους του δεύτερου επιπέδου $\mathbf{W}$ ως πίνακα βάσεων και την ενδιάμεση αναπαράσταση $\mathbf{H}$ ως πίνακα ενεργοποιήσεων.

Για να αξιολογήσουμε διαισθητικά την συμπεριφορά του μοντέλου σε σχέση με την μέθοδο NMF επαναλαμβάνουμε το πείραμα ανακατασκευής μιας ακολουθίας από τρεις νότες πιάνου. Επιλέγουμε μοντέλο NAE τάξης $K = 3$ με την απόλυτη τιμή για την συνάρτηση g και για συνάρτηση σφάλματος εργαζόμαστε με τη εξής:

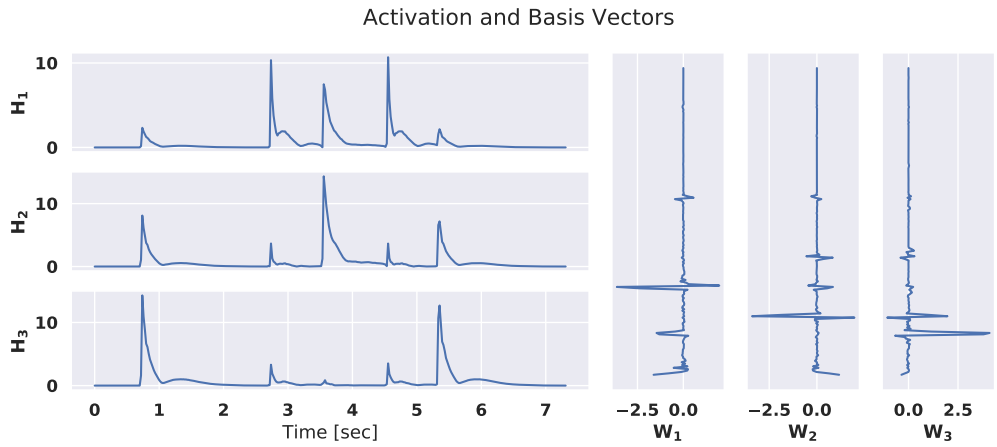
$$\mathcal{L}\left(\mathbf{X}, \hat{\mathbf{X}}\right) = \sum_{i,j} \left(\mathbf{X}_{i,j} \log\left(\frac{\mathbf{X}_{i,j}}{\hat{\mathbf{X}}_{i,j}}\right) - \mathbf{X}_{i,j} + \hat{\mathbf{X}}_{i,j} \right)$$

η οποία είναι η μετρική που χρησιμοποιήσαμε και στην μέθοδο NMF.



Σχήμα 4.1.1: Αρχικό και ανακατασκευασμένο φασματογράφημα με την μέθοδο NAE.

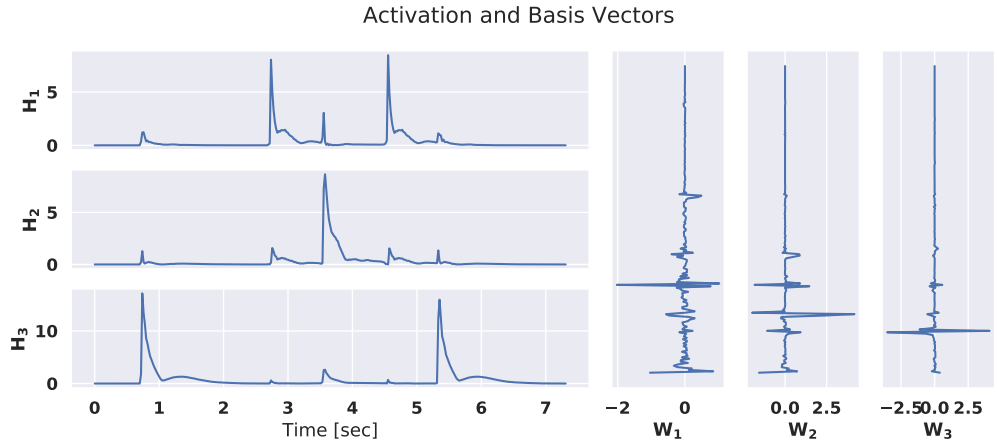
Στο Σχήμα 4.1.1 παρατηρούμε ότι το μοντέλο πετυχαίνει αρκετά καλή ανακατασκευή του αρχικού σήματος. Όμως, σε ορισμένα σημεία στο φασματογράφημα του ανακατασκευασμένου σήματος φαίνεται ότι έχουμε παρεμβολές αρμονικών από την μια νότα στην άλλη.



Σχήμα 4.1.2: Στα αριστερά οι γραμμές του πίνακα $\mathbf{H}$ και στα δεξιά οι K στήλες του πίνακα $\mathbf{W}$.

Παρατηρώντας το Σχήμα 4.1.2, όπου έχουμε τις K γραμμές του πίνακα $\mathbf{H}$ και τις K στήλες του πίνακα $\mathbf{W}$, βλέπουμε ότι πράγματι μπορούμε να ερμηνεύσουμε τους δυο πίνακες ως πίνακα ενεργοποιήσεων και πίνακα βάσεων αντίστοιχα. Ακόμη, επιβεβαιώνουμε τις παρεμβολές ανάμεσα στις συνιστώσες του ανακατασκευασμένου σήματος, από τις κορυφές των γραμμών του πίνακα ενεργοποιήσεων που συμπίπτουν χρονικά. Επίσης, διαπιστώνουμε ότι οι βάσεις περιέχουν και αρνητικές τιμές, οπότε υπάρχει δυνατότητα αλληλοακύρωσης μεταξύ τους. Ωστόσο, παρά την έλλειψη πλήρους αντιστοιχίας με την μέθοδο NMF, κάθε βάση φαίνεται ότι αντιστοιχεί σε μια νότα.

Ένας τρόπος για την μείωση των παρεμβολών είναι κάποιου είδους περιορισμού αραιότητας στον πίνακα ενεργοποιήσεων $\mathbf{H}$. Επαναλαμβάνουμε το παραπάνω πείραμα προσθέτοντας στο σφάλμα την νόρμα $\|\mathbf{H}\|_1$ σταθμισμένη με παράμετρο λ. Τα αποτελέσματα φαίνονται στο Σχήμα 4.1.3 όπου παρατηρούμε ότι πράγματι οι παρεμβολές μειώνονται.



Σχήμα 4.1.3: Αποτέλεσμα πειράματος με περιορισμό αραιότητας. Στα αριστερά οι γραμμές του πίνακα $\mathbf{H}$ και στα δεξιά οι K στήλες του πίνακα $\mathbf{W}$.

4.2 Πλεονεκτήματα

Η οικογένεια μεθόδων των NAE διαθέτει συγκεκριμένα πλεονεκτήματα σε σχέση με τα πλήρως επιβλεπόμενα νευρωνικά δίκτυα και την μέθοδο NMF (Venkataramani 2020). Αυτά όμως δεν είναι αποκλειστικά για τις μεθόδους NAE, δηλαδή μπορεί να υπάρχουν κι άλλες μέθοδοι που προσφέρουν παρόμοια ή τα ίδια πλεονεκτή-

ματα. Αρχικά, στα μοντέλα NAE, όπως και στην περίπτωση των μεθόδων NMF, στη φάση εξέτασης μπορούν προσαρμόσουν τις παραμέτρους τους επαναληπτικά για το τρέχον δείγμα. Ως αποτέλεσμα, μπορούμε να έχουμε καλύτερη προσαρμογή σε δείγματα τα οποία διαφέρουν σε κάποιο βαθμό από εκείνα του συνόλου εκπαίδευσης. Αντίθετα, η κατηγορία των νευρωνικών δικτύων που αναφέραμε συνήθως απαιτεί μεγάλο όγκο δεδομένων για την εκπαίδευση, καθώς δίνουν αποτέλεσμα μόνο με ένα πέρασμα και δεν έχουν την δυνατότητα προσαρμογής σε ένα συγκεκριμένο δείγμα. Ακόμη, τα μοντέλα NAE διαθέτουν την ικανότητα επανάχρησης, για παράδειγμα αφού εκπαιδευτούν για ανακατασκευή ομιλίας μπορούν να χρησιμοποιηθούν για την εξαγωγή ομιλίας από μείγματα.

Παρά τις διαισθητικές και δομικές τους ομοιότητες με τις μεθόδους NMF, οι NAE διαθέτουν ένα ισχυρό πλεονέκτημα σε σχέση με τις μεθόδους NMF παρόλο που έχουν αρκετές ομοιότητες. Χρησιμοποιώντας μεθόδους NAE μπορούμε να εκμεταλλευτούμε το μεγάλο εύρος αρχιτεκτονικών των νευρωνικών δικτύων, ώστε να επεκτείνουμε τις μεθόδους αυτές. Η επέκταση γίνεται με ευκολία καθώς η εκπαίδευση πάλι μπορεί να πραγματοποιηθεί με την μετάδοση σφάλματος προς τα πίσω και την βελτιστοποίηση καθοδικών κλίσεων. Αντίθετα, στην περίπτωση των μεθόδων NMF η επέκτασή τους είναι δυσκολότερη καθώς κάθε αλλαγή συνεπάγεται εξαγωγή νέων κανόνων ενημέρωσης των παραμέτρων.

4.3 Επεκτάσεις

Όπως αναφέραμε προηγουμένως τα μοντέλα NAE όντας νευρωνικά δίκτυα μπορούν να επεκταθούν με σχετική ευκολία. Επιπλέον, τα κριτήρια χαρακτηρισμού ενός autoencoder ως NAE, δηλαδή μη αρνητική είσοδος, έξοδος και ενδιάμεση αναπαράσταση, είναι αρκετά χαλαρά. Συγκεκριμένα, έχουν προταθεί επεκτάσεις ως προς το βάθος τους (Smaragdis and Venkataramani 2017) και ως προς το είδος επιπέδων που αποτελούν το μοντέλο NAE, όπως η χρήση συνελικτικών ή αναδρομικών επιπέδων (Venkataramani 2020).

Μια επέκταση που θα εξετάσουμε σε βάθος είναι η αύξηση του βάθους του μοντέλου. Αυξάνοντας το βάθος του μοντέλου ευελπιστούμε ότι θα μπορεί να αναπτύξει πιο σύνθετες αναπαραστάσεις ώστε να περιγράφει σήματα φωνής με καλύτερη ακρίβεια. Σκοπεύουμε επίσης να εξετάσουμε πως το αυξημένο βάθος επηρεάζει την ικανότητα διαχωρισμού του μοντέλου.

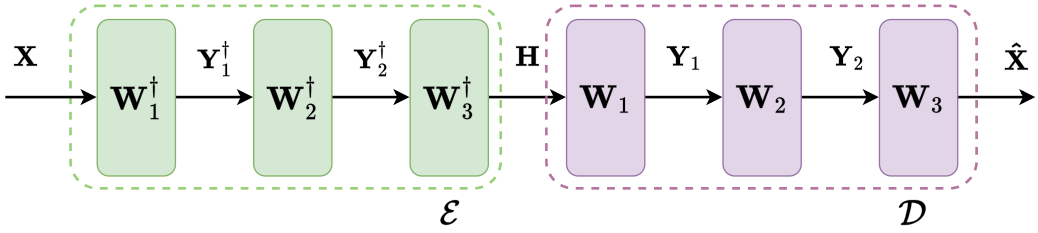
Ορίζουμε το πολυ-επίπεδο μοντέλο NAE με n επίπεδα ($n > 1$) ως εξής:

$$
\begin{aligned}
&\mathbf{Y}_0^\dagger = \mathbf{X} \\
&\mathbf{Y}_i^\dagger = g\left(\mathbf{W}_i^\dagger \cdot \mathbf{Y}_{i-1}^\dagger\right), \quad i = 1, 2, \ldots, n-1 \\
&\mathbf{H} = g\left(\mathbf{W}_n^\dagger \cdot \mathbf{Y}_{n-1}^\dagger\right) \\
&\mathbf{Y}_0 = \mathbf{H} \\
&\mathbf{Y}_i = g\left(\mathbf{W}_i \cdot \mathbf{Y}_{i-1}\right), \quad i = 1, 2, \ldots, n-1 \\
&\hat{\mathbf{X}} = g\left(\mathbf{W}_n \cdot \mathbf{Y}_{n-1}\right)
\end{aligned}
$$

όπου $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ συνάρτηση που δίνει έξοδο μη αρνητικές τιμές και εφαρμόζεται στοιχείο προς στοιχείο. Στα πειράματα χρησιμοποιούμε την συνάρτηση softplus $g(x) = \log\left(1 + e^x\right)$ (Σχήμα 2.3.4). Στη μορφή που το ορίσαμε δεν έχουμε πολώσεις, όμως ο ορισμός αυτός μπορεί να τροποποιηθεί με ευκολία ώστε να προστεθούν. Ακόμη, ισχύουν τα εξής:

$$
\begin{aligned}
&\mathbf{X} \in \mathbb{R}_{\geq 0}^{F \times T} \\
&\mathbf{Y}_i^\dagger \in \mathbb{R}_{\geq 0}^{F \times T} \\
&\mathbf{H} \in \mathbb{R}_{\geq 0}^{K \times T} \\
&\mathbf{Y}_i \in \mathbb{R}_{\geq 0}^{F \times T} \\
&\hat{\mathbf{X}} \in \mathbb{R}_{\geq 0}^{F \times T}
\end{aligned}
$$

Το μοντέλο που ορίσαμε δεν έχει n συνολικά επίπεδα αλλά το ορίζουμε ως μοντέλο NAE τάξης K με n επίπεδα καθώς έχουμε συμμετρικό κωδικοποιητή και αποκωδικοποιητή με n επίπεδα ο καθένας. Στο Σχήμα 4.3.1 έχουμε σχηματική αναπαράσταση για την περίπτωση $n = 3$.



Σχήμα 4.3.1: Σχηματική αναπαράσταση πολυ-επίπεδου μοντέλου NAE με $n = 3$. Με πράσινη διακεκομμένη γραμμή έχουμε τον κωδικοποιητή και με μωβ διακεκομμένη γραμμή έχουμε τον αποκωδικοποιητή.

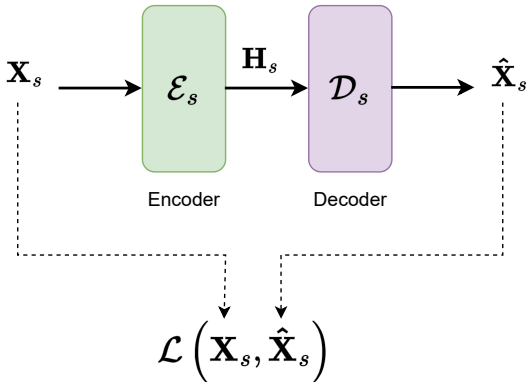
4.4 Εκπαίδευση σε Σήματα Ομιλίας

Η εκπαίδευση των μοντέλων NAE σε σήματα ομιλίας γίνεται με στόχο την ανακατασκευή της εισόδου στην έξοδο. Σκοπός είναι να μάθουμε έναν αποκωδικοποιητή τον οποίο θα επαναχρησιμοποιήσουμε στο πρόβλημα του διαχωρισμού. Επομένως, δεν είναι κύριος στόχος μας η υψηλή απόδοση ανακατασκευής αλλά η εκμάθηση αποκωδικοποιητή που μπορεί να εκφράσει σήματα ομιλίας αλλά ταυτόχρονα να μην είναι πολύ γενικός. Για να το πετύχουμε αυτό περιορίζουμε το μοντέλο NAE χρησιμοποιώντας μικρή τάξη Ks σε σχέση με την διάσταση της εισόδου, δηλαδή με ενδιάμεση αναπαράσταση που είναι μικρότερης διάστασης σε σχέση με την είσοδο ή την έξοδο του μοντέλου.

Η εκμάθηση των παραμέτρων του μοντέλου γίνεται με τροφοδότηση σφάλματος προς τα πίσω και έναν αλγόριθμο βελτιστοποίησης καθοδικής κλίσης όπως Adam (Kingma and Ba 2014) που χρησιμοποιούμε στα πειράματα. Η εκπαίδευση με Adam δεν δουλεύει με ολόκληρο το σύνολο δεδομένων σε κάθε βήμα αλλά κομμάτια αυτού που ονομάζονται batches. Το σφάλμα μπορούμε να το υπολογίσουμε είτε στο πεδίο της συχνότητας είτε στο πεδίο του χρόνου.

4.4.1 Σφάλμα στο Πεδίο της Συχνότητας

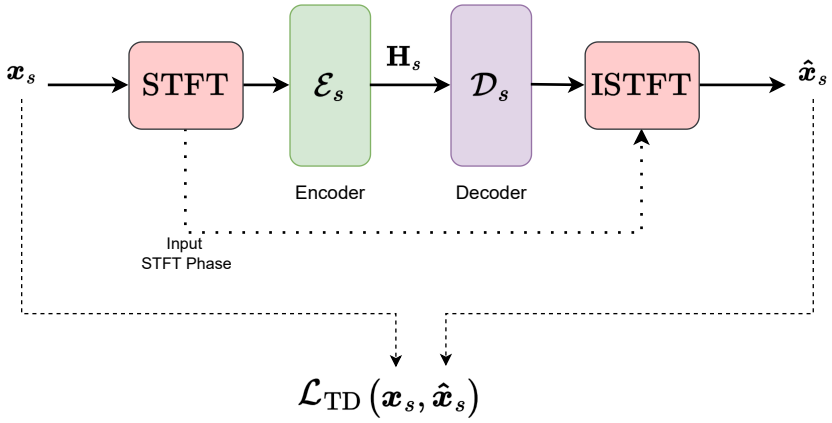
Στην περίπτωση αυτή, η είσοδος $\mathbf{X}_s$ και η έξοδος $\hat{\mathbf{X}}_s$ του μοντέλου είναι το μέτρο του STFT του σήματος εισόδου και το μέτρο του STFT της ανακατασκευής αντίστοιχα. Χρησιμοποιούμε έτσι την συνάρτηση σφάλματος $\mathcal{L}\left(\mathbf{X}_s, \hat{\mathbf{X}}_s\right)$ που ορίσαμε στην ενότητα 4.1. Στο Σχήμα 4.4.1 έχουμε την σχηματική αναπαράσταση των παραπάνω.



Σχήμα 4.4.1: Σχηματική αναπαράσταση εκπαίδευσης μοντέλου NAE στο πεδίο της συχνότητας.

4.4.2 Σφάλμα στο Πεδίο του Χρόνου

Στην περίπτωση αυτή, η είσοδος $\boldsymbol{x}_s$ και η έξοδος $\hat{\boldsymbol{x}}_s$ του μοντέλου βρίσκονται στο πεδίο του χρόνου. Χρησιμοποιούμε έτσι συνάρτηση σφάλματος $\mathcal{L}_{\text{TD}}\left(\boldsymbol{x}_s, \hat{\boldsymbol{x}}_s\right)$ στο πεδίο του χρόνου, η οποία στα πειράματα που πραγματοποιούμε είναι η $\mathcal{L}_{\text{TD}}\left(\boldsymbol{x}_s, \hat{\boldsymbol{x}}_s\right) = \|\boldsymbol{x}_s - \hat{\boldsymbol{x}}_s\|_1$. Πριν δοθεί το σήμα εισόδου στο μοντέλο NAE υπολογίζεται ο STFT του και δίνεται στον κωδικοποιητή του μοντέλου το μέτρο του STFT του. Αφού ο αποκωδικοποιητής δώσει την εκτίμηση του μέτρου του STFT του σήματος εξόδου, συνδυάζεται με την φάση του STFT της εισόδου και με τον αντίστροφο STFT παίρνουμε την έξοδο στο πεδίο του χρόνου. Στο Σχήμα 4.4.2 έχουμε την σχηματική αναπαράσταση των παραπάνω.



Σχήμα 4.4.2: Σχηματική αναπαράσταση εκπαίδευσης μοντέλου NAE στο πεδίο του χρόνου.

4.5 Μεθοδολογία Πλήρως Επιβλεπόμενου Διαχωρισμού με NAE

Πριν περιγράψουμε την μεθοδολογία για τον ημι-επιβλεπόμενο διαχωρισμό με NAE παρουσιάζουμε συνοπτικά την περίπτωση της πλήρους επίβλεψης, με σφάλμα στο πεδίο της συχνότητας όπως έχει προταθεί (Smaragdis and Venkataramani 2017).

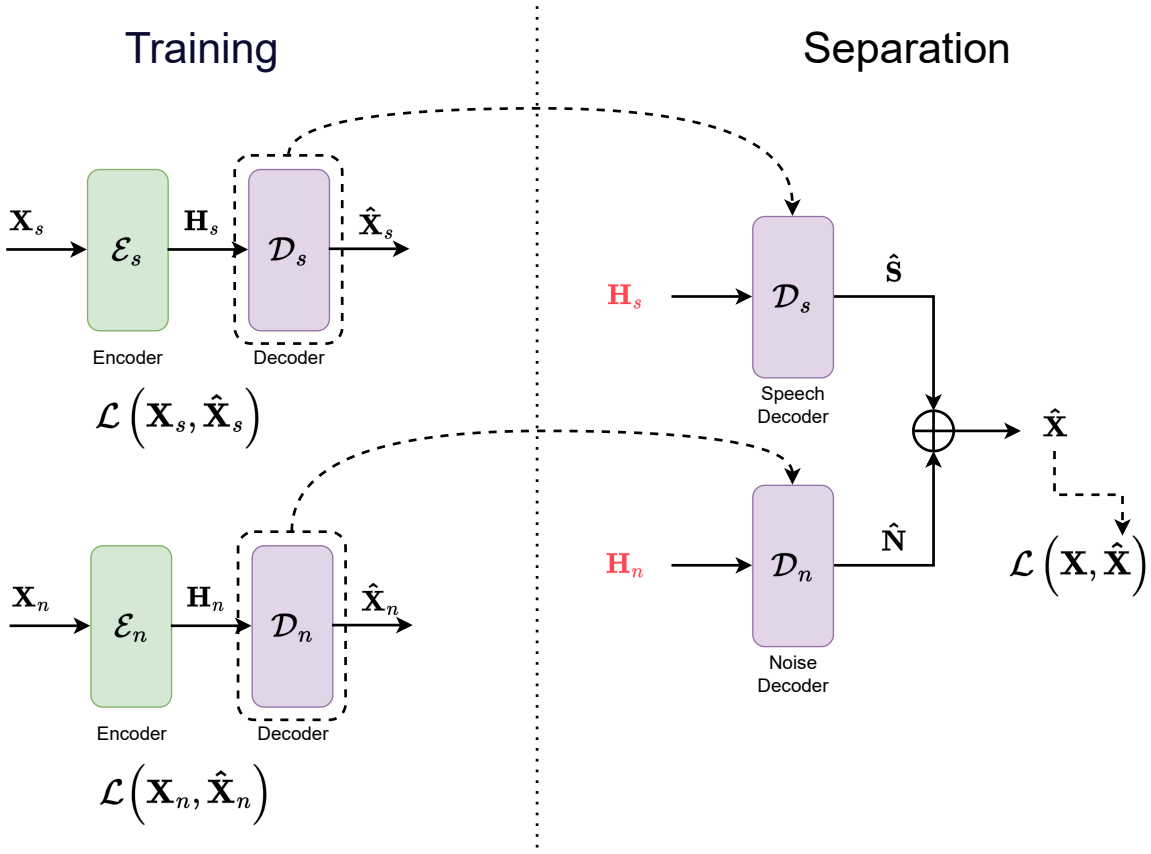
Έχοντας εκπαιδεύσει ένα μοντέλο NAE σε "καθαρά" σήματα φωνής και άλλο ένα σε σήματα θορύβου με τάξεις Ks και Kn αντίστοιχα, κρατάμε τους αποκωδικοποιητές τους $\mathcal{D}_s$ και $\mathcal{D}_n$ ώστε να τους χρησιμοποιήσουμε στον διαχωρισμό. Ο αποκωδικοποιητής $\mathcal{D}_s$ δέχεται ως είσοδο την ενδιάμεση αναπαράσταση $\mathbf{H}_s$ μεγέθους $Ks \times T$ και δίνει έξοδο την εκτίμηση του σήματος $\hat{\mathbf{S}}$, έτσι συμβολίζουμε $\hat{\mathbf{S}} = \mathcal{D}_s\left(\mathbf{H}_s\right)$. Αντίστοιχα, για τον αποκωδικοποιητή θορύβου θα είναι $\hat{\mathbf{N}} = \mathcal{D}_n\left(\mathbf{H}_n\right)$.

Έστω $\mathbf{X} \in \mathbb{R}_{\geq 0}^{F \times T}$ το μέτρο του STFT του θορυβώδους σήματος που επιθυμούμε να διαχωρίσουμε σε σήμα φωνής και σήμα θορύβου. Στόχος μας είναι η παρακάτω προσέγγιση.

$$\mathbf{X} \approx \hat{\mathbf{X}} = \hat{\mathbf{S}} + \hat{\mathbf{N}} = \mathcal{D}_s\left(\mathbf{H}_s\right) + \mathcal{D}_n\left(\mathbf{H}_n\right)$$

Θέλουμε να μάθουμε τις ενδιάμεσες αναπαραστάσεις $\mathbf{H}_s$ και $\mathbf{H}_n$ που δίνονται ως είσοδοι στους αποκωδικοποιητές. Χρησιμοποιούμε έτσι την συνάρτηση σφάλματος $\mathcal{L}\left(\mathbf{X}, \hat{\mathbf{X}}\right)$ που ορίσαμε στην ενότητα 4.1. Έχοντας δώσει ως είσοδο στους αποκωδικοποιητές τις ενδιάμεσες αναπαραστάσεις παίρνουμε την εκτίμηση του μείγματος $\hat{\mathbf{X}}$ που προκύπτει από την πρόσθεση των επιμέρους εκτιμήσεων. Αφού υπολογίσουμε την τιμή του σφάλματος ανάμεσα στο μείγμα και την εκτίμηση του μείγματος, μπορούμε με την τροφοδότηση σφάλματος προς τα πίσω να υπολογίσουμε τις παραγώγους του σφάλματος ως προς τις παραμέτρους που επιθυμούμε να μάθουμε. Επομένως, έχοντας υπολογίσει τις παραγώγους μπορούμε να χρησιμοποιήσουμε έναν αλγόριθμο βελτιστοποίησης καθοδικών κλίσεων ώστε να τις ενημερώσουμε. Η διαδικασία αυτή επαναλαμβάνεται για ένα καθορισμένο αριθμό βημάτων.

Στο Σχήμα 4.5.1 έχουμε την σχηματική αναπαράσταση όσων περιγράψαμε. Με κόκκινο χρώμα έχουμε τις παραμέτρους τις οποίες μαθαίνουμε επαναληπτικά κατά τον διαχωρισμό.



Σχήμα 4.5.1: Σχηματική αναπαράσταση του πλήρως επιβλεπόμενου διαχωρισμού με NAE και σφάλμα στο πεδίο της συχνότητας.

Αφού πραγματοποιήσουμε τον καθορισμένο αριθμό των επαναλήψεων για το μείγμα $\mathbf{X}$ και υπολογιστούν τα $\hat{\mathbf{S}}$ και $\hat{\mathbf{N}}$ υπολογίζουμε την εκτίμηση του σήματος φωνής και την εκτίμηση του θορύβου με την ακόλουθη μάσκα (Smaragdis and Venkataramani 2017).

$$\hat{\mathbf{S}}_{\text{mask}} = \frac{\hat{\mathbf{S}}}{\hat{\mathbf{S}} + \hat{\mathbf{N}}} \odot \mathbf{X} \qquad \hat{\mathbf{N}}_{\text{mask}} = \frac{\hat{\mathbf{N}}}{\hat{\mathbf{S}} + \hat{\mathbf{N}}} \odot \mathbf{X}$$

όπου $\hat{\mathbf{S}}_{\text{mask}}$ και $\hat{\mathbf{N}}_{\text{mask}}$ το μέτρο του STFT του σήματος φωνής και του σήματος θορύβου αντίστοιχα. Ο μετασχηματισμός πίσω στο πεδίο του χρόνου γίνεται με την χρήση της φάσης του μείγματος ακολουθώντας τα βήματα που περιγράψαμε στην ενότητα 2.1.2.

4.6 Μεθοδολογία Ημι-Επιβλεπόμενου Διαχωρισμού

Προτού εξηγήσουμε την μεθοδολογία που προτείνουμε για ημι-επιβλεπόμενο διαχωρισμό στην περίπτωση των μοντέλων NAE, θεωρούμε ότι αξίζει να περιγράψουμε την μεθοδολογία στην περίπτωση της NMF από την οποία αντλούμε έμπνευση και την οποία θα χρησιμοποιήσουμε στο πειραματικό μέρος.

4.6.1 Μεθοδολογία Ημι-Επιβλεπόμενου Διαχωρισμού με NMF

Στην παρούσα ενότητα περιγράφουμε την μεθοδολογία που ακολουθούμε για τον ημι-επιβλεπόμενο διαχωρισμό με NMF, την οποία την οποία περιγράψαμε σύντομα στην ενότητα 3.3. Θα την χρησιμοποιήσουμε στο πειραματικό μέρος σε θορυβώδη σήματα ομιλίας.

Έχοντας μάθει, από τα δεδομένα εκπαίδευσης "καθαρής" ομιλίας, τον πίνακα βάσεων $\mathbf{W}_s \in \mathbb{R}_{\geq 0}^{F \times Ks}$ που αποτελείται από Ks βάσεις, παίρνουμε τον πίνακα

$$\mathbf{W} = [\mathbf{W}_s, \mathbf{W}_n] \in \mathbb{R}_{\geq 0}^{F \times (Ks+Kn)}$$

όπου $\mathbf{W}_n \in \mathbb{R}_{\geq 0}^{F \times Kn}$ ο πίνακας βάσεων θορύβου που αποτελείται από Kn βάσεις. Ο $\mathbf{W}_n$ αρχικοποιείται τυχαία σε μη αρνητικές τιμές που προκύπτουν από την απόλυτη τιμή γκαουσιανού θορύβου όπως στο (Virtanen 2007).

Στην συνέχεια, έχοντας το θορυβώδες σήμα $\boldsymbol{x} = \boldsymbol{s}+\boldsymbol{n}$ παίρνουμε το μέτρο του STFT του το οποίο συμβολίζουμε ως $\mathbf{X} \in \mathbb{R}_{\geq 0}^{F \times T}$. Επιθυμούμε να λύσουμε το πρόβλημα NMF και να προσεγγίζουμε το $\mathbf{X}$ ως:

$$\mathbf{X} \approx \mathbf{WH} = [\mathbf{W}_s, \mathbf{W}_n]\,\mathbf{H}$$

όπου $\mathbf{H} = \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_n \end{bmatrix}$ με $\mathbf{H}_s \in \mathbb{R}_{\geq 0}^{Ks \times T}$ και $\mathbf{H}_n \in \mathbb{R}_{\geq 0}^{Kn \times T}$ δηλαδή $\mathbf{H} \in \mathbb{R}_{\geq 0}^{(Ks+Kn) \times T}$ οπότε έχουμε:

$$\mathbf{X} \approx [\mathbf{W}_s, \mathbf{W}_n] \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_n \end{bmatrix} \implies \mathbf{X} \approx \mathbf{W}_s\mathbf{H}_s + \mathbf{W}_n\mathbf{H}_n$$

Έτσι όπως και στην απλή περίπτωση της NMF, χρησιμοποιούμε την γενικευμένη απόκλιση Kullback Leibler.

$$D\left(\mathbf{X}|\mathbf{WH}\right) = \sum_{i,j} \left(\mathbf{X}_{i,j} \log \left(\frac{\mathbf{X}_{i,j}}{(\mathbf{WH})_{i,j}} \right) - \mathbf{X}_{i,j} + (\mathbf{WH})_{i,j} \right)$$

για την οποία οι κανόνες ενημέρωσης είναι ως εξής:

$$\mathbf{H}_{i,j} \leftarrow \mathbf{H}_{i,j} \frac{\sum_l \left(\frac{\mathbf{W}_{l,i}\mathbf{X}_{l,j}}{(\mathbf{WH})_{l,j}} \right)}{\sum_k \mathbf{W}_{k,i}} \qquad \mathbf{W}_{i,j} \leftarrow \mathbf{W}_{i,j} \frac{\sum_l \left(\frac{\mathbf{H}_{j,l}\mathbf{X}_{i,l}}{(\mathbf{WH})_{i,l}} \right)}{\sum_k \mathbf{H}_{j,k}}$$

Όμως, σε κάθε επανάληψη κρατάμε σταθερές τις πρώτες Ks βάσεις του πίνακα $\mathbf{W}$, δηλαδή τον πίνακα $\mathbf{W}_s$. Συνεπώς, σε κάθε επανάληψη ενημερώνονται μόνο τα $\mathbf{W}_n$, $\mathbf{H}_s$ και $\mathbf{H}_n$.

Αφού πραγματοποιήσουμε τον καθορισμένο αριθμό των επαναλήψεων και υπολογιστούν τα $\mathbf{W}_n$, $\mathbf{H}_s$ και $\mathbf{H}_n$ υπολογίζουμε την εκτίμηση του σήματος φωνής και την εκτίμηση του θορύβου με την ακόλουθη μάσκα, όπως και στο (Mohammadiha, Smaragdis, and Leijon 2013).

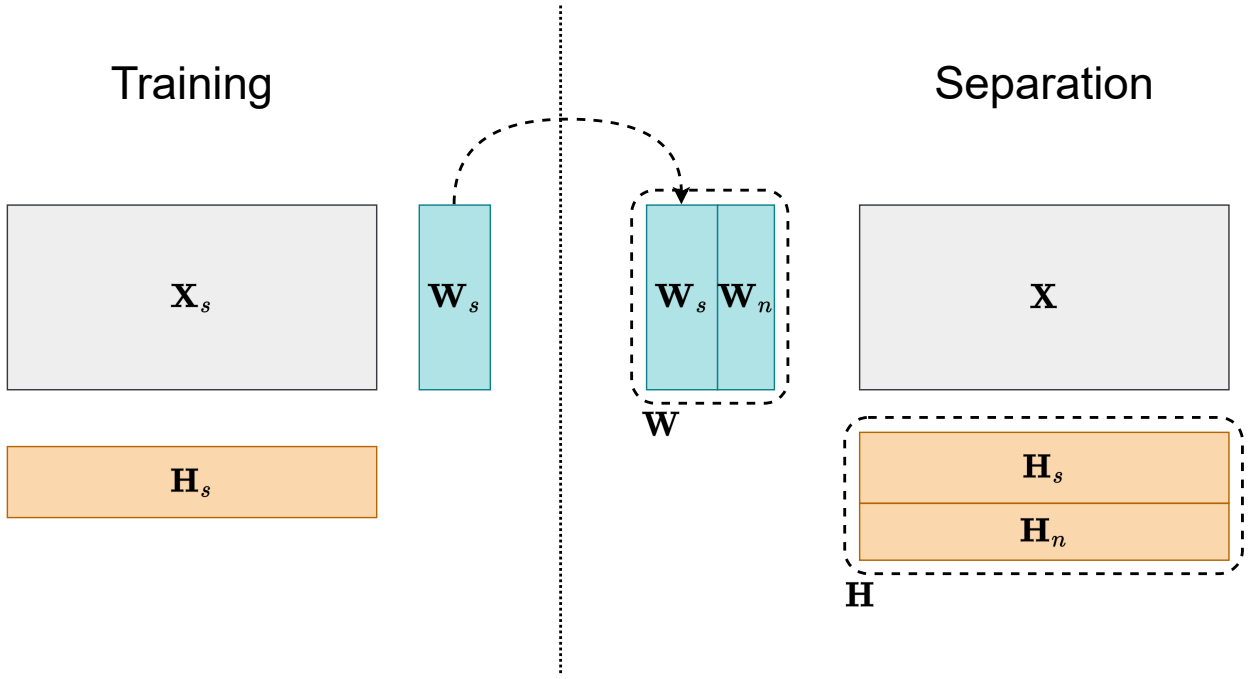
$$\hat{\mathbf{S}} = \frac{\mathbf{H}_s\mathbf{W}_s}{\mathbf{H}_s\mathbf{W}_s + \mathbf{H}_n\mathbf{W}_n} \odot \mathbf{X} \qquad \hat{\mathbf{N}} = \frac{\mathbf{H}_n\mathbf{W}_n}{\mathbf{H}_s\mathbf{W}_s + \mathbf{H}_n\mathbf{W}_n} \odot \mathbf{X}$$

όπου $\hat{\mathbf{S}}$ και $\hat{\mathbf{N}}$ εκτιμώμενο το μέτρο του STFT του σήματος φωνής και του σήματος θορύβου αντίστοιχα.

Στο Σχήμα 4.6.1 έχουμε την σχηματική αναπαράσταση της διαδικασίας που περιγράψουμε. Συγκεκριμένα, αφού μάθουμε τον πίνακα $\mathbf{W}_s$ από δεδομένα ομιλίας τον χρησιμοποιούμε στον πίνακα $\mathbf{W}$. Κατά τον διαχωρισμό μαθαίνουμε μόνο τους $\mathbf{W}_n$, $\mathbf{H}_s$ και $\mathbf{H}_n$ ενώ ο $\mathbf{W}_s$ παραμένει σταθερός.

Τροποποίηση Μεθοδολογίας Ημι-Επιβλεπόμενου Διαχωρισμού με NMF

Στο πειραματικό μέρος εξετάζουμε μια τροποποίηση της παραπάνω μεθοδολογίας, όπου υπολογίζουμε μια προς μια τις βάσεις θορύβου. Αρχικά, ακολουθούμε την παραπάνω διαδικασία για μια βάση θορύβου. Στη συνέχεια,



Σχήμα 4.6.1: Σχηματική αναπαράσταση του ημι-επιβλεπόμενου διαχωρισμού με NMF.

έχοντας μάθει την μια βάση θορύβου την συμπεριλαμβάνουμε στον πίνακα με τις σταθερές βάσεις και επαναλαμβάνουμε την διαδικασία που περιγράψαμε παραπάνω. Επαναλαμβάνουμε τα βήματα αυτά μέχρι να μάθουμε και την τελευταία βάση θορύβου. Αφού την μάθουμε με βάση τον πίνακα ενεργοποιήσεων $\mathbf{H}$ που μάθαμε πραγματοποιούμε τον τελικό διαχωρισμό. Συνεπώς, η διαδικασία αυτή έχει πολλαπλάσιο κόστος σε σχέση με την προηγούμενη αφού την εφαρμόζει Kn φορές.

4.6.2 Μεθοδολογία Ημι-Επιβλεπόμενου Διαχωρισμού με NAE

Με βάση την παραπάνω μεθοδολογία, προτείνουμε την αντίστοιχη μεθοδολογία για μοντέλα NAE.

Σφάλμα στο Πεδίο της Συχνότητας

Έχοντας εκπαιδεύσει ένα μοντέλο NAE σε "καθαρά" σήματα φωνής κρατάμε τον αποκωδικοποιητή του $\mathcal{D}_s$ ώστε να τον χρησιμοποιήσουμε στον διαχωρισμό. Συνδυάζουμε τον προ-εκπαιδευμένο αποκωδικοποιητή φωνής με ένα τυχαία αρχικοποιημένο αποκωδικοποιητή θορύβου $\mathcal{D}_n$. Όπως και πριν, ο αποκωδικοποιητής $\mathcal{D}_s$ δέχεται ως είσοδο την ενδιάμεση αναπαράσταση $\mathbf{H}_s$ μεγέθους $Ks \times T$ και δίνει έξοδο την εκτίμηση του σήματος $\hat{\mathbf{S}}$, έτσι συμβολίζουμε $\hat{\mathbf{S}} = \mathcal{D}_s(\mathbf{H}_s)$. Αντίστοιχα, για τον αποκωδικοποιητή θορύβου θα είναι $\hat{\mathbf{N}} = \mathcal{D}_n(\mathbf{H}_n)$, όπου ενδιάμεση αναπαράσταση $\mathbf{H}_s$ μεγέθους $Kn \times T$.

Έστω πάλι $\mathbf{X} \in \mathbb{R}_{\geq 0}^{F \times T}$ το μέτρο του STFT του θορυβώδους σήματος που επιθυμούμε να διαχωρίσουμε σε σήμα φωνής και σήμα θορύβου. Στόχος μας είναι η παρακάτω προσέγγιση.

$$\mathbf{X} \approx \hat{\mathbf{X}} = \hat{\mathbf{S}} + \hat{\mathbf{N}} = \mathcal{D}_s(\mathbf{H}_s) + \mathcal{D}_n(\mathbf{H}_n)$$

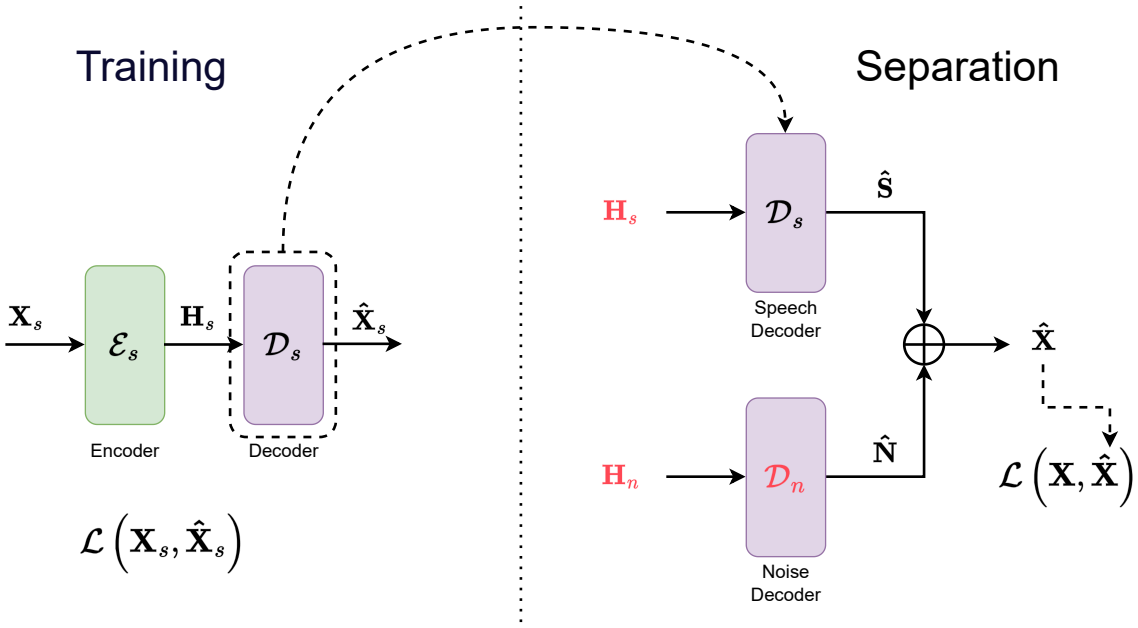
Λόγω της μη αρνητικότητας των δύο σημάτων που προστίθενται στόχος είναι να πάρουμε την έξοδο $\hat{\mathbf{X}}$ ως σύνθεση των δυο πηγών.

Επιθυμούμε να μάθουμε τις ενδιάμεσες αναπαραστάσεις $\mathbf{H}_s$ και $\mathbf{H}_n$ καθώς και τις παραμέτρους του αποκωδικοποιητή θορύβου $\mathcal{D}_n$. Χρησιμοποιούμε λοιπόν την συνάρτηση σφάλματος $\mathcal{L}\left(\mathbf{X}, \hat{\mathbf{X}}\right)$ που ορίσαμε στην ενότητα 4.1. Αφού δώσουμε ως είσοδο στους αποκωδικοποιητές τις ενδιάμεσες αναπαραστάσεις παίρνουμε την εκτίμηση του

μείγματος $\hat{\mathbf{X}}$ που προκύπτει από την πρόσθεση των επιμέρους εκτιμήσεων. Έπειτα, αφού υπολογίσουμε την τιμή του σφάλματος ανάμεσα στο μείγμα και την εκτίμηση του μείγματος, μπορούμε με την τροφοδότηση σφάλματος προς τα πίσω να υπολογίσουμε τις παραγώγους του σφάλματος ως προς τις παραμέτρους που επιθυμούμε να μάθουμε. Επομένως, αφού υπολογίσουμε τις παραγώγους χρησιμοποιούμε τον αλγόριθμο Adam ώστε να τις ενημερώσουμε.

Η διαδικασία προσέγγισης του μείγματος είναι επαναληπτική και πραγματοποιείται για ένα συγκεκριμένο αριθμό επαναλήψεων. Η αρχικοποίηση της ενδιάμεσης αναπαράστασης του θορύβου γίνεται με ομοιόμορφη κατανομή στο $[0, 1)$, ενώ η ενδιάμεση αναπαράσταση της ομιλίας αρχικοποιείται στην έξοδο του κωδικοποιητή της ομιλίας με είσοδο το μείγμα, δηλαδή $\mathbf{H}_s = \mathcal{E}_s(\mathbf{X})$. Πέραν της αρχικοποίησης δεν κάνουμε κάποια προσπάθεια ώστε να κρατήσουμε τις τιμές των ενδιάμεσων αναπαραστάσεων θετικές. Για κάθε μείγμα $\mathbf{X}$ που επιθυμούμε να διαχωρίσουμε η διαδικασία αυτή επαναλαμβάνεται από την αρχή και οι παράμετροι αρχικοποιούνται και πάλι.

Στο Σχήμα 4.6.2 έχουμε την σχηματική αναπαράσταση όσων περιγράψαμε. Με κόκκινο χρώμα έχουμε τις παραμέτρους τις οποίες μαθαίνουμε κατά τον διαχωρισμό. Κατά την εκπαίδευση μαθαίνουμε τις παραμέτρους του μοντέλου NAE ώστε να προσεγγίσουμε την είσοδο ομιλίας στο πεδίο της συχνότητας, ενώ κατά τον διαχωρισμό μαθαίνουμε τις παραμέτρους με το κόκκινο χρώμα ώστε να προσεγγίσουμε το μείγμα στο πεδίο της συχνότητας.



Σχήμα 4.6.2: Σχηματική αναπαράσταση του ημι-επιβλεπόμενου διαχωρισμού με NAE και σφάλμα στο πεδίο της συχνότητας.

Αφού πραγματοποιήσουμε τον καθορισμένο αριθμό των επαναλήψεων για το μείγμα $\mathbf{X}$ και υπολογιστούν τα $\hat{\mathbf{S}}$ και $\hat{\mathbf{N}}$ υπολογίζουμε την εκτίμηση του σήματος φωνής και την εκτίμηση του θορύβου με την ακόλουθη μάσκα, όπως και στην πλήρως επιβλεπόμενη περίπτωση (Smaragdis and Venkataramani 2017).

$$\hat{\mathbf{S}}_{\text{mask}} = \frac{\hat{\mathbf{S}}}{\hat{\mathbf{S}} + \hat{\mathbf{N}}} \odot \mathbf{X} \qquad \hat{\mathbf{N}}_{\text{mask}} = \frac{\hat{\mathbf{N}}}{\hat{\mathbf{S}} + \hat{\mathbf{N}}} \odot \mathbf{X}$$

όπου $\hat{\mathbf{S}}_{\text{mask}}$ και $\hat{\mathbf{N}}_{\text{mask}}$ το μέτρο του STFT του σήματος φωνής και του σήματος θορύβου αντίστοιχα. Ο μετασχηματισμός πίσω στο πεδίο του χρόνου γίνεται με την χρήση της φάσης του μείγματος ακολουθώντας τα βήματα που περιγράψαμε στην ενότητα 2.1.2.

Σφάλμα στο Πεδίο του Χρόνου

Σε αντίθεση με την παραπάνω μεθοδολογία μπορούμε να πάρουμε το σφάλμα στο πεδίο του χρόνου. Έχοντας εκπαιδεύσει ένα μοντέλο NAE σε "καθαρά" σήματα φωνής με σφάλμα στο πεδίο του χρόνου, κρατάμε πάλι τον αποκωδικοποιητή του $\mathcal{D}_s$ ώστε να τον χρησιμοποιήσουμε στον διαχωρισμό.

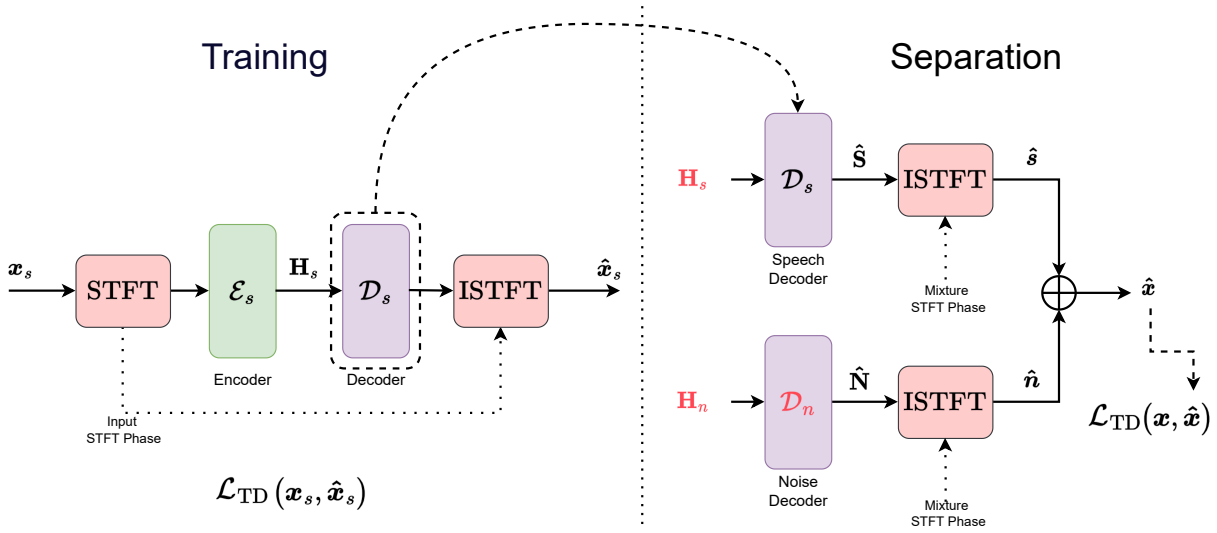
Όπως πριν, συνδυάζουμε τον προ-εκπαιδευμένο αποκωδικοποιητή φωνής με ένα τυχαία αρχικοποιημένο αποκωδικοποιητή θορύβου $\mathcal{D}_n$. Όμως οι δυο αποκωδικοποιητές που δίνουν τις εκτιμήσεις του μέτρου του STFT των σημάτων $\hat{\mathbf{S}}$ και $\hat{\mathbf{N}}$ αντίστοιχα, ακολουθούνται από τον αντίστροφο STFT που συνδυάζει το μέτρο STFT της εκτίμησης και την φάση STFT του μείγματος και δίνει ως έξοδο την εκτίμηση του σήματος στο πεδίο του χρόνου με βάση τα τελευταία βήματα που περιγράψαμε στην ενότητα 2.1.2. Έτσι, προκύπτουν οι εκτιμήσεις στο πεδίο του χρόνου $\hat{\boldsymbol{s}}$ και $\hat{\boldsymbol{n}}$ για το σήμα φωνής και το σήμα θορύβου αντίστοιχα.

Έστω $\boldsymbol{x}$ η κυματομορφή του μείγματος και $\mathbf{X} \in \mathbb{R}_{\geq 0}^{F \times T}$ το μέτρο του STFT του. Στόχος μας αυτή την φορά είναι η προσέγγιση του $\boldsymbol{x}$ στο πεδίο του χρόνου ως εξής:

$$\boldsymbol{x} \approx \hat{\boldsymbol{x}} = \hat{\boldsymbol{s}} + \hat{\boldsymbol{n}}$$

Αντίστοιχα με πριν, επιθυμούμε να μάθουμε τις ενδιάμεσες αναπαραστάσεις $\mathbf{H}_s$ και $\mathbf{H}_n$ καθώς και τις παραμέτρους του αποκωδικοποιητή θορύβου $\mathcal{D}_n$. Χρησιμοποιούμε λοιπόν την συνάρτηση σφάλματος $\mathcal{L}_{\text{TD}}(\boldsymbol{x}, \hat{\boldsymbol{x}})$ στο πεδίο του χρόνου. Στα πειράματα που πραγματοποιούμε στην συνέχεια χρησιμοποιούμε την νόρμα ℓ_1, δηλαδή $\mathcal{L}_{\text{TD}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1$.

Αφού υπολογίσουμε την τιμή του σφάλματος, μπορούμε με την τροφοδότηση σφάλματος προς τα πίσω να υπολογίσουμε τις παραγώγους του σφάλματος ως προς τις παραμέτρους που επιθυμούμε να μάθουμε. Επομένως, αφού υπολογίσουμε της παραγώγους πάλι χρησιμοποιούμε τον αλγόριθμο Adam ώστε να τις ενημερώσουμε.



Σχήμα 4.6.3: Σχηματική αναπαράσταση του ημι-επιβλεπόμενου διαχωρισμού με NAE στο πεδίο του χρόνου. Με κόκκινο χρώμα έχουμε τις παραμέτρους οι οποίες μαθαίνονται κατά τον διαχωρισμό.

Τέλος, έχοντας πραγματοποιήσει τον καθορισμένο αριθμό των επαναλήψεων, με βάση τα $\hat{\mathbf{S}}$ και $\hat{\mathbf{N}}$ υπολογίζουμε την εκτίμηση του σήματος φωνής και την εκτίμηση του θορύβου με την ακόλουθη μάσκα στο πεδίο της συχνότητας.

$$\hat{\mathbf{S}}_{\text{mask}} = \frac{\hat{\mathbf{S}}}{\hat{\mathbf{S}} + \hat{\mathbf{N}}} \odot \mathbf{X} \qquad \hat{\mathbf{N}}_{\text{mask}} = \frac{\hat{\mathbf{N}}}{\hat{\mathbf{S}} + \hat{\mathbf{N}}} \odot \mathbf{X}$$

όπου $\hat{\mathbf{S}}_{\text{mask}}$ και $\hat{\mathbf{N}}_{\text{mask}}$ το μέτρο του STFT του σήματος φωνής και του σήματος θορύβου αντίστοιχα. Ο

μετασχηματισμός πίσω στο πεδίο του χρόνου γίνεται με την χρήση της φάσης του μείγματος ακολουθώντας τα βήματα που περιγράψαμε στην ενότητα 2.1.2.

Στο Σχήμα 4.6.3 έχουμε την σχηματική αναπαράσταση της διαδικασίας διαχωρισμού με σφάλμα στο πεδίο του χρόνου, την οποία περιγράψαμε. Με κόκκινο χρώμα έχουμε τις παραμέτρους τις οποίες μαθαίνουμε κατά τον διαχωρισμό. Για τον μετασχηματισμό πίσω στο πεδίο του χρόνου χρησιμοποιούμε την φάση του STFT της αρχικής εισόδου στην περίπτωση της εκπαίδευσης και την φάση του STFT του μείγματος στην περίπτωση του διαχωρισμού.

Κεφάλαιο 5

Πειράματα και Αποτελέσματα

5.1 Πειραματικό Πλαίσιο

Στο πειραματικό μέρος της εργασίας, εκπαιδεύουμε πρώτα μοντέλα NMF και NAE σε "καθαρά" σήματα ομιλίας και στη συνέχεια τα αξιολογούμε στον διαχωρισμό θορυβωδών σημάτων ομιλίας με βάση τις ημι-επιβλεπόμενες μεθοδολογίες που περιγράψαμε στο Κεφάλαιο 4.

5.1.1 Σύνολα Δεδομένων

Για το πειραματικό μέρος της εργασίας κατασκευάζουμε μονοφωνικά μείγματα ομιλίας και διαφόρων ειδών θορύβου. Βασιζόμαστε στο σύνολο δεδομένων TIMIT για καθαρά σήματα ομιλίας και στα σύνολα DEMAND και MUSDB18 για σήματα θορύβου, τα οποία παρουσιάσαμε αναλυτικά στην ενότητα 3.5.

Σύνολο Σημάτων Ομιλίας

Η εκπαίδευση των μοντέλων που θα χρησιμοποιηθούν στον διαχωρισμό γίνεται στο σύνολο TIMIT. Συγκεκριμένα, χωρίζουμε το σύνολο train αυτού σε σύνολο εκπαίδευσης και σε σύνολο επαλήθευσης (για την εκπαίδευση) μεγέθους μερικών δεκάδων δειγμάτων.

Σύνολα Θορυβωδών Σημάτων

Για τα πειράματα του διαχωρισμού, κατασκευάζουμε δύο σύνολα δεδομένων, ένα από τον συνδυασμό TIMIT και DEMAND και ακόμη ένα από τον συνδυασμό TIMIT και MUSDB. Καθένα από τα δύο σύνολα χωρίζεται σε σύνολο επαλήθευσης (`Validation`) και σύνολο εξέτασης (`Test`). Χρησιμοποιούμε το σύνολο επαλήθευσης ώστε να προσαρμόσουμε και να ρυθμίσουμε τα μοντέλα με τα οποία πειραματιζόμαστε. Ακόμη, το χρησιμοποιούμε για να επιλέξουμε το καλύτερο μοντέλο ανά κατηγορία το οποίο δοκιμάζουμε στο σύνολο εξέτασης.

Πρώτα, χωρίζουμε το προϋπάρχον σύνολο test του TIMIT σε σύνολα επαλήθευσης και εξέτασης, ώστε στο κάθε σύνολο να έχουμε διαφορετικούς ομιλητές αλλά και διαφορετικές προτάσεις, τόσο μεταξύ των δύο συνόλων αλλά και με το σύνολο εκπαίδευσης. Έπειτα, για το σύνολο DEMAND πραγματοποιούμε τον χωρισμό σε σύνολα επαλήθευσης και εξέτασης, χωρίζοντας τα σήματα διάρκειας 5 λεπτών στη μέση. Τέλος, για το σύνολο MUSDB18 ως σύνολο επαλήθευσης παίρνουμε το προϋπάρχον σύνολο train και ως σύνολο εξέτασης παίρνουμε το προϋπάρχον σύνολο test.

Συνεπώς, κατασκευάζουμε το σύνολο δεδομένων TIMIT-DEMAND με μέγεθος 256 μείγματα στο σύνολο επαλήθευσης και 256 μείγματα στο σύνολο εξέτασης με SNR στο εύρος $[-5, 5]$ dB. Η επιλογή των δειγμάτων για την κατασκευή του μείγματος από το αντίστοιχο σύνολο γίνεται τυχαία όπως και η επιλογή του SNR του κάθε μείγματος. Ομοίως, κατασκευάζουμε το σύνολο δεδομένων TIMIT-MUSDB, αφού αφαιρέσουμε πρώτα τα φωνητικά από τα μουσικά κομμάτια του MUSDB18. Αφού δημιουργήσουμε τα σύνολα αυτά τα διατηρούμε σταθερά για όλα τα πειράματα που θα πραγματοποιήσουμε.

5.1.2 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων αποτελείται από τα βήματα που ακολουθούν. Σε περίπτωση που χρειάζεται μειώνουμε τον ρυθμό δειγματοληψίας στα 16 kHz. Περικόπτουμε το σήμα ώστε να έχει διάρκεια 3.5 δευτερόλεπτα. Στη συνέχεια κανονικοποιούμε το σήμα ώστε να έχει μηδενική μέση τιμή και μοναδιαία διακύμανση. Όπου χρησιμοποιούμε STFT χρησιμοποιούμε παράθυρο μήκους 64 ms (1024 δείγματα) με 75% επικάλυψη, ενώ το παράθυρο είναι τύπου root-Hann που περιγράψαμε στην ενότητα 2.1.2.

Κατασκευή Μειγμάτων

Η κατασκευή των μειγμάτων γίνεται προσθέτοντας τα δυο κανονικοποιημένα σήματα, αφού πρώτα το σήμα θορύβου έχει πολλαπλασιαστεί με παράγοντα $10^{-\mathrm{SNR}/20}$ όπου SNR το επιθυμητό Signal-to-Noise Ratio του μείγματος. Τέλος, το σήμα που προκύπτει κανονικοποιείται ώστε να έχει μηδενική μέση τιμή και μοναδιαία διακύμανση.

5.2 Εκπαίδευση NMF σε Σήματα Ομιλίας

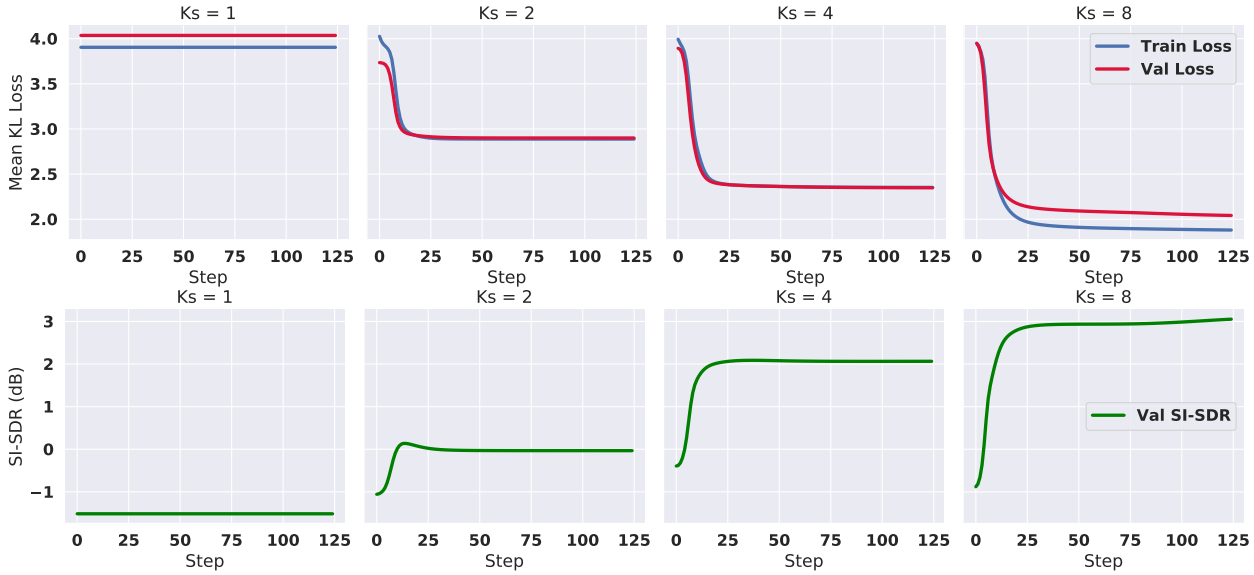
Στην ενότητα αυτή εκπαιδεύουμε πίνακες βάσεων διαφόρων μεγεθών στο σύνολο δεδομένων TIMIT. Η προεπεξεργασία των δεδομένων αποτελείται από τα βήματα που περιγράψαμε στην υποενότητα 5.1.2. Έτσι, για κάθε δείγμα του συνόλου εκπαίδευσης έχουμε πίνακα μεγέθους $F \times T$. Στην συνέχεια, συνενώνουμε τα n στο πλήθος δείγματα σε ένα πίνακα $\mathbf{X}_s$ μεγέθους $F \times nT$. Για την εκμάθηση κάθε πίνακα βάσεων μεγέθους $F \times Ks$ χρησιμοποιούμε την μέθοδο NMF όπως την περιγράψαμε στην ενότητα 2.4, για την γενικευμένη απόκλιση Kullback Leibler πραγματοποιώντας 125 βήματα. Σε κάθε βήμα υπολογίζουμε το σφάλμα του μοντέλου και στο σύνολο επαλήθευσης.

Μετά την ολοκλήρωση της εκπαίδευσης των μοντέλων NMF με διαφορετικό αριθμό βάσεων Ks, στον Πίνακα 5.1 έχουμε την απόδοση στο σύνολο επαλήθευσης. Παρατηρούμε ότι όσο μεγαλώνει ο αριθμός των βάσεων Ks η απόδοση στο σύνολο επαλήθευσης αυξάνεται.

	Mean SI-SDR (dB)									
Ks	1	2	4	8	10	16	20	32	64	128
	-1.5	0.0	2.1	3.1	4.2	5.6	6.9	10.0	14.4	**19.1**

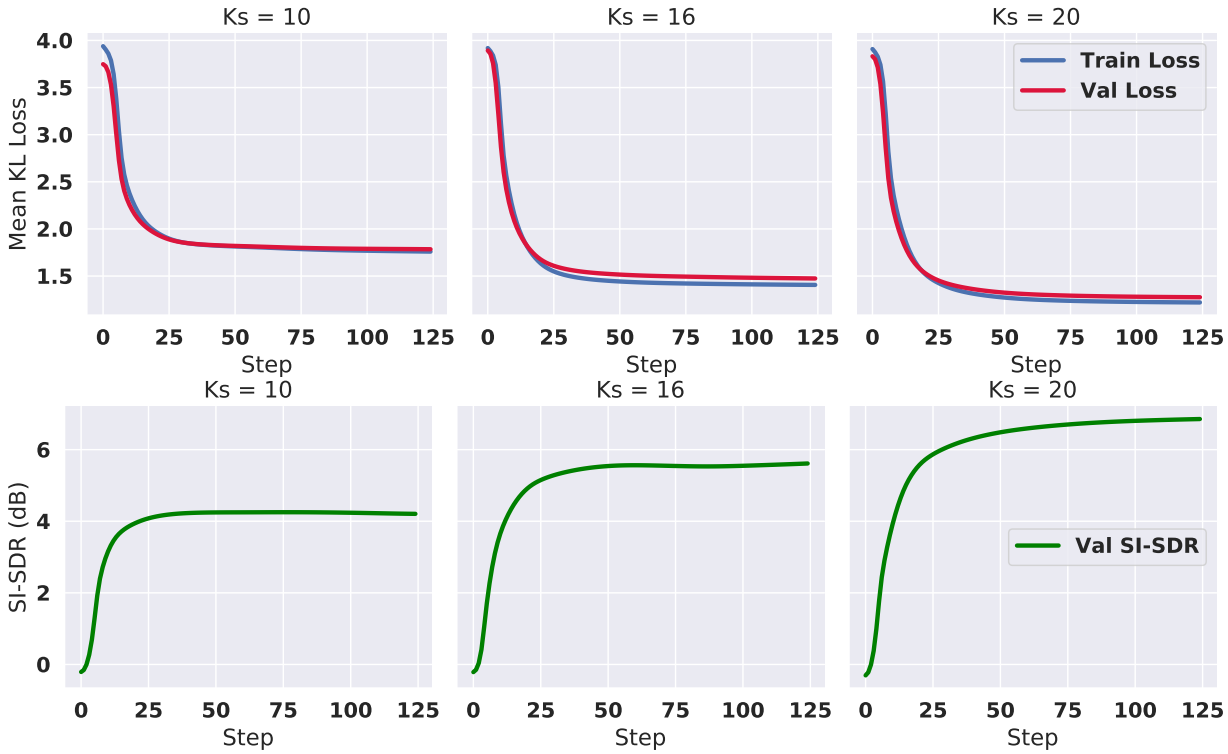
Πίνακας 5.1: Εκπαίδευση NMF σε σήματα ομιλίας του συνόλου δεδομένων TIMIT. Μέση απόδοση στο σύνολο επαλήθευσης σε SI-SDR.

Στα Σχήματα 5.2.1, 5.2.2 και 5.2.3 στην πάνω σειρά έχουμε τις καμπύλες του σφάλματος στο σύνολο εκπαίδευσης και επαλήθευσης ως προς το βήμα εκπαίδευσης, για τα διάφορα μεγέθη πινάκων βάσεων. Στην κάτω σειρά έχουμε τις καμπύλες απόδοσης σε SI-SDR στο σύνολο επαλήθευσης ως προς το βήμα εκπαίδευσης, για τα διάφορα μεγέθη πινάκων βάσεων.

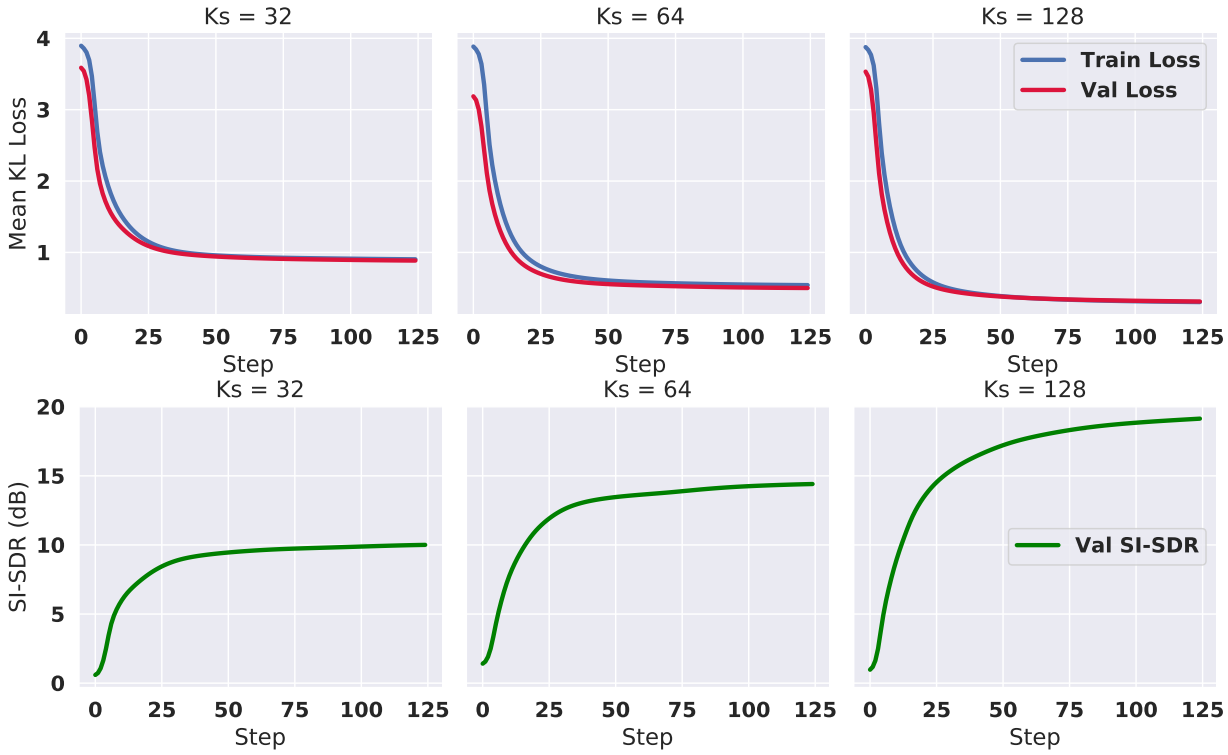


Σχήμα 5.2.1: Στο πάνω σχήμα έχουμε τις καμπύλες σφάλματος της γενικευμένης Kullback Leibler απόκλισης ως προς το βήμα εκπαίδευσης, στο σύνολο εκπαίδευσης (μπλε) και το σύνολο επαλήθευσης (κόκκινο). Στο κάτω σχήμα έχουμε τις καμπύλες της μετρικής SI-SDR σε dB ως προς το βήμα εκπαίδευσης στο σύνολο επαλήθευσης (πράσινο). Για Ks ίσο με 1, 2, 4 και 8.

Στο Σχήμα 5.2.1 παρατηρούμε ότι για $Ks = 1$ από το πρώτο βήμα βρισκόμαστε σε τοπικό ελάχιστο. Για $Ks = 2$ η απόδοση στο σύνολο επαλήθευσης σε SI-SDR εμφανίζει την μέγιστη τιμή στα αρχικά βήματα, το οποίο όμως δεν ισχύει και για το σφάλμα με την γενικευμένη Kullback Leibler απόκλιση, που έχει ελάχιστη



Σχήμα 5.2.2: Όπως στο Σχήμα 5.2.1 έχουμε τις καμπύλες σφάλματος και απόδοσης για Ks ίσο με 10, 16, 20.



Σχήμα 5.2.3: Όπως στο Σχήμα 5.2.1 έχουμε τις καμπύλες σφάλματος και απόδοσης για Ks ίσο με 32, 64, 128.

τιμή στα τελευταία βήματα. Γενικά, το σφάλμα εκπαίδευσης είναι χαμηλότερο από το σφάλμα επαλήθευσης και η διαφορά τους μικρή.

Στα Σχήματα 5.2.2 και 5.2.3 παρατηρούμε ότι όλες οι καμπύλες έχουν ομαλή μορφή με το σφάλμα και την απόδοση σχεδόν να σταθεροποιούνται και να συγκλίνουν όσο περνάνε τα βήματα. Πάλι, το σφάλμα εκπαίδευσης και επαλήθευσης είναι αρκετά κοντά. Βλέπουμε ότι όσο μεγαλώνει το Ks τόσο βελτιώνεται η απόδοση αλλά η σύγκλιση σε ορισμένες περιπτώσεις απαιτεί περισσότερα βήματα. Ακόμη, διαπιστώνουμε ότι το σφάλμα εκπαίδευσης δεν αυξάνεται από βήμα σε βήμα, όπως εγγυάται θεωρητικά ο αλγόριθμος που χρησιμοποιούμε.

5.3 Πειράματα Διαχωρισμού με NMF

Πρώτα, δοκιμάζουμε την ημι-επιβλεπόμενη μέθοδο NMF που περιγράψαμε στην ενότητα 4.6.1 στο σύνολο δεδομένων TIMIT-DEMAND για διαφορετικούς συνδυασμούς αριθμού βάσεων ομιλίας Ks και αριθμού βάσεων θορύβου Kn. Στον Πίνακα 5.2 έχουμε τα αποτελέσματα στο σύνολο επαλήθευσης και στο σύνολο εξέτασης σε SI-SDR.

			Mean SI-SDR (dB)									
	Kn	Ks	1	2	4	8	10	16	20	32	64	128
Validation	$Kn = 1$		7.9	9.9	11.1	**12.1**	**12.0**	**12.4**	**12.3**	**12.2**	10.5	6.8
	$Kn = 10$					-1.7	-1.4	1.2	2.1	4.6	7.0	7.5
	$Kn = 16$					-3.1		-0.3		3.2	5.9	7.1
	$Kn = Ks\ /\ 2$					2.6	1.4	2.0	2.0	3.2	4.8	6.2
	$Kn = Ks\ /\ 8$					**12.1**		10.3		8.8	7.5	7.1
Test	$Kn = 1$		8.1	9.9	11.2	**12.5**	**12.5**	**12.9**	**12.7**	**12.6**	11.0	7.1
	$Kn = 10$					-2.4	-1.9	0.6	1.6	4.1	6.9	7.8
	$Kn = 16$					-3.8		-0.9		2.7	5.8	7.4
	$Kn = Ks\ /\ 2$					1.7	0.8	1.5	1.7	2.7	4.5	6.4
	$Kn = Ks\ /\ 8$					**12.5**		9.9		8.7	7.5	7.4

Πίνακας 5.2: Απόδοση της μεθόδου NMF στο TIMIT-DEMAND

Έπειτα, πραγματοποιούμε πειράματα για τους ίδιους συνδυασμούς αριθμού βάσεων ομιλίας Ks και αριθμού βάσεων θορύβου Kn, στο σύνολο δεδομένων TIMIT-MUSDB. Στον Πίνακα 5.3 έχουμε τα αποτελέσματα στο σύνολο επαλήθευσης και στο σύνολο εξέτασης σε SI-SDR.

			Mean SI-SDR (dB)									
	Kn	Ks	1	2	4	8	10	16	20	32	64	128
Validation	$Kn = 1$		6.3	7.4	**7.8**	**8.1**	**7.9**	7.5	7.1	6.5	3.7	2.6
	$Kn = 10$					-0.6	-0.6	2.2	3.0	5.0	5.2	4.0
	$Kn = 16$					-2.2	-2.2	0.4	1.2	3.7	4.7	4.0
	$Kn = Ks\ /\ 2$					3.9	2.6	3.0	2.8	3.7	3.9	3.9
	$Kn = Ks\ /\ 8$					**8.1**		**7.7**		7.1	5.3	4.0
Test	$Kn = 1$		6.6	7.5	**7.9**	**8.3**	**8.1**	7.7	7.5	6.8	3.8	2.8
	$Kn = 10$					-0.8	-0.5	2.2	3.0	5.1	5.8	4.4
	$Kn = 16$					-2.5	-2.2	0.4	1.3	3.7	5.3	4.5
	$Kn = Ks\ /\ 2$					3.8	2.7	2.9	3.0	3.7	4.4	4.4
	$Kn = Ks\ /\ 8$					**8.3**		**7.9**		7.5	5.8	4.5

Πίνακας 5.3: Απόδοση της μεθόδου NMF στο TIMIT-MUSDB

Στο TIMIT-DEMAND καλύτερη απόδοση παίρνουμε για μία βάση θορύβου και 8, 10, 16, 20 ή 32 βάσεις

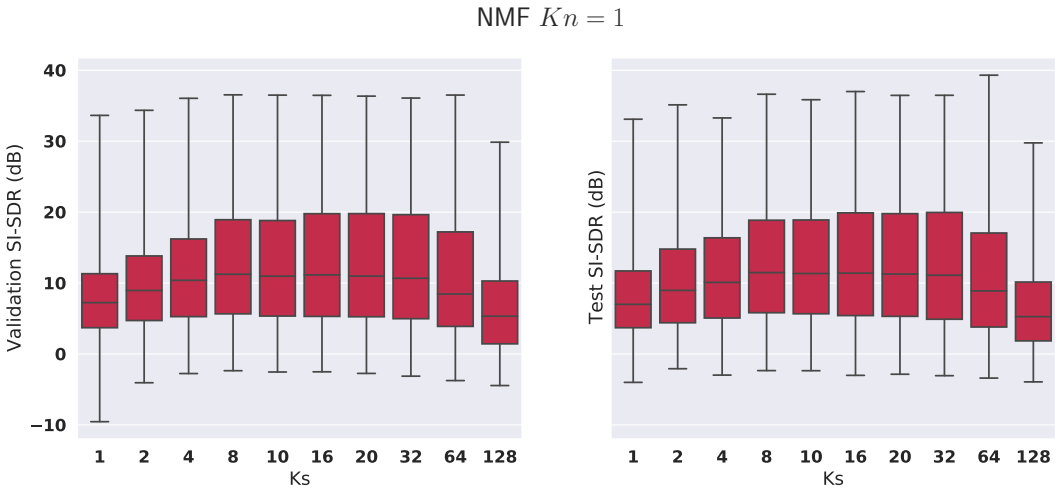
ομιλίας. Στο TIMIT-MUSDB καλύτερη απόδοση παίρνουμε για μία βάση θορύβου και 4, 8, ή 10 βάσεις ομιλίας ή για 2 βάσεις θορύβου και 16 βάσεις ομιλίας. Γενικά, παρατηρούμε ότι όσο μειώνονται οι βάσεις θορύβου τόσο καλύτερη απόδοση έχουμε. Ακόμη, αρκετά καλά είναι τα αποτελέσματα για πολύ λίγες βάσεις ομιλίας με μια βάση θορύβου, το οποίο οφείλεται εν μέρει στο γεγονός ότι στην ανακατασκευή με μάσκα η ποιότητα του εκτιμώμενου σήματος δεν εξαρτάται τόσο έντονα από την εκφραστικότητα των βάσεων ομιλίας. Συγκρίνοντας την απόδοση μεταξύ TIMIT-DEMAND και TIMIT-MUSDB μπορούμε να συμπεράνουμε ότι το δεύτερο περιέχει πιο δύσκολους θορύβους καθώς η απόδοση της μεθόδου είναι αρκετά χαμηλότερη.

Στη συνέχεια, στον Πίνακα 5.4 συγκρίνουμε στο TIMIT-DEMAND την μέθοδο NMF που χρησιμοποιήσαμε παραπάνω και την τροποποιημένη που περιγράψαμε στην υποενότητα 4.6.1 με στόχο την βελτίωση της απόδοσης για μεγάλα Kn. Συγκρίνουμε για διαφορετικά Ks με Kn ίσο με 10 ή το μισό του Ks.

			Mean SI-SDR (dB)						
	$Kn = 10$	Ks	8	10	16	20	32	64	128
Val	Baseline		-1.7	-1.4	1.2	2.1	4.6	7.0	7.5
	Iterative		**-0.3**	**-0.3**	**2.1**	**3.0**	**5.2**	**7.2**	**8.0**
Test	Baseline		-2.4	-1.9	0.6	1.6	4.1	6.9	7.8
	Iterative		**-1.1**	**-0.8**	**1.6**	**2.6**	**4.8**	**7.2**	**8.4**
	$Kn = Ks$ / 2	Ks	8	10	16	20	32	64	128
Val	Baseline		**2.6**	**1.4**	2.0	2.0	3.2	4.8	6.2
	Iterative		2.0	1.2	**2.6**	**3.1**	**4.3**	**6.0**	**7.1**
Test	Baseline		**1.7**	**0.8**	1.5	1.7	2.7	4.5	6.4
	Iterative		1.5	0.7	**2.0**	**2.7**	**4.0**	**5.8**	**7.4**

Πίνακας 5.4: Συγκριση αλγορίθμων NMF στο TIMIT-DEMAND

Παρατηρούμε ότι, για $Kn = 10$ πράγματι με την χρήση της τροποποιημένης μεθόδου βελτιώνεται σε ένα βαθμό απόδοση αλλά παραμένουμε μακριά από το μέγιστο του Πίνακα 5.2. Για $Kn = Ks/2$ βλέπουμε ότι όταν το Kn είναι μικρό, δηλαδή ίσο με 4 ή 5, η τροποποιημένη μέθοδος μειώνει την απόδοση, σε αντίθεση με τις υπόλοιπες περιπτώσει όπου την βελτιώνει.



Σχήμα 5.3.1: Θηκόγράμματα για την απόδοση στο σύνολο επαλήθευσης και εξέτασης του TIMIT-DEMAND με NMF στην περίπτωση όπου $Kn = 1$.

Στο Σχήμα 5.3.1 έχουμε θηκογράμματα με τις αποδόσεις στο TIMIT-DEMAND για διαφορετικά Ks και σταθερό

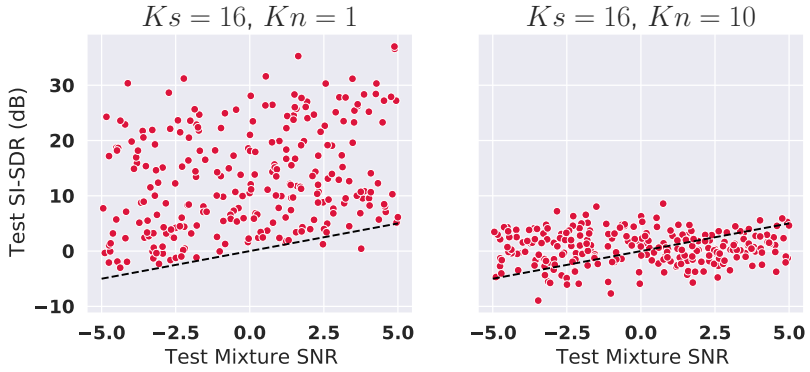
Kn ίσο με 1. Παρατηρώντας το σχήμα αυτό, συμπεραίνουμε ότι η απόδοση στον διαχωρισμό δεν εξαρτάται έντονα από την ικανότητα ανακατασκευής "καθαρής" ομιλίας, η οποία μεγιστοποιείται για $Ks = 128$.

Επιλέγουμε το μοντέλο με $Ks = 16$ και $Kn = 1$ που έχει την καλύτερη απόδοση στο TIMIT-DEMAND ώστε να εξετάσουμε την συμπεριφορά του για διαφορετικούς τύπους θορύβου. Στον Πίνακα 5.5 έχουμε την απόδοση ανά τύπο θορύβου όπου βλέπουμε ότι υπάρχουν διαφοροποιήσεις ανάμεσά τους.

Noise Type	Domestic	Nature	Office	Public	Street	Transportation
Test Mean SI-SDR (dB)	13.4	11.0	13.6	11.8	11.7	**15.3**

Πίνακας 5.5: Απόδοση ανά τύπο θορύβου μεθόδου NMF στο σύνολο εξέτασης του TIMIT-DEMAND με $Ks = 16$ και $Kn = 1$.

Στο Σχήμα 5.3.2 εξετάζουμε δυο μοντέλα NMF με $Ks = 16$, $Kn = 1$ και $Ks = 16$, $Kn = 10$ στο TIMIT-DEMAND, όπου απεικονίζουμε την απόδοση σε κάθε δείγμα σε σχέση με το SNR αυτού. Όσα δείγματα είναι πάνω από την διακεκομμένη γραμμή σημαίνει ότι παρουσιάζουν βελτίωση σε σχέση με το θορυβώδες σήμα. Στην περίπτωση $Kn = 1$ παρατηρούμε ότι σχεδόν για όλα τα δείγματα είμαστε πάνω από αυτή την γραμμή, ενώ μπορούμε να πούμε ότι όσο αυξάνει το SNR τόσο αυξάνεται η απόδοση. Αντίθετα, στην περίπτωση $Kn = 10$ έχουμε συνολικά χαμηλή απόδοση και μάλιστα σε υψηλά SNR βρισκόμαστε κάτω από την διακεκομμένη γραμμή, δηλαδή δεν βελτιώνουμε σε σχέση με το μείγμα.



Σχήμα 5.3.2: Απόδοση μοντέλου και SNR μείγματος για κάθε δείγμα στο σύνολο εξέτασης του TIMIT-DEMAND με NMF στις περιπτώσεις όπου $Ks = 16$, $Kn = 1$ και $Ks = 16$, $Kn = 10$. Η διακεκομμένη γραμμή έχει κλίση 1 και διέρχεται από το $(0, 0)$.

5.4 Εκπαίδευση NAE σε Σήματα Ομιλίας

Στην ενότητα αυτή εκπαιδεύουμε μοντέλα NAE διαφόρων τάξεων και βάθους στο σύνολο δεδομένων TIMIT. Η προεπεξεργασία των δεδομένων είναι αυτή που περιγράψαμε στην υποενότητα 5.1.2. Εκπαιδεύουμε μοντέλα NAE στο πεδίο της συχνότητας και στο πεδίο του χρόνου, όπως περιγράψαμε στην ενότητα 4.4.

5.4.1 Εκπαίδευση στο TIMIT με Σφάλμα στο Πεδίο της Συχνότητας

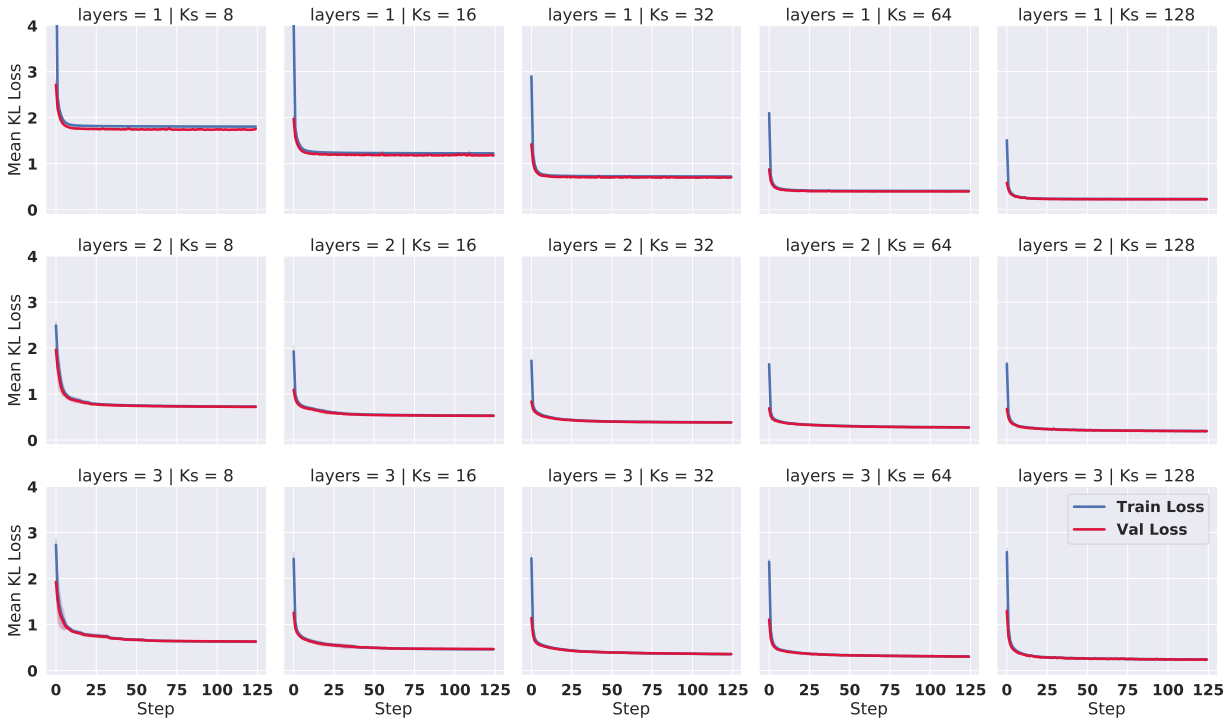
Στα πειράματα εκπαίδευσης που πραγματοποιούμε χρησιμοποιούμε τον αλγόριθμο Adam με ρυθμό εκπαίδευσης ίσο με 0.001. Τα κομμάτια με τα οποία δουλεύει ο αλγόριθμος Adam σε κάθε βήμα αποτελούνται από 2048 χρονικά παράθυρα μέτρου STFT τυχαία επιλεγμένα από τα δείγματα του συνόλου εκπαίδευσης. Επομένως, η είσοδος που δίνεται στο μοντέλο δεν προέρχεται μόνο από ένα σήμα ομιλίας. Ο αριθμός των βημάτων εκπαίδευσης είναι σταθερός και ίσος με 125.

Επίσης, πριν δοθεί η είσοδος στο μοντέλο υπολογίζεται η ρίζα του μέτρου του STFT η οποία και δίνεται ως είσοδος, ενώ υψώνουμε την έξοδο στο τετράγωνο ώστε να πάρουμε την εκτίμηση του μέτρου του STFT. Ο λόγος που γίνεται αυτό είναι ο περιορισμός του εύρος του σήματος εισόδου.

		Mean SI-SDR (dB)				
		Mean of Models				
layers	Ks	8	16	32	64	128
1		4.5	8.1	12.6	17.6	22.6
2		11.7	14.1	17.1	20.6	**23.3**
3		12.9	15.3	17.8	19.5	21.6

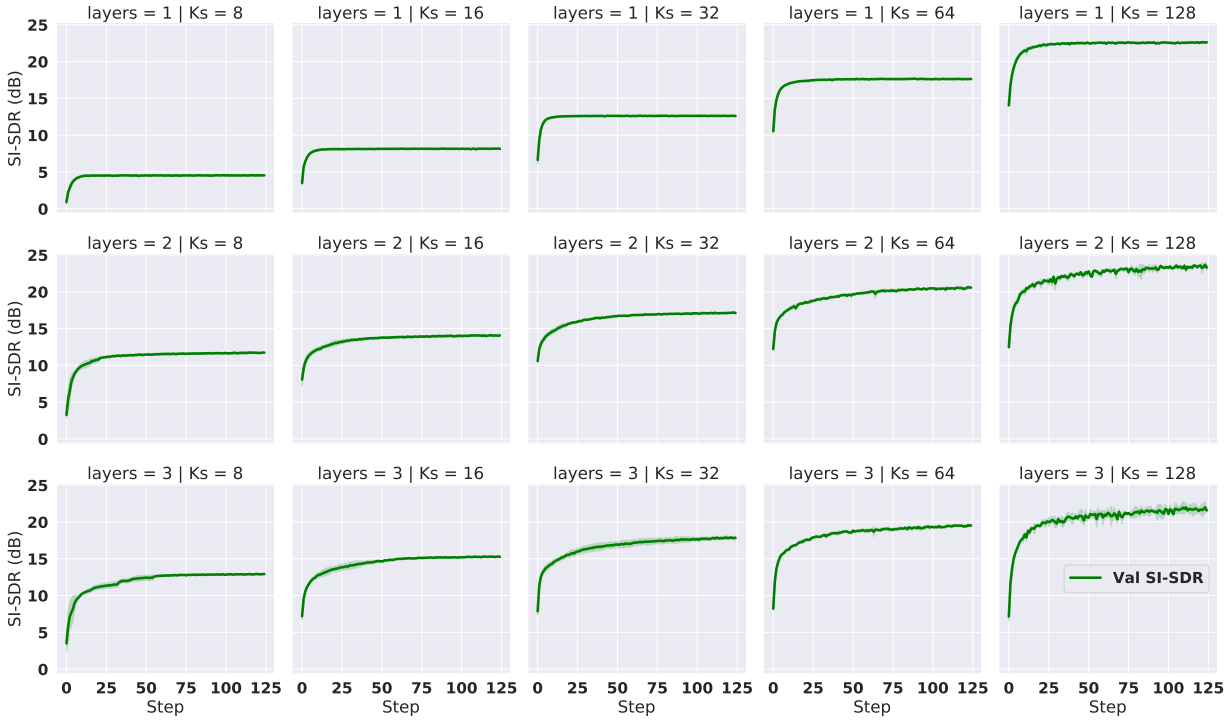
Πίνακας 5.6: Εκπαίδευση NAE σε σήματα ομιλίας του συνόλου δεδομένων TIMIT στο πεδίο της συχνότητας, για κάθε συνδυασμό παραμέτρων εκπαιδεύουμε τρία μοντέλα. Μέση απόδοση των μοντέλων που εκπαιδεύουμε για κάθε συνδυασμό τάξης και βάθους, στο σύνολο επαλήθευσης σε SI-SDR.

Στον Πίνακα 5.6 έχουμε την μέση απόδοση, στο σύνολο επαλήθευσης του TIMIT, των μοντέλων NAE με διαφορετικούς συνδυασμούς τάξης Ks και αριθμού επιπέδων. Παρατηρούμε ότι όταν αυξάνεται η τάξη του μοντέλου έχουμε και αύξηση στην απόδοση ανακατασκευής. Επίσης, στις περισσότερες περιπτώσεις, αύξηση της απόδοσης έχουμε, όταν αυξάνουμε τον αριθμό των επιπέδων. Επισημαίνουμε, ότι για κάθε συνδυασμό παραμέτρων έχουμε εκπαιδεύσει τρία μοντέλα και τα αποτελέσματα στον πίνακα αποτελούν τον μέσο όρο σε κάθε περίπτωση.



Σχήμα 5.4.1: Καμπύλες μέσου σφάλματος στο σύνολο εκπαίδευσης (μπλε) και στο σύνολο επαλήθευσης (κόκκινο) κατά την εκπαίδευση μοντέλων NAE στο πεδίο της συχνότητας.

Στο Σχήμα 5.4.1 έχουμε τις καμπύλες του σφάλματος στο σύνολο εκπαίδευσης και επαλήθευσης ως προς το βήμα εκπαίδευσης, για τους διάφορους συνδυασμούς τάξης Ks και αριθμό επιπέδων. Οι καμπύλες προκύπτουν από το μέσο όρο των επιμέρους καμπυλών και η σκιαγράφηση απεικονίζει την τυπική απόκλισή τους. Στο Σχήμα 5.4.2 έχουμε τις αντίστοιχες μέσες καμπύλες απόδοσης σε SI-SDR στο σύνολο επαλήθευσης ως προς το βήμα εκπαίδευσης. Παρατηρούμε ότι οι καμπύλες είναι αρκετά ομαλές και έχουν περάσει αρκετά βήματα ώστε να



Σχήμα 5.4.2: Καμπύλη μέσης απόδοσης στο σύνολο επαλήθευσης (πράσινο) κατά την εκπαίδευση μοντέλων NAE στο πεδίο της συχνότητας. Η σκιαγράφηση όπου υπάρχει αναπαριστά την τυπική απόκλιση.

έχουμε σύγκλιση.

		Mean SI-SDR (dB)		
layers	Ks	1	2	4
1		-10.1	0.2	1.4
2		0.5	2.6	7.1
3		-5.5	4.7	**9.1**

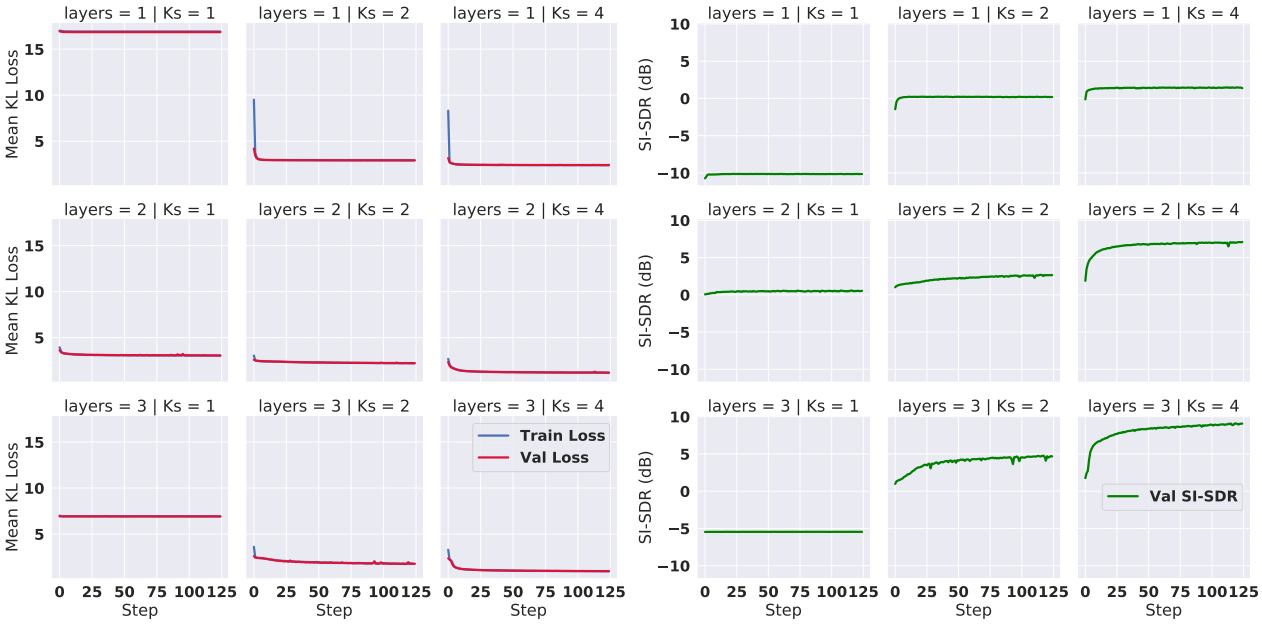
Πίνακας 5.7: Απόδοση των μοντέλων NAE με τάξη Ks ίση με 1, 2 ή 4 και βάθους ενός μέχρι τρία επίπεδα, στο σύνολο επαλήθευσης TIMIT σε SI-SDR.

Ακόμη, εκπαιδεύουμε μοντέλα με τάξη Ks ίση με 1, 2 ή 4. Όμως, στην περίπτωση αυτή εκπαιδεύουμε ένα μοντέλο ανά συνδυασμό. Τα αποτελέσματα ανακατασκευής παρατίθενται στον Πίνακα 5.7. Παρατηρούμε ότι η απόδοση για $Ks = 1$ είναι πολύ κακή ενώ μόνο για $Ks = 4$ έχουμε ικανοποιητική απόδοση. Αντίστοιχα, στο Σχήμα 5.4.3 έχουμε τις καμπύλες σφάλματος και απόδοσης για την συγκεκριμένη περίπτωση. Βλέπουμε ότι οι καμπύλες σχετικά είναι ομαλές, όμως στην περίπτωση $Ks = 1$ φαίνεται σαν να μην μαθαίνει το μοντέλο.

5.4.2 Εκπαίδευση στο TIMIT με Σφάλμα στο Πεδίο του Χρόνου

Στη συνέχεια, εκπαιδεύουμε μοντέλα NAE με σφάλμα στο πεδίο του χρόνου. Χρησιμοποιούμε τον αλγόριθμο Adam με ρυθμό εκπαίδευσης ίσο με 0.005. Τα κομμάτια με τα οποία δουλεύει ο αλγόριθμος Adam σε κάθε βήμα αποτελούνται από 16 σήματα ομιλίας. Δηλαδή αφού υπολογίσουμε το σφάλμα για κάθε ένα από αυτά πραγματοποιούμε ένα βήμα ενημέρωσης των παραμέτρων του μοντέλου.

Επίσης, αφού το μοντέλο υπολογίσει το μέτρο του STFT παίρνει την ρίζα αυτού την οποία και επεξεργάζεται το μοντέλο, ενώ υψώνουμε την έξοδο στο τετράγωνο ώστε να πάρουμε την εκτίμηση του μέτρου του STFT. Όπως πριν λόγος που γίνεται αυτό είναι ο περιορισμός του εύρος του σήματος εισόδου.



Σχήμα 5.4.3: Εκπαιδεύουμε μοντέλα NAE στο πεδίο της συχνότητας, με τάξη Ks ίση με 1, 2 ή 4. Στα αριστερά έχουμε τις καμπύλες σφάλματος στο σύνολο εκπαίδευσης (μπλε) και στο σύνολο επαλήθευσης (κόκκινο). Στα δεξιά έχουμε τις καμπύλες απόδοσης στο σύνολο επαλήθευσης (πράσινο).

Εκτός από την περίπτωση των $Ks = 1, 2, 4$, αν έχουμε απόδοση στο σύνολο επαλήθευσης κάτω από 0 dB τότε απορρίπτουμε το μοντέλο. Κατά την εκπαίδευση μοντέλων παρατηρήσαμε τέτοια πτώση της απόδοσης σε λίγες περιπτώσεις με τρία επίπεδα.

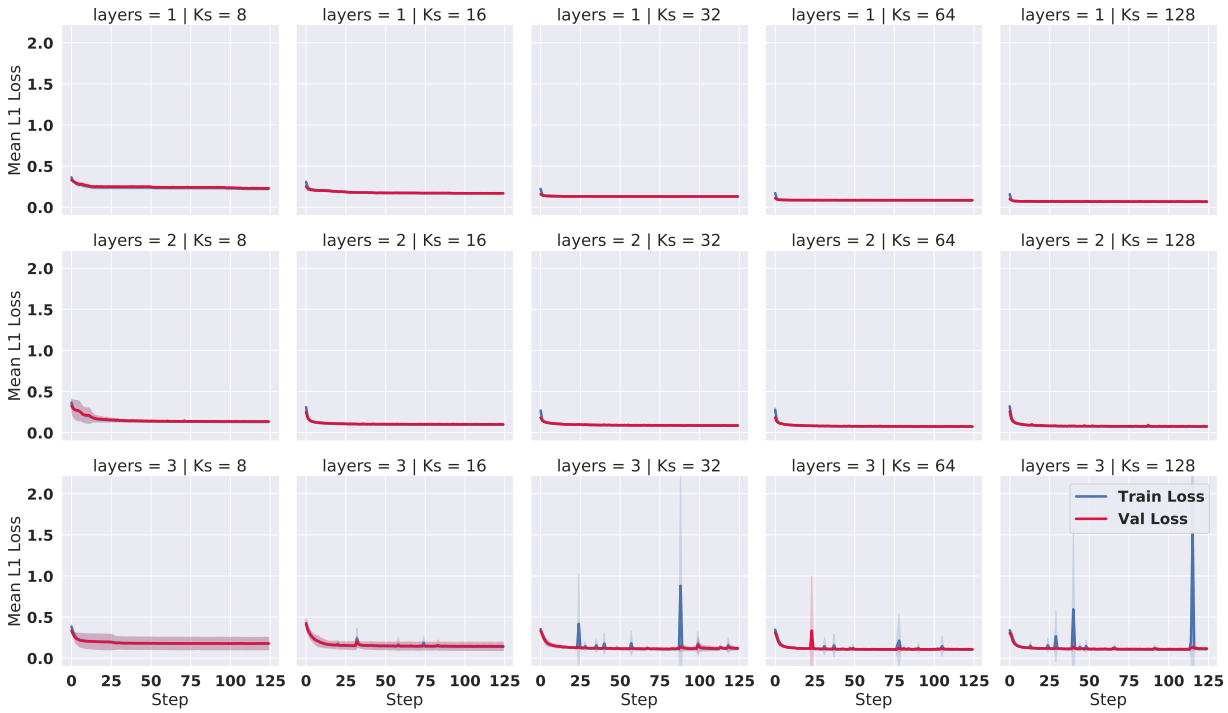
		Mean SI-SDR (dB)				
		Mean of Models				
layers	Ks	8	16	32	64	128
1		5.9	9.0	11.8	15.8	**17.9**
2		11.5	14.4	15.9	17.3	17.1
3		8.7	11.3	12.7	13.7	12.8

Πίνακας 5.8: Εκπαίδευση NAE σε σήματα ομιλίας του συνόλου δεδομένων TIMIT στο πεδίο του χρόνου. Μέση απόδοση των μοντέλων που εκπαιδεύουμε σε κάθε κατηγορία, στο σύνολο επαλήθευσης σε SI-SDR.

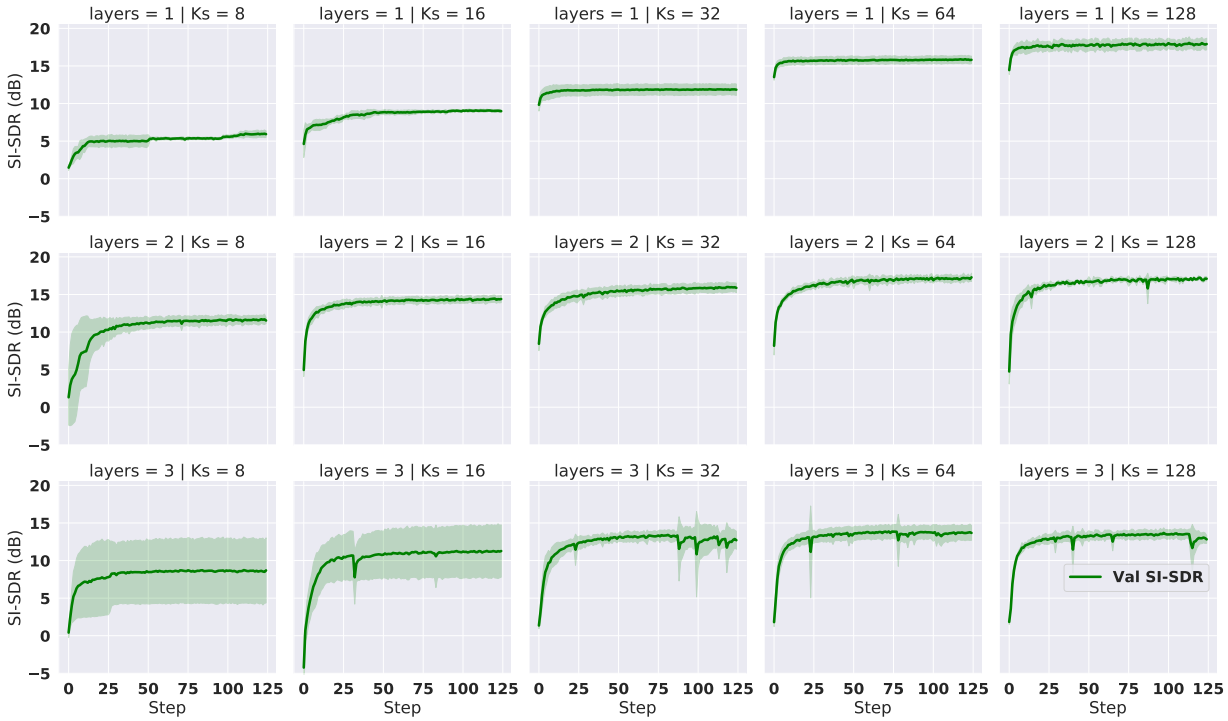
Στον Πίνακα 5.8 έχουμε την μέση απόδοση, στο σύνολο επαλήθευσης του TIMIT, των μοντέλων NAE με διαφορετικούς συνδυασμούς τάξης Ks και αριθμού επιπέδων. Για κάθε συνδυασμό παραμέτρων έχουμε εκπαιδεύσει τουλάχιστον τρία μοντέλα. Παρατηρούμε ότι όταν αυξάνεται η τάξη του μοντέλου έχουμε και αύξηση στην απόδοση ανακατασκευής. Όμως, από τα δυο επίπεδα στα τρία, δεν έχουμε αύξηση της απόδοσης. Σε σχέση με τα μοντέλα του πεδίου συχνότητας η απόδοση είναι μειωμένη.

Στο Σχήμα 5.4.4 έχουμε τις καμπύλες του σφάλματος στο σύνολο εκπαίδευσης και επαλήθευσης ως προς το βήμα εκπαίδευσης, για τους διάφορους συνδυασμούς τάξης Ks και αριθμό επιπέδων. Οι καμπύλες προκύπτουν από το μέσο όρο των επιμέρους καμπυλών και η σκιαγράφηση απεικονίζει την τυπική απόκλισή τους. Στο Σχήμα 5.4.5 έχουμε τις αντίστοιχες μέσες καμπύλες απόδοσης σε SI-SDR στο σύνολο επαλήθευσης ως προς το βήμα εκπαίδευσης. Παρατηρούμε ότι οι καμπύλες είναι λιγότερο ομαλές σε σχέση με την περίπτωση των μοντέλων NAE εκπαιδευμένα με σφάλμα στο πεδίο συχνότητας. Ειδικά για την περίπτωση των τριών επιπέδων, έχουμε απότομες αλλαγές στο σφάλμα εκπαίδευσης, ενώ οι καμπύλες απόδοσης έχουν μεγάλη τυπική απόκλιση.

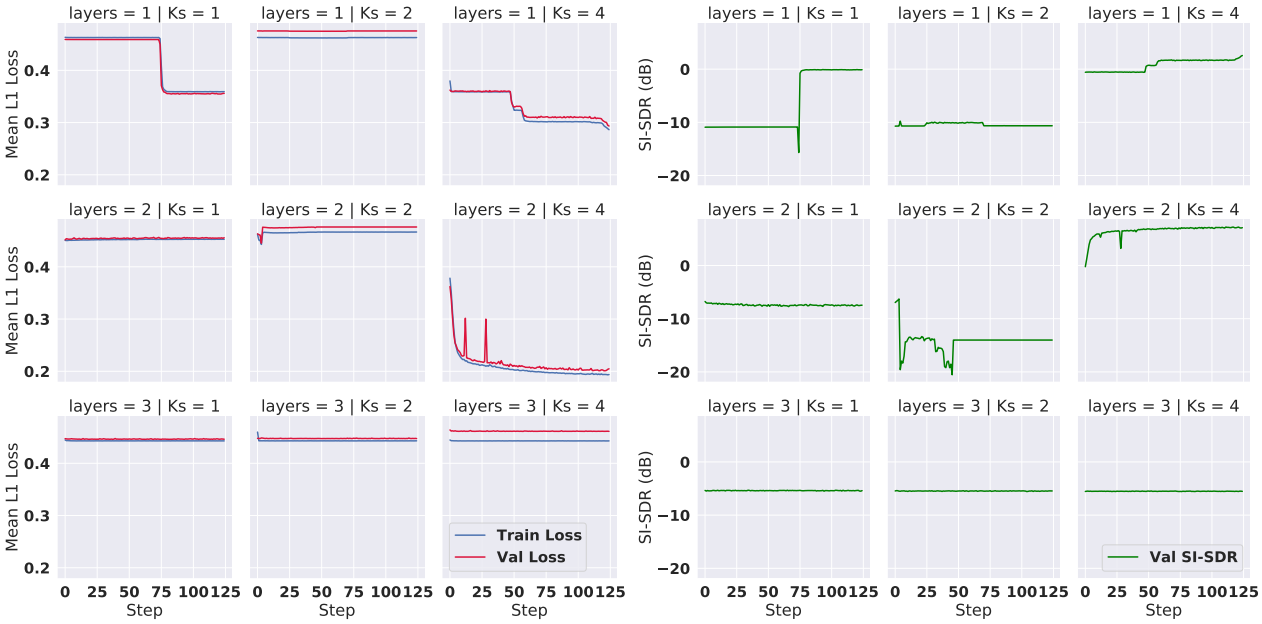
Έπειτα, εκπαιδεύουμε μοντέλα με $Ks = 1, 2, 4$ και από ένα έως τρία επίπεδα. Εκπαιδεύουμε ένα μοντέλο



Σχήμα 5.4.4: Καμπύλες σφάλματος στο σύνολο εκπαίδευσης (μπλε) και στο σύνολο επαλήθευσης (κόκκινο) κατά την εκπαίδευση μοντέλων NAE στο πεδίο του χρόνου. Η σκιαγράφηση αναπαριστά την τυπική απόκλιση.



Σχήμα 5.4.5: Καμπύλη μέσης απόδοσης στο σύνολο επαλήθευσης (πράσινο) κατά την εκπαίδευση μοντέλων NAE στο πεδίο του χρόνου. Η σκιαγράφηση αναπαριστά την τυπική απόκλιση.



Σχήμα 5.4.6: Εκπαιδεύουμε μοντέλα NAE στο πεδίο του χρόνου, με τάξη Ks ίση με 1, 2 ή 4. Στα αριστερά έχουμε τις καμπύλες σφάλματος στο σύνολο εκπαίδευσης (μπλε) και στο σύνολο επαλήθευσης (κόκκινο). Στα δεξιά έχουμε τις καμπύλες απόδοσης στο σύνολο επαλήθευσης (πράσινο).

ανά συνδυασμό παραμέτρων. Στον Πίνακα 5.9 έχουμε την απόδοση στο σύνολο επαλήθευσης του TIMIT. Βλέπουμε ότι μόνο τα μοντέλα με $Ks = 4$ και ένα ή δυο επίπεδα έχουν θετικό SI-SDR. Στο Σχήμα 5.4.6 έχουμε τις αντίστοιχες καμπύλες, από τις οποίες παρατηρούμε ότι κατά την εκπαίδευση η συμπεριφορά είναι ακανόνιστη.

		Mean SI-SDR (dB)		
layers	Ks	1	2	4
1		-0.1	-10.6	2.6
2		-7.5	-14.0	**7.1**
3		-5.4	-5.5	-5.5

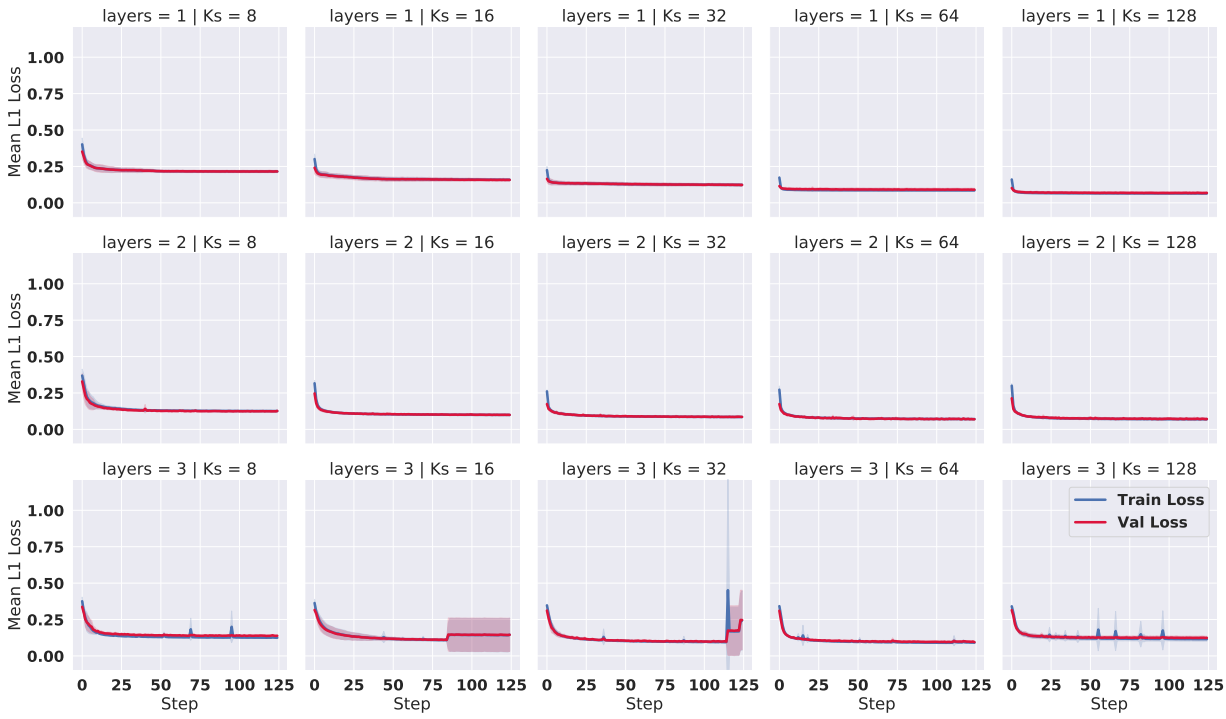
Πίνακας 5.9: Εκπαίδευση μοντέλων NAE, με πολύ μικρή τάξη Ks, στο TIMIT στο πεδίο του χρόνου. Απόδοση στο σύνολο επαλήθευσης σε SI-SDR.

Τέλος, εκπαιδεύουμε μοντέλα με επίπεδα με πολώσεις. Εκπαιδεύουμε τουλάχιστον τρία μοντέλα ανά συνδυασμό παραμέτρων. Στον Πίνακα 5.10 έχουμε την απόδοση στο σύνολο επαλήθευσης του TIMIT. Παρατηρούμε ότι σε σχέση με την περίπτωση χωρίς πολώσεις τα αποτελέσματα είναι αρκετά κοντά, με κάποιες διαφορές για τις περιπτώσεις όπου ο αριθμός των επιπέδων είναι ίσος με τρία.

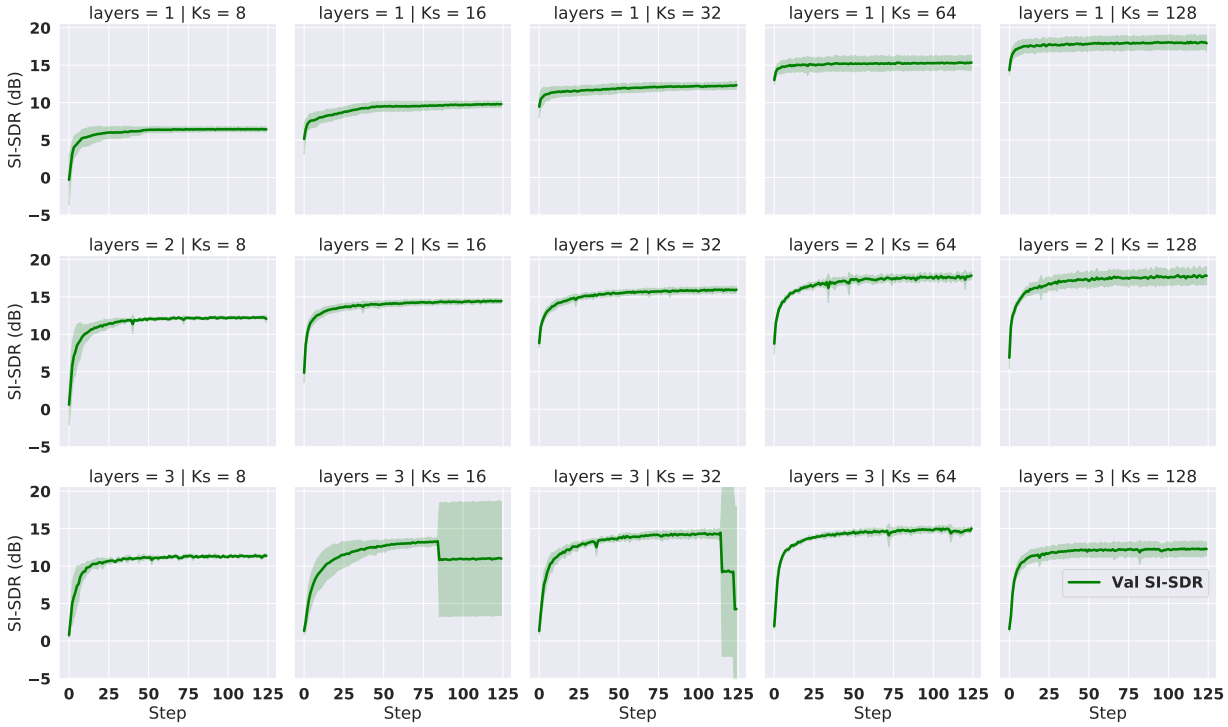
Στα Σχήματα 5.4.7 και 5.4.8 έχουμε τις αντίστοιχες καμπύλες σφάλματος και απόδοσης, από τις οποίες παρατηρούμε ότι σε ορισμένα σημεία ειδικά για την περίπτωση με τρία επίπεδα η τυπική απόκλιση είναι αρκετά μεγάλη. Ακόμη, σε ορισμένες περιπτώσεις παρατηρούμε απότομες πτώσεις της απόδοσης.

		Mean SI-SDR (dB)				
		Mean of Models				
layers	Ks	8	16	32	64	128
1		6.4	9.8	12.3	15.3	**17.9**
2		12.1	14.4	16.0	**17.8**	**17.8**
3		11.4	11.0	4.3	15.0	12.3

Πίνακας 5.10: Εκπαίδευση μοντέλων NAE με πολώσεις στο TIMIT στο πεδίο του χρόνου. Μέση απόδοση των μοντέλων που εκπαιδεύουμε σε κάθε συνδυασμό παραμέτρων, στο σύνολο επαλήθευσης σε SI-SDR.



Σχήμα 5.4.7: Καμπύλες σφάλματος στο σύνολο εκπαίδευσης (μπλε) και στο σύνολο επαλήθευσης (κόκκινο) κατά την εκπαίδευση μοντέλων NAE στο πεδίο του χρόνου με πολώσεις. Η σκιαγράφηση αναπαριστά την τυπική απόκλιση.



Σχήμα 5.4.8: Καμπύλη μέση απόδοσης στο σύνολο επαλήθευσης (πράσινο) κατά την εκπαίδευση μοντέλων NAE στο πεδίο του χρόνου με πολώσεις. Η σκιαγράφηση αναπαριστά την τυπική απόκλιση.

5.5 Πειράματα Προσαρμογής Μοντέλου NAE για τον Διαχωρισμό

Πριν πραγματοποιήσουμε πειράματα διαχωρισμού με μοντέλα NAE, δοκιμάζουμε μερικά μοντέλα με διαφορετικές ρυθμίσεις στο σύνολο επαλήθευσης του TIMIT-DEMAND. Στόχος μας είναι να βρούμε τις κατάλληλες ρυθμίσεις για τα πειράματα διαχωρισμού που θα ακολουθήσουν. Συγκεκριμένα, επιλέγουμε $Ks = 16$ και $Kn = 10$ και αριθμό επιπέδων από ένα έως τρία. Για κάθε συνδυασμό ρυθμίσεων και παραμέτρων εκπαιδεύουμε τρία μοντέλα. Οι ρυθμίσεις που εξετάζουμε είναι οι εξής:

- **Σφάλμα**: Δοκιμάζουμε σφάλμα $\mathrm{L}_1\left(\boldsymbol{x}, \hat{\boldsymbol{x}}\right) = \left\|\boldsymbol{x} - \hat{\boldsymbol{x}}\right\|_1$ στο πεδίο του χρόνου, σφάλμα SI-SDR στο πεδίο του χρόνου και γενικευμένη απόκλιση Kullback Leibler στο πεδίο της συχνότητας (KLD).
- **Πόλωση**: Δοκιμάζουμε μοντέλα που έχουν ή δεν έχουν πόλωση στα επίπεδά τους.
- **Συμπίεση εύρους μέτρου STFT**: Δοκιμάζουμε το μοντέλο να δουλεύει με το μέτρο STFT να δουλεύει αντί με την τετραγωνική ρίζα αυτού.

Για κάθε συνδυασμό ρυθμίσεων που δοκιμάζουμε εκπαιδεύουμε ένα μοντέλο με αυτές τις ρυθμίσεις και μετά πραγματοποιούμε τον διαχωρισμό στο TIMIT-DEMAND, με τις ίδιες ρυθμίσεις.

Στον Πίνακα 5.11 έχουμε τους συνδυασμούς ρυθμίσεων που δοκιμάσαμε με τις αντίστοιχες αποδόσεις στον διαχωρισμό στο TIMIT-DEMAND.

Παρατηρούμε αρχικά ότι δουλεύοντας με ρίζα του μέτρου του STFT έχουμε καλύτερη απόδοση σε σχέση με την αντίθετη περίπτωση. Η χρήση πόλωσης δοκιμάζεται μόνο για την περίπτωση L_1 σφάλματος στο πεδίο του χρόνου και δεν παρουσιάζει μεγάλη διαφορά εκτός από την περίπτωση με τρία επίπεδα όπου δίνει χειρότερη απόδοση. Η χρήση SI-SDR στο πεδίο του χρόνου επιτυγχάνει την καλύτερη απόδοση για δυο επίπεδα αλλά για ένα επίπεδο έχει πολύ χαμηλή απόδοση. Έπειτα, η χρήση γενικευμένης απόκλισης Kullback Leibler στο πεδίο της συχνότητας, επιτυγχάνει την βέλτιστη απόδοση με ένα επίπεδο. Όμως, για δυο και τρία επίπεδα η απόδοση

	Loss	$\sqrt{\lvert\text{STFT}\rvert}$	Bias	layers	Mean SI-SDR (dB) 1	2	3
Validation	L_1 time domain				2.2	2.2	2.5
	L_1 time domain	✓			9.1	6.4	**8.6**
	L_1 time domain	✓	✓		9.1	5.9	4.8
	SI-SDR time domain	✓			3.8	**7.1**	6.9
	KLD frequency domain	✓			**11.0**	5.2	4.9
	KLD frequency domain				8.2	0.9	2.4
Test	L_1 time domain				2.4	2.4	2.3
	L_1 time domain	✓			9.3	6.4	**8.7**
	L_1 time domain	✓	✓		9.3	6.1	5.2
	SI-SDR time domain	✓			4.0	**7.5**	7.3
	KLD frequency domain	✓			**11.2**	5.2	4.9
	KLD frequency domain				8.4	1.1	2.6

Πίνακας 5.11: Προσαρμογή των ρυθμίσεων της μεθόδου NAE στο TIMIT-DEMAND για Ks=16 και Kn=10

είναι σχετικά χαμηλή. Τέλος, με L_1 σφάλμα στο πεδίο του χρόνου πετυχαίνουμε την καλύτερη απόδοση για τρία επίπεδα και ικανοποιητική απόδοση για ένα και δυο.

Συνεπώς, επειδή μας ενδιαφέρει κυρίως η επέκταση με περισσότερα από ένα επίπεδα επιλέγουμε να ασχοληθούμε κυρίως με L_1 σφάλμα στο πεδίο του χρόνου, χωρίς πόλωση και με ρίζα του μέτρου του STFT. Ακόμη, λόγω της καλής απόδοσης πραγματοποιούμε κάποια πειράματα με την χρήση γενικευμένης απόκλισης Kullback Leibler στο πεδίο της συχνότητας. Τέλος, δοκιμάσουμε κάποια πειράματα με L_1 σφάλμα στο πεδίο του χρόνου και πολώσεις.

5.6 Πειράματα Διαχωρισμού με NAE

Όπως, αναφέραμε δοκιμάζουμε μοντέλα NAE με L_1 σφάλμα στο πεδίο του χρόνου και ρίζα μέτρου STFT καθώς και μοντέλα NAE με γενικευμένη απόκλιση Kullback Leibler στο πεδίο της συχνότητας και ρίζα μέτρου STFT. Σε όλα τα παρακάτω πειράματα χρησιμοποιούμε Adam με ρυθμό εκπαίδευσης ίσο με 0.001.

5.6.1 Σφάλμα στο Πεδίο του Χρόνου

Αρχικά, δοκιμάζουμε μοντέλα NAE στο TIMIT-DEMAND με τάξεις μοντέλου ομιλίας $Ks = 8$, 16, 32, 64, 128 και ένα έως τρία επίπεδα στο μοντέλο ομιλίας, στις εξής περιπτώσεις:

- $Kn = 10$ και ένα επίπεδο στο μοντέλο θορύβου.
- $Kn = Ks/2$ και ένα επίπεδο στο μοντέλο θορύβου.
- $Kn = 10$ και ίσα επίπεδα στο μοντέλο θορύβου με τα επίπεδα στο μοντέλο ομιλίας.
- $Kn = Ks/2$ και ίσα επίπεδα στο μοντέλο θορύβου με τα επίπεδα στο μοντέλο ομιλίας.

Για κάθε συνδυασμό παραμέτρων και περίπτωση αξιολογούμε τουλάχιστον τρία μοντέλα, έτσι στον Πίνακα 5.12 έχουμε τα αποτελέσματα του καλύτερου μοντέλου με βάση το σύνολο επαλήθευσης και τη μέση απόδοση των μοντέλων.

Ακόμη, ορμώμενοι από τα καλά αποτελέσματα της NMF, στον Πίνακα 5.13 έχουμε τα αποτελέσματα της περίπτωσης:

- $Kn = 1$ και ένα επίπεδο στο μοντέλο θορύβου.

					Mean SI-SDR (dB)									
					Best Model					Mean of Models				
	Kn	layers	layers noise	Ks	8	16	32	64	128	8	16	32	64	128
Val	10	1	1		9.2	9.5	6.4	2.7	3.3	7.1	9.0	3.3	2.3	1.5
	10	2	1		10.5	7.3	5.5	4.5	3.0	9.5	6.3	3.9	2.7	1.8
	10	3	1		**12.0**	9.0	7.0	5.0	4.1	**10.2**	8.5	5.9	3.7	3.4
Test	10	1	1		9.5	9.8	6.7	3.1	3.7	6.8	9.6	3.4	2.5	1.7
	10	2	1		10.8	7.4	5.9	4.4	3.5	9.8	6.4	4.0	2.8	2.1
	10	3	1		**12.4**	9.3	6.8	5.1	4.4	**10.5**	8.7	6.0	4.0	3.6
Val	Ks / 2	1	1		10.1	9.6	6.6	3.6	4.3	7.7	9.3	3.2	2.6	2.2
	Ks / 2	2	1		10.6	7.3	5.7	4.6	3.4	8.8	6.3	4.0	2.7	2.9
	Ks / 2	3	1		**11.0**	8.9	7.7	7.2	7.0	**9.5**	8.2	6.8	5.8	6.4
Test	Ks / 2	1	1		10.2	10.0	7.1	3.5	4.7	7.3	9.5	3.4	2.7	2.4
	Ks / 2	2	1		10.8	7.4	6.0	4.6	3.5	9.1	6.4	3.9	2.8	3.1
	Ks / 2	3	1		**11.9**	9.0	7.6	7.7	7.2	**9.8**	8.4	6.9	6.2	6.7
Val	10	1	1		9.2	9.5	6.4	2.7	3.3	7.1	9.0	3.3	2.3	1.5
	10	2	2		9.9	7.0	6.9	5.5	5.0	8.6	6.3	4.3	3.3	3.7
	10	3	3		**11.3**	10.0	9.6	8.7	8.4	**10.1**	9.2	9.0	7.4	7.4
Test	10	1	1		9.5	9.8	6.7	3.1	3.7	6.8	9.6	3.4	2.5	1.7
	10	2	2		10.2	7.1	7.1	5.4	5.2	8.8	6.4	4.3	3.4	3.9
	10	3	3		**11.3**	10.7	10.2	8.9	8.8	**10.2**	9.8	9.6	8.0	7.8
Val	Ks / 2	1	1		10.1	9.6	6.6	3.6	4.3	7.7	9.3	3.2	2.6	2.2
	Ks / 2	2	2		10.2	7.1	6.8	5.5	5.3	9.1	6.3	4.2	3.2	3.8
	Ks / 2	3	3		**11.1**	10.5	9.8	8.8	8.6	**10.2**	9.3	9.1	7.8	7.6
Test	Ks / 2	1	1		10.2	10.0	7.1	3.5	4.7	7.3	9.5	3.4	2.7	2.4
	Ks / 2	2	2		10.6	7.1	7.0	5.4	5.5	9.2	6.4	4.2	3.3	4.0
	Ks / 2	3	3		**11.5**	10.7	9.4	9.4	8.9	**10.3**	9.4	9.3	8.1	7.9

Πίνακας 5.12: Απόδοση της μεθόδου NAE στο TIMIT-DEMAND.

					Mean SI-SDR (dB)									
					Best Model					Mean of Models				
	Kn	layers	layers noise	Ks	8	16	32	64	128	8	16	32	64	128
Val	1	1	1		6.4	6.0	1.3	0.8	0.9	4.5	3.8	0.5	0.5	0.4
	1	2	1		**7.8**	2.5	1.1	1.5	0.3	**5.1**	1.5	0.5	0.1	0.1
	1	3	1		3.1	3.3	2.5	0.7	1.0	3.1	2.9	1.4	0.2	0.2
Test	1	1	1		4.1	4.8	1.1	1.0	0.7	4.1	3.8	0.5	0.7	0.4
	1	2	1		**5.9**	1.8	0.8	1.4	0.3	**4.9**	1.5	0.6	0.3	0.2
	1	3	1		3.4	3.7	1.8	0.6	0.9	3.8	2.7	1.2	0.5	0.4

Πίνακας 5.13: Απόδοση της μεθόδου NAE στο TIMIT-DEMAND.

Στην συνέχεια, για να διερευνήσουμε μοντέλα NAE με μικρές τάξεις, δοκιμάζουμε στο TIMIT-DEMAND μοντέλα με τάξεις ομιλίας $Ks = 1, 2, 4$ και ένα έως τρία επίπεδα στο μοντέλο ομιλίας, στις εξής περιπτώσεις:

- $Kn = 1$ και ένα επίπεδο στο μοντέλο θορύβου.
- $Kn = 10$ και ένα επίπεδο στο μοντέλο θορύβου.

Αυτή την φορά, για κάθε συνδυασμό παραμέτρων και περίπτωση αξιολογούμε ένα μοντέλο. Στον Πίνακα 5.14 έχουμε τα αποτελέσματα των περιπτώσεων αυτών.

					Mean SI-SDR (dB)		
	Kn	layers	layers noise	Ks	1	2	4
Val	1	1	1		-3.5	5.3	1.3
	1	2	1		2.2	-10.1	**7.4**
	1	3	1		6.6	5.7	5.1
Test	1	1	1		-4.4	4.9	3.3
	1	2	1		2.1	-11.6	**9.5**
	1	3	1		6.5	7.2	6.3
Val	10	1	1		-3.5	5.6	2.5
	10	2	1		0.8	-14.5	**8.8**
	10	3	1		1.6	1.5	2.2
Test	10	1	1		-3.6	6.9	2.5
	10	2	1		0.6	-16.3	**8.8**
	10	3	1		1.0	1.3	2.8

Πίνακας 5.14: Απόδοση της μεθόδου NAE στο TIMIT-DEMAND.

Έπειτα, δοκιμάζουμε μοντέλα NAE στο TIMIT-DEMAND με πολώσεις και τάξεις μοντέλου ομιλίας $Ks = 8, 16, 32, 64, 128$ και ένα έως τρία επίπεδα στο μοντέλο ομιλίας, στην εξής περίπτωση:

- $Kn = 10$ και ένα επίπεδο στο μοντέλο θορύβου.

Για κάθε συνδυασμό παραμέτρων αξιολογούμε τουλάχιστον τρία μοντέλα, έτσι στον Πίνακα 5.15 έχουμε τα αποτελέσματα του καλύτερου μοντέλου με βάση το σύνολο επαλήθευσης και τη μέση απόδοση των μοντέλων.

					Mean SI-SDR (dB)									
					Best Model					Mean of Models				
	Kn	layers	layers noise	Ks	8	16	32	64	128	8	16	32	64	128
Val	10	1	1		6.8	9.7	7.1	5.4	2.3	6.5	8.8	4.2	3.4	1.2
	10	2	1		10.4	9.1	5.1	6.4	4.2	9.7	5.8	3.5	3.1	2.4
	10	3	1		**11.6**	8.9	7.5	5.3	4.5	**11.3**	7.3	6.1	3.5	3.3
Test	10	1	1		6.7	10.2	7.6	5.8	2.5	6.5	9.2	4.5	3.6	1.4
	10	2	1		10.4	9.3	5.2	6.7	4.2	9.9	6.0	3.5	3.2	2.6
	10	3	1		**12.2**	9.1	7.6	5.1	4.5	**11.9**	7.6	6.2	3.6	3.6

Πίνακας 5.15: Απόδοση της μεθόδου NAE με πολώσεις στο TIMIT-DEMAND.

Τέλος, δοκιμάζουμε μοντέλα NAE στο TIMIT-MUSDB για να εξετάσουμε την απόδοση σε διαφορετικούς τύπους θορύβου, με τάξεις μοντέλου ομιλίας $Ks = 8, 16, 32, 64, 128$ και ένα έως τρία επίπεδα στο μοντέλο ομιλίας, στις εξής περιπτώσεις:

- $Kn = 10$ και ένα επίπεδο στο μοντέλο θορύβου.
- $Kn = Ks/2$ και ένα επίπεδο στο μοντέλο θορύβου.
- $Kn = 10$ και ίσα επίπεδα στο μοντέλο θορύβου με τα επίπεδα στο μοντέλο ομιλίας.
- $Kn = Ks/2$ και ίσα επίπεδα στο μοντέλο θορύβου με τα επίπεδα στο μοντέλο ομιλίας.

Για κάθε συνδυασμό παραμέτρων και περίπτωση αξιολογούμε τουλάχιστον τρία μοντέλα, έτσι στον Πίνακα 5.16 έχουμε τα αποτελέσματα του καλύτερου μοντέλου με βάση το σύνολο επαλήθευσης και τη μέση απόδοση των μοντέλων.

					Mean SI-SDR (dB)									
					Best Model					Mean of Models				
	Kn	layers	layers noise	Ks	8	16	32	64	128	8	16	32	64	128
Val	10	1	1		**6.0**	5.0	3.8	2.2	1.6	**5.6**	4.8	3.4	1.9	1.3
	10	2	1		**6.2**	3.6	2.7	2.0	1.6	**5.9**	3.3	2.3	1.8	1.2
	10	3	1		**6.1**	5.6	4.4	3.1	2.2	**5.7**	4.8	3.5	2.2	2.0
Test	10	1	1		**6.2**	5.7	4.4	2.5	1.7	**5.8**	5.5	3.8	2.1	1.3
	10	2	1		**6.8**	3.9	2.8	2.1	1.6	**6.5**	3.6	2.3	1.8	1.1
	10	3	1		**6.7**	5.9	4.7	3.2	2.1	**6.1**	5.0	3.7	2.1	1.9
Val	Ks / 2	1	1		**6.5**	5.2	3.6	2.6	1.9	**6.2**	4.9	3.3	2.4	1.7
	Ks / 2	2	1		**6.3**	3.5	2.9	2.4	2.1	**5.9**	3.2	2.5	2.2	1.7
	Ks / 2	3	1		5.9	5.5	4.8	4.2	3.8	**5.7**	4.6	3.9	3.0	3.3
Test	Ks / 2	1	1		**6.8**	5.9	4.3	2.9	2.2	**6.5**	5.7	3.9	2.6	1.9
	Ks / 2	2	1		**7.0**	3.8	3.1	2.7	2.3	**6.5**	3.5	2.5	2.3	1.8
	Ks / 2	3	1		**6.6**	5.9	5.1	4.5	4.1	**6.2**	4.9	4.1	3.2	3.6
Val	10	1	1		**6.0**	5.0	3.8	2.2	1.6	**5.6**	4.8	3.4	1.9	1.3
	10	2	2		**5.6**	3.6	3.1	2.6	1.9	**5.2**	3.3	2.7	2.3	1.8
	10	3	3		**6.0**	**6.0**	5.2	4.1	3.4	**5.6**	**5.1**	4.4	3.2	3.1
Test	10	1	1		**6.2**	5.7	4.4	2.5	1.7	**5.8**	**5.5**	3.8	2.1	1.3
	10	2	2		**6.2**	3.9	3.5	2.9	2.1	**5.7**	3.7	2.8	2.4	1.9
	10	3	3		**6.5**	**6.5**	5.5	4.2	3.6	**6.0**	**5.5**	4.7	3.4	3.2
Val	Ks / 2	1	1		**6.5**	5.2	3.6	2.6	1.9	**6.2**	4.9	3.3	2.4	1.7
	Ks / 2	2	2		5.7	3.6	3.1	2.7	2.1	5.3	3.3	2.7	2.4	2.0
	Ks / 2	3	3		**6.0**	**6.0**	5.3	4.2	3.7	**5.6**	5.1	4.4	3.3	3.2
Test	Ks / 2	1	1		**6.8**	5.9	4.3	2.9	2.2	**6.5**	5.7	3.9	2.6	1.9
	Ks / 2	2	2		**6.3**	3.9	3.5	3.1	2.3	5.9	3.7	2.8	2.5	2.2
	Ks / 2	3	3		**6.6**	**6.5**	5.6	4.4	3.7	**6.2**	5.5	4.7	3.5	3.4

Πίνακας 5.16: Απόδοση της μεθόδου NAE στο TIMIT-MUSDB

Εξετάζοντας τους πίνακες αποτελεσμάτων κάνουμε κάποιες παρατηρήσεις:

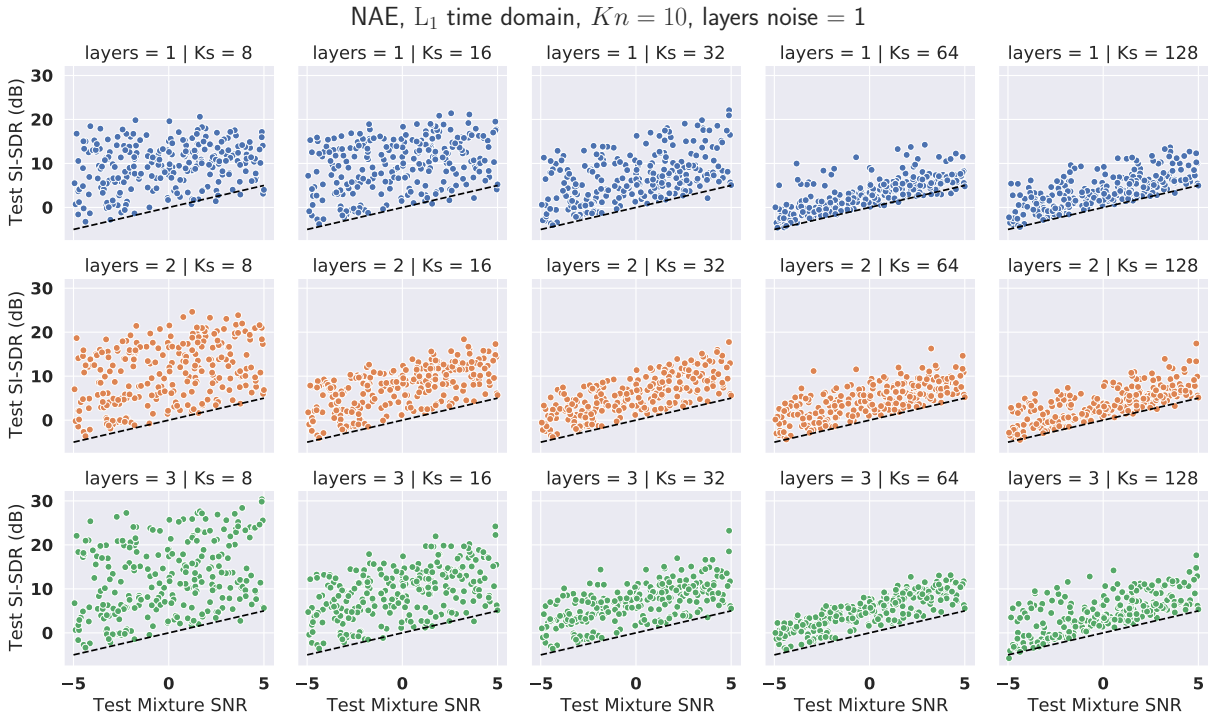
- Από τον Πίνακα 5.12 βλέπουμε ότι το μοντέλο με $Ks = 8$ και τρία επίπεδα επιτυγχάνει την υψηλότερη απόδοση σε όλες τις περιπτώσεις. Ακόμα, βελτίωση παρατηρούμε για δυο ή τρία επίπεδα όταν θέτουμε $Kn = Ks/2$ ή έχουμε ίσο αριθμό επιπέδων σε μοντέλο θορύβου και ομιλίας.
- Από τους Πίνακες 5.13 και 5.14 βλέπουμε ότι σε αντίθεση με την NMF για $Kn = 1$ δεν πετυχαίνουμε καλή απόδοση, εκτός αν το Ks είναι 4 ή 8.

- Στον Πίνακα 5.15 παρατηρούμε ότι η χρήση πολώσεων δεν διαφέρει σημαντικά από την περίπτωση χωρίς πολώσεις.

- Έπειτα, από τον Πίνακα 5.16 συμπεραίνουμε ότι με $Ks = 8$ μεγιστοποιούμε την απόδοση. Παρατηρούμε πάλι βελτίωση στην περίπτωση που έχουμε δυο ή τρία επίπεδα και θέτουμε $Kn = Ks/2$ ή έχουμε ίσο αριθμό επιπέδων σε μοντέλο θορύβου και ομιλίας. Ακόμη, σε σχέση με το TIMIT-DEMAND η απόδοση είναι αρκετά χαμηλότερη.

Συνολικά, παρατηρούμε ότι στο TIMIT-DEMAND φτάνουμε την απόδοση της NMF αλλά στο TIMIT-MUSDB η μέθοδος αυτή υστερεί σε σχέση με την NMF. Παρατηρούμε επίσης ότι η ικανότητα ανακατασκευής σημάτων ομιλίας δεν σχετίζεται τόσο άμεσα με την απόδοση στον διαχωρισμό. Για παράδειγμα, τα μοντέλα με $Ks = 8$ και τρία επίπεδα έχουν σχετικά χαμηλή απόδοση ανακατασκευής αλλά σε πολλές περιπτώσεις πετυχαίνουν τα καλύτερα αποτελέσματα.

Στο Σχήμα 5.6.1 έχουμε μια καλύτερη επισκόπηση των αποτελεσμάτων του Πίνακα 5.12. Συγκεκριμένα, απεικονίζουμε την απόδοση στο σύνολο εξέτασης του TIMIT-DEMAND σε κάθε δείγμα, σε σχέση με το SNR αυτού, για τα καλύτερα μοντέλα με $Ks = 8, 16, 32, 64, 128$, $Kn = 1$ και ένα επίπεδο στο μοντέλο θορύβου. Παρατηρούμε ότι, τα περισσότερα δείγματα έχουν απόδοση πάνω από την διακεκομμένη γραμμή που σημαίνει ότι το εκτιμώμενο σήμα φωνής αποτελεί βελτίωση σε σχέση με το μείγμα.



Σχήμα 5.6.1: Απόδοση μοντέλου και SNR μείγματος για κάθε δείγμα στο σύνολο εξέτασης του TIMIT-DEMAND των καλύτερων μοντέλων του Πίνακα 5.12 με $Kn = 10$ και ένα επίπεδο στο μοντέλο θορύβου. Η διακεκομμένη γραμμή έχει κλίση 1 και διέρχεται από το $(0, 0)$

Τέλος, επιλέγουμε από τον Πίνακα 5.12 τα μοντέλα με $Ks = 8, 16$, $Kn = 10$ και ένα επίπεδο θορύβου ώστε να εξετάσουμε την μέση απόδοσή τους ανά κατηγορία θορύβου του συνόλου εξέτασης του TIMIT-DEMAND. Στον Πίνακα 5.17 έχουμε τα αντίστοιχα αποτελέσματα. Βλέπουμε ότι η κατηγορία στην οποία επιτυγχάνεται καλύτερη απόδοση και η κατηγορία που έχουμε την χειρότερη απόδοση, είναι οι ίδιες για όλα τα μοντέλα.

			Test Mean SI-SDR (dB)					
Ks	layers	Noise Type	Domestic	Nature	Office	Public	Street	Transportation
8	1		6.8	6.0	7.2	6.7	6.2	**7.5**
8	2		9.7	7.6	10.0	8.9	8.1	**10.4**
8	3		10.5	8.8	11.1	9.8	9.1	**12.0**
16	1		10.0	7.8	10.2	9.1	8.9	**10.7**
16	2		6.8	5.0	7.1	6.1	5.3	**7.2**
16	3		9.7	7.4	10.0	8.9	8.1	**10.7**

Πίνακας 5.17: Μέση απόδοση μοντέλων ανά κατηγορία θορύβων στο σύνολο εξέτασης του TIMIT-DEMAND για $Ks = 8$, 16, $Kn = 10$, ένα έως τρία επίπεδα και ένα επίπεδο θορύβου (Από Πίνακα 5.12).

5.6.2 Σφάλμα στο Πεδίο της Συχνότητας

Αρχικά, δοκιμάζουμε μοντέλα NAE στο TIMIT-DEMAND με τάξεις μοντέλου ομιλίας $Ks = 8$, 16, 32, 64, 128 και ένα έως τρία επίπεδα στο μοντέλο ομιλίας, στην εξής περίπτωση:

- $Kn = 10$ και ένα επίπεδο στο μοντέλο θορύβου.

Για κάθε συνδυασμό παραμέτρων αξιολογούμε τουλάχιστον τρία μοντέλα, έτσι στον Πίνακα 5.18 έχουμε τα αποτελέσματα του καλύτερου μοντέλου με βάση το σύνολο επαλήθευσης και τη μέση απόδοση των μοντέλων.

				Mean SI-SDR (dB)									
				Best Model					Mean of Models				
	Kn	layers	Ks	8	16	32	64	128	8	16	32	64	128
Val	10	1		10.5	**11.3**	**11.3**	9.3	3.6	10.4	**11.0**	**11.3**	9.1	3.1
	10	2		10.9	7.8	1.9	1.3	1.6	10.7	5.2	1.7	1.1	1.4
	10	3		10.1	6.6	5.2	4.1	4.4	9.3	4.9	5.0	3.8	3.7
Test	10	1		10.9	**11.4**	**11.7**	9.6	3.9	10.6	**11.3**	**11.7**	9.3	3.3
	10	2		10.9	7.7	1.9	1.4	1.8	10.8	5.2	1.7	1.2	1.6
	10	3		10.1	6.5	5.0	4.2	4.3	9.3	4.9	5.0	4.0	3.8

Πίνακας 5.18: Απόδοση της μεθόδου NAE με σφάλμα στο πεδίο της συχνότητας στο TIMIT-DEMAND

Στη συνέχεια, δοκιμάζουμε μοντέλα NAE στο TIMIT-DEMAND με τάξεις μοντέλου ομιλίας $Ks = 1$, 2, 4 και ένα έως τρία επίπεδα στο μοντέλο ομιλίας, στις εξής περιπτώσεις:

- $Kn = 1$ και ένα επίπεδο στο μοντέλο θορύβου.
- $Kn = 10$ και ένα επίπεδο στο μοντέλο θορύβου.

Αυτή την φορά, για κάθε συνδυασμό παραμέτρων και περίπτωση αξιολογούμε ένα μοντέλο. Στον Πίνακα 5.19 έχουμε τα αποτελέσματα των περιπτώσεων αυτών.

Παρατηρούμε ότι πετυχαίνουμε υψηλή απόδοση όταν έχουμε ένα επίπεδο και τάξη μοντέλου Ks ίση με 16 ή 32. Για μεγάλο Ks ή πάνω από ένα επίπεδο βλέπουμε ότι η απόδοση πέφτει κατακόρυφα. Για μικρά Ks η συμπεριφορά στον διαχωρισμό είναι πολύ καλύτερη σε σχέση με την περίπτωση μοντέλων NAE με σφάλμα στο πεδίο του χρόνου. Συγκεκριμένα, για $Ks = 4$, $Kn = 10$ και τρία επίπεδα πετυχαίνουμε πολύ υψηλή απόδοση. Συνολικά, φτάνουμε αρκετά κοντά στην απόδοση της NMF αλλά απ' ότι φαίνεται τα μοντέλα αυτά δεν χρησιμοποιούν το βάθος τους στο έπακρο. Συνεπώς, παρόλο που για μεγαλύτερα Ks και περισσότερα επίπεδα έχουμε καλύτερη απόδοση στην ανακατασκευή σημάτων ομιλίας, αυτό δεν μεταφράζεται σε υψηλότερη απόδοση στον διαχωρισμό.

Στο Σχήμα 5.6.2 απεικονίζουμε την απόδοση στο σύνολο εξέτασης του TIMIT-DEMAND σε κάθε δείγμα σε σχέση με το SNR αυτού για τα καλύτερα μοντέλα με $Ks = 8$, 16, 32, 64, 128, $Kn = 1$ και ένα επίπεδο στο

	Kn	layers	layers noise	Ks	Mean SI-SDR (dB) 1	2	4
Val	1	1	1		-0.5	6.8	8.6
	1	2	1		4.9	3.6	5.5
	1	3	1		8.5	**9.5**	8.8
Test	1	1	1		-1.7	5.3	**8.0**
	1	2	1		5.7	3.0	**7.9**
	1	3	1		7.2	**7.9**	6.4
Val	10	1	1		-10.0	4.1	8.1
	10	2	1		1.0	4.8	10.1
	10	3	1		2.6	9.1	**10.8**
Test	10	1	1		-10.4	4.6	8.5
	10	2	1		1.9	5.6	10.3
	10	3	1		3.2	9.5	**11.8**

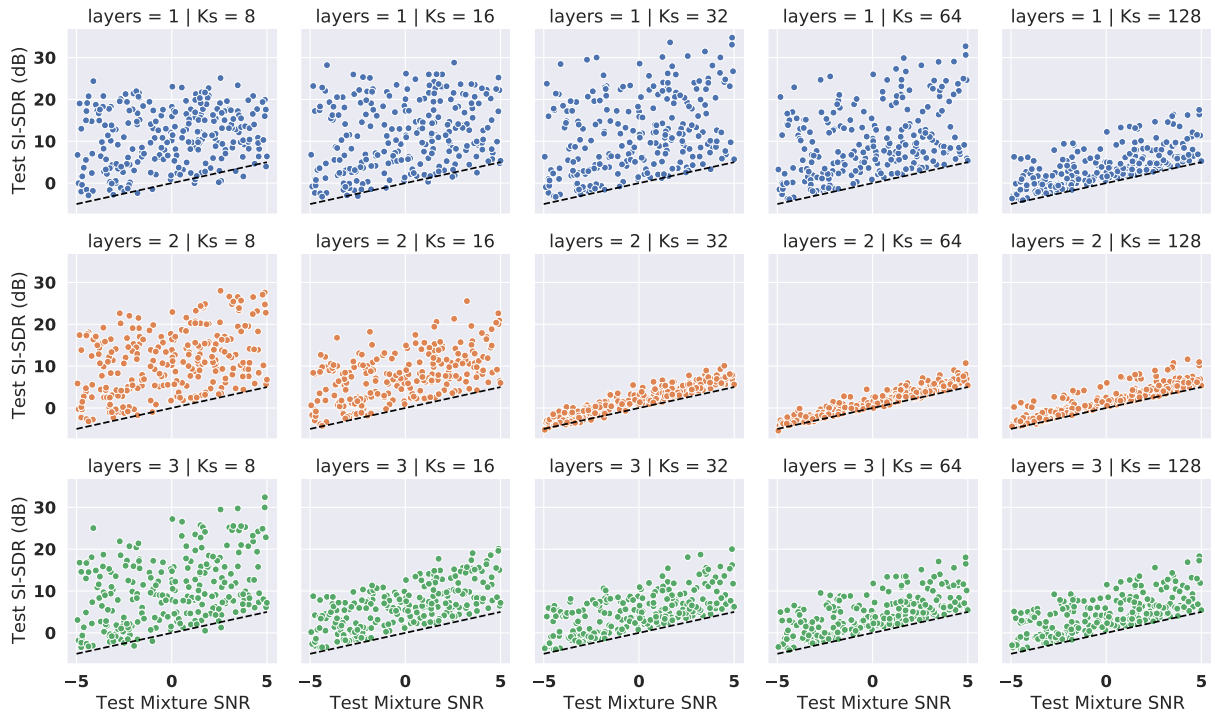
Πίνακας 5.19: Απόδοση της μεθόδου NAE με σφάλμα στο πεδίο της συχνότητας στο TIMIT-DEMAND.

μοντέλο θορύβου. Το σχήμα αυτό προσφέρει μια καλύτερη επισκόπηση των αποτελεσμάτων του Πίνακα 5.18. Παρατηρούμε ότι, η απόδοση βρίσκεται σχεδόν πάντα πάνω από την διακεκομμένη γραμμή που σημαίνει ότι το εκτιμώμενο σήμα φωνής αποτελεί βελτίωση σε σχέση με το μείγμα. Εμφανής είναι επίσης η μειωμένη απόδοση για την περίπτωση των δυο επιπέδων και $Ks = 32, 64, 128$.

Ks	layers	Noise Type	Test Mean SI-SDR (dB) Domestic	Nature	Office	Public	Street	Transportation
16	1		11.1	9.9	12.3	10.7	10.7	**12.6**
16	2		5.4	4.0	**6.2**	4.9	4.0	5.9
16	3		5.1	3.9	5.5	4.6	3.9	**5.9**
32	1		12.0	9.9	12.5	10.8	10.7	**13.7**
32	2		2.0	1.0	**2.4**	1.6	0.5	1.9
32	3		5.2	3.9	5.5	4.6	4.0	**6.2**

Πίνακας 5.20: Μέση απόδοση μοντέλων ανά κατηγορία θορύβων στο σύνολο εξέτασης του TIMIT-DEMAND για $Ks = 16, 32$, $Kn = 10$, ένα έως τρία επίπεδα και ένα επίπεδο θορύβου (Από Πίνακα 5.18).

Τέλος, επιλέγουμε από τον Πίνακα 5.18 τα μοντέλα με $Ks = 16, 32$, $Kn = 10$ και με ένα επίπεδο θορύβου με σκοπό να εξετάσουμε την μέση απόδοσή τους ανά κατηγορία θορύβου στο σύνολο εξέτασης του TIMIT-DEMAND. Στον Πίνακα 5.20 έχουμε τα αποτελέσματα αυτά.

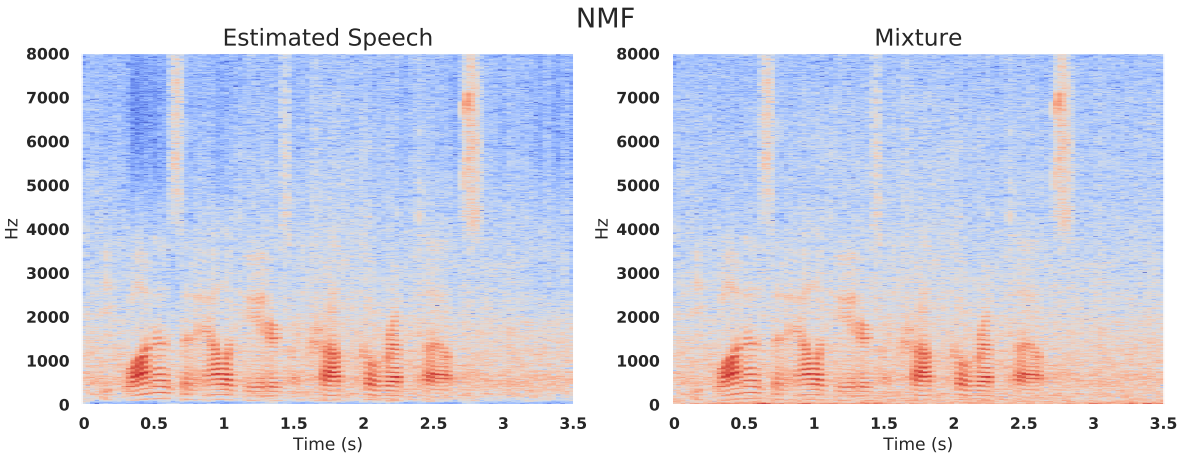


Σχήμα 5.6.2: Απόδοση μοντέλου και SNR μείγματος για κάθε δείγμα στο σύνολο εξέτασης του TIMIT-DEMAND των καλύτερων μοντέλων του Πίνακα 5.18. Η διακεκομμένη γραμμή έχει κλίση 1 και διέρχεται από το $(0, 0)$

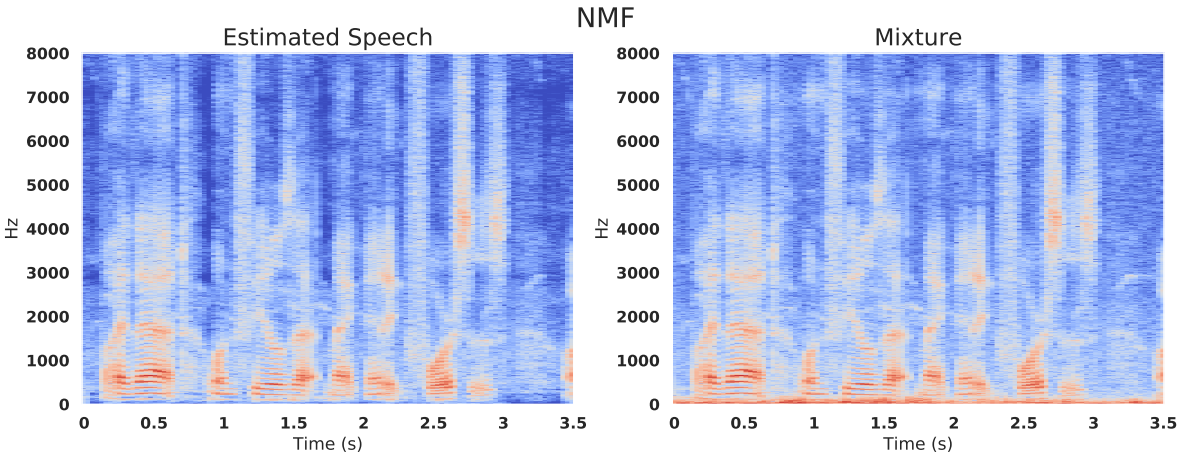
5.7 Ενδεικτικά Φασματογραφήματα

Σε αυτή την ενότητα παρουσιάζουμε μερικά ενδεικτικά φασματογραφήματα από τα εκτιμώμενα σήματα φωνής και τα αντίστοιχα θορυβώδη σήματα, χρησιμοποιώντας τα καλύτερα μοντέλα από κάθε κατηγορία. Πρώτα, στα Σχήματα 5.7.1 και 5.7.2 χρησιμοποιούμε το μοντέλο NMF με $Ks = 16$ και $Kn = 1$. Έπειτα, στα Σχήματα 5.7.3 και 5.7.4 έχουμε το μοντέλο NAE με σφάλμα στο πεδίο της συχνότητας με $Ks = 32$, $Kn = 10$ και από ένα επίπεδο σε κάθε αποκωδικοποιητή. Τέλος, στα Σχήματα 5.7.5 και 5.7.6 έχουμε το μοντέλο NAE με σφάλμα στο πεδίο του χρόνου με $Ks = 8$, $Kn = 10$, τρία επίπεδα στον αποκωδικοποιητή ομιλίας και ένα στον αποκωδικοποιητή θορύβου.

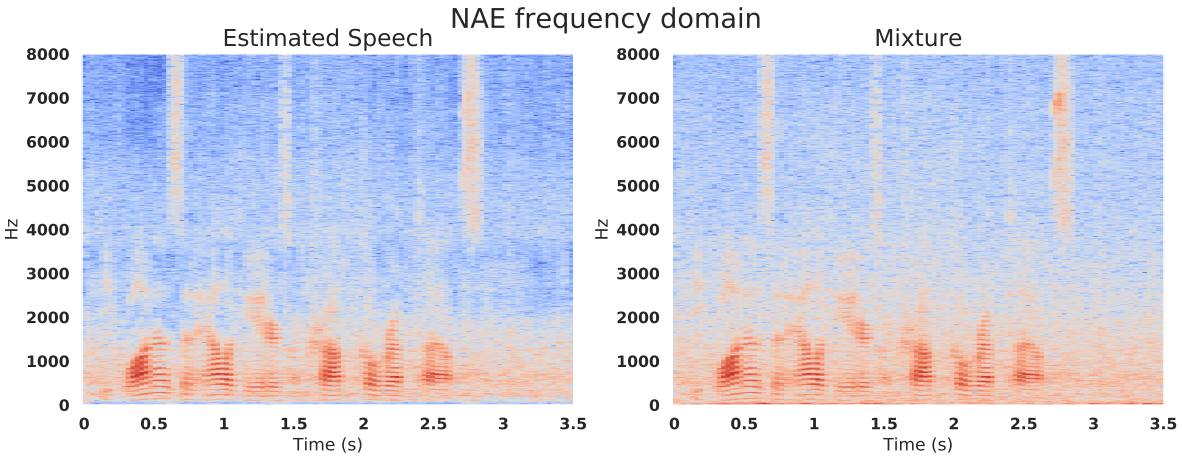
Συγκρίνοντας τα επιλεγμένα μοντέλα σε δυο θορυβώδη σήματα κάνουμε τις παρακάτω παρατηρήσεις. Αρχικά, στο πρώτο θορυβώδες σήμα, το σήμα θορύβου έχει μεγάλο μέρος της ενέργειάς του συγκεντρωμένο σε σχετικά χαμηλές συχνότητες, στο ίδιο εύρος με το σήμα φωνής. Αντίθετα, στο δεύτερο θορυβώδες σήμα, έχουμε συγκέντρωση της ενέργειας του σήματος θορύβου σε πολύ χαμηλές συχνότητες, οπότε δεν έχουμε μεγάλη επικάλυψη με το σήμα φωνής. Επομένως, όπως περιμέναμε παρατηρούμε ότι όλα τα μοντέλα αποδίδουν καλύτερα στο δεύτερο θορυβώδες σήμα. Ακόμη, βλέπουμε ότι τα μοντέλα και στα δυο θορυβώδη σήματα αφαιρούν την ενέργεια από τις πολύ χαμηλές συχνότητες. Την καλύτερη απόδοση στο πρώτο σήμα την επιτυγχάνει το μοντέλο NAE με σφάλμα στο πεδίο του χρόνου και πράγματι παρατηρούμε στο φασματογράφημα του εκτιμώμενου σήματος ότι έχουμε καλύτερη αφαίρεση του θορύβου σε σχέση με τις υπόλοιπες περιπτώσεις. Για την περίπτωση του δεύτερου σήματος, το μοντέλο NMF επιτυγχάνει την καλύτερη απόδοση. Πράγματι, συγκρίνοντας με τα υπόλοιπα μοντέλα, βλέπουμε ότι το μοντέλο NAE με σφάλμα στο πεδίο της συχνότητας αφαιρεί παραπάνω ενέργεια από όσο χρειάζεται στο εύρος συχνοτήτων 2000 Hz με 4000 Hz, ενώ το μοντέλο NAE με σφάλμα στο πεδίο του χρόνου δεν αφαιρεί τόσο καλά την ενέργεια από τις πολύ χαμηλές συχνότητες.



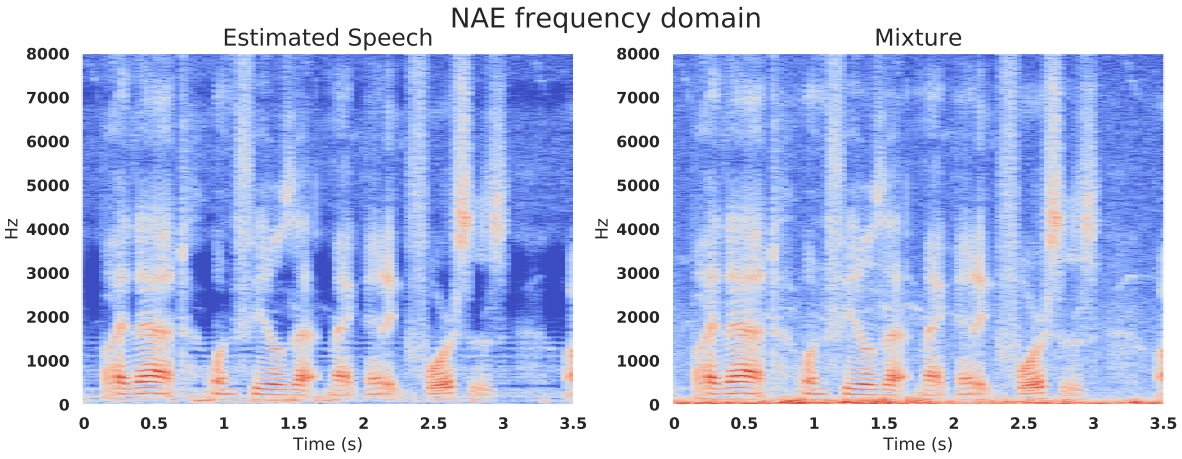
Σχήμα 5.7.1: Φασματογράφημα εκτιμώμενου σήματος φωνής με NMF στα αριστερά και θορυβώδους σήματος στα δεξιά.



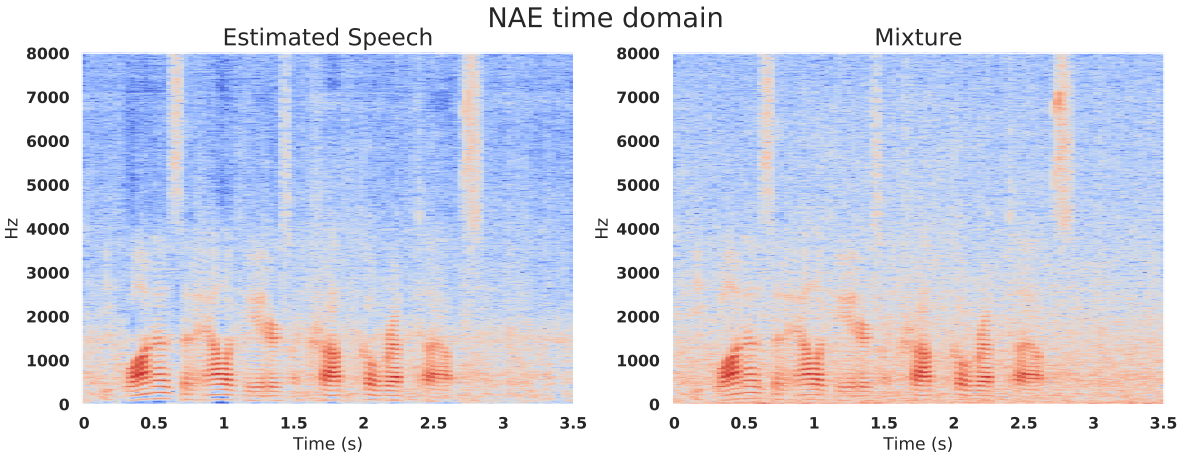
Σχήμα 5.7.2: Φασματογράφημα εκτιμώμενου σήματος φωνής με NMF στα αριστερά και θορυβώδους σήματος στα δεξιά.



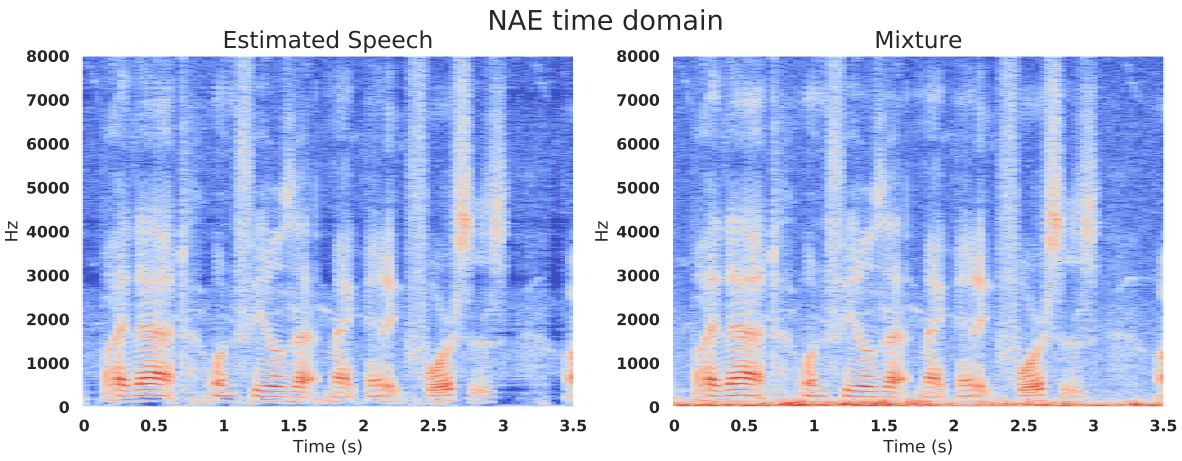
Σχήμα 5.7.3: Φασματογράφημα εκτιμώμενου σήματος φωνής με NAE, με σφάλμα στο πεδίο της συχνότητας, στα αριστερά και θορυβώδους σήματος στα δεξιά.



Σχήμα 5.7.4: Φασματογράφημα εκτιμώμενου σήματος φωνής με NAE, με σφάλμα στο πεδίο της συχνότητας, στα αριστερά και θορυβώδους σήματος στα δεξιά.



Σχήμα 5.7.5: Φασματογράφημα εκτιμώμενου σήματος φωνής με NAE, με σφάλμα στο πεδίο του χρόνου, στα αριστερά και θορυβώδους σήματος στα δεξιά.



Σχήμα 5.7.6: Φασματογράφημα εκτιμώμενου σήματος φωνής με NAE, με σφάλμα στο πεδίο του χρόνου, στα αριστερά και θορυβώδους σήματος στα δεξιά.

5.8 Συζήτηση Αποτελεσμάτων

5.8.1 Εκπαίδευση σε Σήματα Ομιλίας

Αρχικά συγκρίνουμε την απόδοση ανακατασκευής "καθαρής" ομιλίας από τον Πίνακα 5.1 για τα μοντέλα NMF, τους Πίνακες 5.6 και 5.7 για τα μοντέλα NAE με σφάλμα στο πεδίο της συχνότητας και τους Πίνακες 5.8, 5.9 και 5.10 για τα μοντέλα NAE με σφάλμα στο πεδίο του χρόνου.

- Γενικά βλέπουμε ότι όσο αυξάνει η τάξη Ks των μοντέλων αυξάνεται και η απόδοση.
- Ακόμη, τα μοντέλα NAE με ένα επίπεδο, εκτός από κάποιες εξαιρέσεις για πολύ μεγάλα ή πολύ μικρά Ks, ξεπερνάνε την απόδοση της NMF, το οποίο είναι αναμενόμενο αφού οι βάσεις των μοντέλων NAE δεν περιορίζονται από την μη αρνητικότητα.
- Επίσης, όσο αυξάνουμε το βάθος των μοντέλων NAE έχουμε βελτίωση σε σχέση με την περίπτωση του ενός επιπέδου. Όμως, παρατηρούμε πτώση της απόδοσης από τα δυο επίπεδα σε σχέση με τα τρία επίπεδα στην περίπτωση των μοντέλων NAE με σφάλμα στο πεδίο του χρόνου.
- Η χρήση πολώσεων στα μοντέλα NAE με σφάλμα στο πεδίο του χρόνου δεν έχει μεγάλη επίδραση στη συμπεριφορά τους.
- Τέλος, η εκπαίδευση των μοντέλων NAE με μικρές τάξεις μοντέλου $Ks = 1, 2, 4$ πολλές φορές αποτυγχάνει ιδιαίτερα στην περίπτωση του σφάλματος στο πεδίο του χρόνου.

Συγκρίνοντας τις καμπύλες σφάλματος στο σύνολο εκπαίδευσης ανάμεσα στα μοντέλα NAE με σφάλμα στο πεδίο της συχνότητας και χρόνου, παρατηρούμε ότι στην πρώτη κατηγορία μοντέλων έχουμε πιο ομαλές καμπύλες και συνολικά καλύτερη συμπεριφορά κατά την εκπαίδευση. Αντίθετα, για την περίπτωση του σφάλματος στο πεδίο του χρόνου έχουμε αυξημένη διασπορά ανάμεσά τους και ορισμένες φορές απότομες μεταβολές στο σφάλμα. Πιθανώς, για αυτή την συμπεριφορά να οφείλεται σε κάποιο βαθμό ο αυξημένος ρυθμός εκμάθησης του αλγορίθμου Adam.

Εναλλακτικά, θα μπορούσαμε να είχαμε χρησιμοποιήσει την τεχνική της πρόωρης διακοπής της εκπαίδευσης (early stopping), όπου με βάση το σφάλμα στο σύνολο επαλήθευσης διακόπτουμε την εκπαίδευση ώστε να πάρουμε το βέλτιστο μοντέλο. Πιθανώς έτσι να αποφεύγαμε μοντέλα τα οποία έχουν υποστεί απότομες μεταβολές στο σφάλμα εκπαίδευσης.

5.8.2 Πειράματα Διαχωρισμού

Εξετάζοντας τα αποτελέσματα του διαχωρισμού συνολικά, παρατηρούμε ότι οι μέθοδοι NAE φτάνουν την απόδοση της NMF στο TIMIT-DEMAND αλλά στο TIMIT-MUSDB υστερούν σε σχέση με αυτήν. Συμπεραίνουμε έτσι ότι η ικανότητα ανακατασκευής σημάτων ομιλίας δεν σχετίζεται τόσο άμεσα με την απόδοση στον διαχωρισμό. Για παράδειγμα, τα μοντέλα NAE με δυο ή τρία επίπεδα και σφάλμα στο πεδίο της συχνότητας έχουν χειρότερη απόδοση στον διαχωρισμό σε σχέση με τα αντίστοιχα μοντέλα NAE με σφάλμα στο πεδίο του χρόνου, παρόλο που τα δεύτερα δεν έχουν τόσο καλή απόδοση στην ανακατασκευή. Συνεπώς, τα μοντέλα NAE δεν χρησιμοποιούν το βάθος τους στο έπακρο. Πιθανώς αυτό να οφείλεται στην αυξημένη τους εκφραστικότητα και ως αποτέλεσμα να μπορούν να περιγράφουν και μέρος του σήματος θορύβου. Η υπόθεση αυτή υποστηρίζεται από το γεγονός ότι έχουμε αύξηση της απόδοση όταν ο αποκωδικοποιητής θορύβου έχει τον ίδιο αριθμό επιπέδων με τον αποκωδικοποιητή φωνής, δηλαδή όταν αυξάνουμε την εκφραστικότητα του αποκωδικοποιητή θορύβου. Ένας πιθανός τρόπος αντιμετώπισης είναι η χρήση περιορισμού αραιότητας για τις ενδιάμεσες αναπαραστάσεις τόσο κατά την εκπαίδευση αλλά και κατά τον διαχωρισμό.

Στο (Leglaive, Girin, and Horaud 2018) το οποίο αξιολογεί το προτεινόμενο VAE μοντέλο στο TIMIT-DEMAND με μείγματα με SNR 0 dB, μέγιστη απόδοση επιτυγχάνεται για τάξη $Ks = 16, 32, 64$ η οποία φτάνει στα 14 dB SDR ξεπερνώντας την NMF. Παρόλο που δεν έχουμε ακριβή σύγκριση, μπορούμε να πούμε με αρκετή βεβαιότητα ότι η προτεινόμενη μέθοδος με VAE ξεπερνά την απόδοση των μοντέλων NAE. Στην περίπτωση των VAEs το γεγονός ότι εκπαιδεύονται ώστε οι ενδιάμεσες κρυφές μεταβλητές να ακολουθούν

μια προκαθορισμένη κατανομή, πιθανώς συμβάλει στην αντιμετώπιση του προβλήματος της εκφραστικότητας. Τέλος, εμπειρικά παρατηρήσαμε ότι ο προτεινόμενος επαναληπτικός αλγόριθμος διαχωρισμού με VAE απαιτεί περισσότερο χρόνο από αυτόν που προτείναμε με NAE.

Στο (Venkataramani 2020) εξετάζεται ο συνδυασμός των NAE με ευθύ μετασχηματισμό που λαμβάνει είσοδο στο πεδίο του χρόνου και δίνει ως έξοδο δισδιάστατη αναπαράσταση και τον αντίστοιχο αντίστροφο, που μαθαίνονται κατά την εκπαίδευση. Με βάση αυτό, μπορούμε τροποποιήσουμε την περίπτωση με σφάλμα στο πεδίο του χρόνου ώστε να αντικαταστήσουμε τον STFT με τον μετασχηματισμό αυτόν, μαθαίνοντας τον κατά την εκπαίδευση. Κατά τον διαχωρισμό για το μοντέλο του θορύβου μπορούμε να κρατήσουμε τον αντίστροφο STFT, ενώ για το μοντέλο φωνής κρατάμε τον αντίστροφο μετασχηματισμό που μάθαμε. Με τον τρόπο αυτό πιθανώς να περιορίσουμε την ικανότητα του αποκωδικοποιητή ομιλίας να περιγράφει σήματα θορύβου.

Τέλος, εξετάζοντας τους Πίνακες 5.5, 5.20 και 5.17 παρατηρούμε ότι η αποδόσεις των μοντέλων ανά τύπο θορύβων είναι αρκετά παρόμοιες. Για παράδειγμα, τα περισσότερα μοντέλα επιτυγχάνουν μέγιστη απόδοση στην κατηγορία Transportation. Μια εξήγηση είναι ότι όλες οι κατηγορίες θορύβου δεν διαφέρουν σε μεγάλο βαθμό ως προς την δυσκολία αλλά παίρνουμε αυτά τα αποτελέσματα λόγω του τρόπου που έχουν κατανεμηθεί τα δύσκολα δείγματα στο σύνολο εξέτασης που κατασκευάσαμε.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές Επεκτάσεις

6.1 Σύνοψη και Συμπεράσματα

Στην παρούσα διπλωματική εργασία εξετάσαμε το πρόβλημα της Αποθορυβοποίησης Σήματος Φωνής μέσω του Διαχωρισμού Πηγών. Θεωρήσαμε το πρόβλημα στην ημι-επιβλεπόμενη περίπτωση όπου τα δεδομένα εκπαίδευσης αποτελούνται από "καθαρά" σήματα ομιλίας, ενώ κατά την αξιολόγηση οι θόρυβοι είναι άγνωστοι. Θέσαμε το πρόβλημα σε αυτή την μορφή ώστε η μέθοδος επίλυσης που αναπτύσσουμε να μην υποφέρει από προβλήματα γενίκευσης ως προς το είδος και το περιβάλλον θορύβου.

Η μελέτη μας ξεκίνησε με την παρουσίαση του σχετικού θεωρητικού υποβάθρου σε θέματα Ψηφιακής Επεξεργασίας Σήματος, Μηχανικής Μάθησης και Νευρωνικών Δικτύων. Στη συνέχεια, πραγματοποιήσαμε μια βιβλιογραφική επισκόπηση για το γενικό πρόβλημα της Αποθορυβοποίησης Σήματος Φωνής, για το πρόβλημα του Διαχωρισμού Πηγών σε ηχητικά σήματα καθώς και για ημι-επιβλεπόμενες μεθόδους για το πρόβλημα Αποθορυβοποίησης Σήματος Φωνής μέσω του Διαχωρισμού Πηγών. Εστιάσαμε στις μεθόδους NAE και τις παλαιότερες μεθόδους NMF, τις οποίες μελετήσαμε διεξοδικά. Με βάση την ημι-επιβλεπόμενη μεθοδολογία με NMF για το πρόβλημα και παλαιότερη έρευνα για τα μοντέλα NAE σχεδιάσαμε και προτείναμε ημι-επιβλεπόμενη μεθοδολογία με NAE.

Συγκεκριμένα, η μεθοδολογία που προτείναμε περιλαμβάνει δυο στάδια. Στο πρώτο στάδιο εκπαιδεύουμε ένα μοντέλο NAE σε "καθαρά" σήματα ομιλίας με στόχο την ανακατασκευή τους μέσω μιας ενδιάμεσης αναπαράστασης μικρότερης διαστατικότητας. Στη συνέχεια, συνδυάζουμε τον αποκωδικοποιητή ομιλίας του εκπαιδευμένου μοντέλου με έναν τυχαία αρχικοποιημένο αποκωδικοποιητή θορύβου για τον διαχωρισμό, κατά τον οποίον προσαρμόζουμε κατάλληλα, μέσω ενός επαναληπτικού αλγορίθμου, τις παραμέτρους του αποκωδικοποιητή θορύβου καθώς και τις εισόδους των δυο αποκωδικοποιητών.

Στο πειραματικό μέρος της εργασίας, πρώτα εκπαιδεύσαμε μοντέλα NMF με διαφορετικούς αριθμούς βάσεων και μοντέλα NAE με διαφορετικούς αριθμούς τάξης και ένα ή περισσότερα επίπεδα, σε "καθαρά" σήματα ομιλίας. Έπειτα, συγκρίναμε τα μοντέλα αυτά στο πρόβλημα. Συγκεκριμένα, δοκιμάσαμε την ημι-επιβλεπόμενη μέθοδο NMF σε δυο σύνολα δεδομένων που καλύπτουν ένα μεγάλο εύρος τύπων θορύβου, με μεταβαλλόμενα επίπεδα θορύβου. Πραγματοποιήσαμε τροποποιήσεις στην ημι-επιβλεπόμενη μέθοδο NMF που είχαν ως αποτέλεσμα αύξηση της απόδοσης σε ορισμένες περιπτώσεις, αλλά με αυξημένο υπολογιστικό κόστος. Στη συνέχεια, πραγματοποιήσαμε πειράματα ώστε να ρυθμίσουμε την ημι-επιβλεπόμενη μέθοδο NAE, καταλήγοντας σε ένα συνδυασμό από ρυθμίσεις οι οποίες μεγιστοποιούν την απόδοση. Δοκιμάσαμε τελικά τις διάφορες τροποποιήσεις των μοντέλων NAE στα δυο σύνολα δεδομένων.

Μια από της αρχικές παρατηρήσεις μας είναι ότι η εκπαίδευση μοντέλων NAE με σφάλμα στο πεδίο της συχνότητας πετυχαίνει καλύτερη απόδοση στην ανακατασκευή σημάτων φωνής και εμφανίζει πιο ομαλή συμπεριφορά στην εκπαίδευση, σε σχέση με την εκπαίδευση μοντέλων NAE με σφάλμα στο πεδίο του χρόνου. Προσαρμόζοντας την ημι-επιβλεπόμενη μέθοδο διαχωρισμού με το μοντέλο NAE παρατηρούμε ότι όταν εργαζόμαστε με την ρίζα του μέτρου του STFT αυξάνεται σημαντικά η απόδοση στον διαχωρισμό. Καταφέραμε λοιπόν να ρυθμίσουμε την μέθοδο NAE ώστε να λειτουργεί ικανοποιητικά στο πρόβλημα του διαχωρισμού και να φτάνει την απόδοση της NMF στο σύνολο δεδομένων TIMIT-DEMAND. Όμως, στο δεύτερο σύνολο δεδομένων, το TIMIT-MUSDB, η απόδοση της προτεινόμενης μεθόδου υστερεί σε σχέση με την NMF. Εξετάζοντας τα αποτελέσματα της μεθόδου NAE παρατηρούμε ότι ανεξάρτητα το SNR του μείγματος στις περισσότερες περιπτώσεις καταφέρνει να δώσει βελτίωση σε σχέση με το μείγμα. Τέλος, συμπεραίνουμε ότι η απόδοση στο πρόβλημα του διαχωρισμού δεν εξαρτάται άμεσα από την απόδοση στην ανακατασκευή σημάτων ομιλίας.

6.2 Μελλοντικές Επεκτάσεις

Με αφετηρία τα αποτελέσματα της ημι-επιβλεπόμενης μεθόδου NAE που προτείνουμε για το πρόβλημα, ορισμένες επεκτάσεις αξίζει να μελετηθούν μελλοντικά.

Αρχικά, παρατηρήσαμε την μέθοδο NAE να επιτυγχάνει καλύτερη απόδοση ανακατασκευής σημάτων ομιλίας σε σχέση με την μέθοδο NMF αλλά να υστερεί σε σχέση με αυτή στο πρόβλημα του διαχωρισμού. Συνεπώς, θα πρέπει να εξεταστούν τρόποι ώστε να περιοριστεί η εκφραστικότητα των μοντέλων NAE ώστε να περιγράφουν μόνο σήματα φωνής. Ένας πιθανός τρόπος είναι η χρήση περιορισμού αραιότητας για τις ενδιάμεσες αναπαραστάσεις τόσο κατά την εκπαίδευση αλλά και κατά τον διαχωρισμό. Όπως είδαμε στη βιβλιογραφική επισκόπηση ο περιορισμός αραιότητας έχει ευρεία χρήση σε μεθόδους NMF. Επίσης, ένας άλλος τρόπος αντιμετώπισης του προβλήματος πιθανώς να είναι η αντικατάσταση των γραμμικών επιπέδων των μοντέλων NAE με συνελικτικά ή αναδρομικά επίπεδα. Σκοπός στη συγκεκριμένη περίπτωση είναι η ενσωμάτωση των χρονικών χαρακτηριστικών των σημάτων φωνής, ώστε να αυξηθεί η εκφραστικότητα για την κατηγορία σημάτων φωνής και να μειωθεί για τις υπόλοιπες κατηγορίες σημάτων.

Τέλος, είδαμε ότι η ενσωμάτωση του μετασχηματισμού Fourier βραχέως χρόνου στο μοντέλο NAE και ο υπολογισμός του σφάλματος στο πεδίο του χρόνου, οδηγεί σε λιγότερο ομαλή συμπεριφορά κατά την εκπαίδευση σε σήματα φωνής και δεν δίνει πολύ υψηλότερη απόδοση στον διαχωρισμό σε σχέση με την περίπτωση που το σφάλμα υπολογίζεται στο πεδίο της συχνότητας. Η χρήση μετασχηματισμού που λαμβάνει είσοδο στο πεδίο του χρόνου και δίνει ως έξοδο δισδιάστατη αναπαράσταση, ο οποίος μαθαίνεται κατά την εκπαίδευση (Venkataramani, Casebeer, and Smaragdis 2017), πιθανώς να λύσει το παραπάνω πρόβλημα αλλά και αυτό της εκφραστικότητας.

Παράρτημα Α

Βιβλιογραφία

Loizou, P. C. (2007). *Speech enhancement: theory and practice*. CRC press.

Oppenheim, A. and Schafer, R. (2010). *Discrete time signal processing*. 3rd. Prentice Hall Press.

Rabiner, L. and Schafer, R. (2010). *Theory and applications of digital speech processing*. Prentice Hall Press.

Smith, J. O. (2011). *Spectral audio signal processing*. W3K.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Hornik, K., Stinchcombe, M., and White, H. (1989). "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366.

Leshno, M. et al. (1993). "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". In: *Neural networks* 6.6, pp. 861–867.

Kingma, D. P. and Ba, J. (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Lee, D. and Seung, H. S. (1999). "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755, pp. 788–791.

Smaragdis, P. and Brown, J. C. (2003). "Non-negative matrix factorization for polyphonic music transcription". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE.

Févotte, C. and Idier, J. (2011). "Algorithms for nonnegative matrix factorization with the β-divergence". In: *Neural computation* 23.9, pp. 2421–2456.

Cichocki, A. et al. (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons.

Lee, D. and Seung, H. S. (2000). "Algorithms for non-negative matrix factorization". In: *Advances in Neural Information Processing Systems* 13.

Virtanen, T. (2007). "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3, pp. 1066–1074.

Boll, S. (1979). "Suppression of acoustic noise in speech using spectral subtraction". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27.2, pp. 113–120.

Lim, J. and Oppenheim, A. V. (1979). "Enhancement and bandwidth compression of noisy speech". In: *Proceedings of the IEEE* 67.12, pp. 1586–1604.

Ephraim, Y. and Malah, D. (1984). "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.6, pp. 1109–1121.

Ephraim, Y. and Van Trees, H. L. (1995). "A signal subspace approach for speech enhancement". In: *IEEE Transactions on Speech and Audio Processing* 3.4, pp. 251–266.

Wang, D. and Chen, J. (2018). "Supervised speech separation based on deep learning: An overview". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10, pp. 1702–1726.

Wang, D. (2005). "On ideal binary mask as the computational goal of auditory scene analysis". In: *Speech separation by humans and machines*. Springer, pp. 181–197.

Lu, X. et al. (2013). "Speech enhancement based on deep denoising autoencoder." In: *Interspeech*. Vol. 2013.

Maas, A. et al. (2012). "Recurrent Neural Networks for Noise Reduction in Robust ASR". In: *Interspeech*.

Weninger, F., Erdogan, H., et al. (2015). "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR". In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer.

Hochreiter, S. and Schmidhuber, J. (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Fu, S.-W., Tsao, Y., et al. (2017). "Raw waveform-based speech enhancement by fully convolutional networks". In: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE.

Pandey, A. and Wang, D. (2019). "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Rethage, D., Pons, J., and Serra, X. (2018). "A wavenet for speech denoising". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Oord, A. v. d. et al. (2016). "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499*.

Goodfellow, I., Pouget-Abadie, J., et al. (2014). "Generative adversarial nets". In: *Advances in Neural Information Processing Systems* 27.

Pascual, S., Bonafonte, A., and Serrà, J. (2017). "SEGAN: Speech Enhancement Generative Adversarial Network". In: *Interspeech*.

Fu, S.-W., Liao, C.-F., et al. (2019). "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement". In: *International Conference on Machine Learning*. PMLR.

Zhang, Q. et al. (2020). "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, pp. 1404–1415.

Defossez, A., Synnaeve, G., and Adi, Y. (2020). "Real time speech enhancement in the waveform domain". In: *arXiv preprint arXiv:2006.12847*.

Fujimura, T. et al. (2021). "Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech". In: *European Signal Processing Conference (EUSIPCO)*. IEEE.

Fu, S.-W., Yu, C., et al. (2022). "MetricGAN-U: Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Xiang, Y. and Bao, C. (2020). "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, pp. 1826–1838.

Hyvärinen, A. and Oja, E. (2000). "Independent component analysis: algorithms and applications". In: *Neural networks* 13.4-5, pp. 411–430.

Smaragdis, P. (1998). "Blind separation of convolved mixtures in the frequency domain". In: *Neurocomputing* 22.1-3, pp. 21–34.

Greenberg, J. (2007). "Blind Source Separation: Principal & Independent Component Analysis". In: *MIT OpenCourseWare*.

Casey, M. A. and Westner, A. (2000). "Separation of mixed audio sources by independent subspace analysis". In: *International Computer Music Conference (ICMC)*.

Hu, G. and Wang, D. (2010). "A tandem algorithm for pitch estimation and voiced speech segregation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.8, pp. 2067–2079.

Kristjansson, T., Attias, H., and Hershey, J. R. (2004). "Single microphone source separation using high resolution signal reconstruction". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE.

Roweis, S. (2000). "One microphone source separation". In: *Advances in Neural Information Processing Systems* 13.

Smaragdis, P., Fevotte, C., et al. (2014). "Static and dynamic source separation using nonnegative factorizations: A unified view". In: *IEEE Signal Processing Magazine* 31.3, pp. 66–75.

Duong, N. Q., Ozerov, A., and Chevallier, L. (2014). "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization". In: *IEEE International Conference on Consumer Electronics Berlin (ICCE)*. IEEE.

Mohammadiha, N., Smaragdis, P., and Leijon, A. (2013). "Supervised and unsupervised speech enhancement using nonnegative matrix factorization". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.10, pp. 2140–2151.

Le Roux, J., Weninger, F., and Hershey, J. R. (Mar. 2015). *Sparse NMF – half-baked or well done?* Tech. rep. TR2015-023. Cambridge, MA, USA: Mitsubishi Electric Research Laboratories (MERL).

Smaragdis, P. (2006). "Convolutive speech bases and their application to supervised speech separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.1, pp. 1–12.

Mysore, G. J., Smaragdis, P., and Raj, B. (2010). "Non-negative hidden Markov modeling of audio with application to source separation". In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer.

Le Roux, J., Hershey, J. R., and Weninger, F. (2015). "Deep NMF for speech separation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Smaragdis, P. and Venkataramani, S. (2017). "A neural network alternative to non-negative audio models". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Grais, E. M., Sen, M. U., and Erdogan, H. (2014). "Deep neural networks for single channel source separation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Wang, Y. and Wang, D. (2013). "Towards scaling up classification-based speech separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.7, pp. 1381–1390.

Weninger, F., Hershey, J. R., et al. (2014). "Discriminatively trained recurrent neural networks for single-channel speech separation". In: *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE.

Huang, P.-S. et al. (2015). "Joint optimization of masks and deep recurrent neural networks for monaural source separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.12, pp. 2136–2147.

Hershey, J. R. et al. (2016). "Deep clustering: Discriminative embeddings for segmentation and separation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Yu, D. et al. (2017). "Permutation invariant training of deep models for speaker-independent multi-talker speech separation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Luo, Y. and Mesgarani, N. (2018). "Tasnet: time-domain audio separation network for real-time, single-channel speech separation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Luo, Y., Chen, Z., and Yoshioka, T. (2020). "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Luo, Y. and Mesgarani, N. (2019). "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation". In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8, pp. 1256–1266.

Lea, C. et al. (2016). "Temporal convolutional networks: A unified approach to action segmentation". In: *European Conference on Computer Vision*. Springer.

Venkataramani, S., Casebeer, J., and Smaragdis, P. (2018). "End-to-end source separation with adaptive front-ends". In: *52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE.

Stoller, D., Ewert, S., and Dixon, S. (2018). "Wave-u-net: A multi-scale neural network for end-to-end audio source separation". In: *arXiv preprint arXiv:1806.03185*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer.

Wisdom, S. et al. (2020). "Unsupervised Speech Separation Using Mixtures of Mixtures". In: *ICML 2020 Workshop on Self-Supervision for Audio and Speech*.

Smaragdis, P., Raj, B., and Shashanka, M. (2007). "Supervised and semi-supervised separation of sounds from single-channel mixtures". In: *International Conference on Independent Component Analysis and Signal Separation*. Springer.

Mysore, G. J. and Smaragdis, P. (2011). "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Bando, Y. et al. (2018). "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Leglaive, S., Girin, L., and Horaud, R. (2018). "A variance modeling framework based on variational autoencoders for speech enhancement". In: *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE.

Pariente, M., Deleforge, A., and Vincent, E. (2019). "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders". In: *arXiv preprint arXiv:1905.01209*.

Kingma, D. P. and Welling, M. (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Leglaive, S., Alameda-Pineda, X., et al. (2020). "A recurrent variational autoencoder for speech enhancement". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Girin, L. et al. (2021). "Dynamical Variational Autoencoders: A Comprehensive Review". In: *Foundations and Trends® in Machine Learning* 15.1-2, pp. 1–175. ISSN: 1935-8237. DOI: 10.1561/2200000089.

Nugraha, A. A., Sekiguchi, K., and Yoshii, K. (2020). "A flow-based deep latent variable model for speech spectrogram modeling and enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, pp. 1104–1117.

Taal, C. H. et al. (2011). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7, pp. 2125–2136.

Papamakarios, G. et al. (2021). "Normalizing Flows for Probabilistic Modeling and Inference". In: *Journal of Machine Learning Research* 22, pp. 1–64.

Leglaive, S., Girin, L., and Horaud, R. (2019). "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Sekiguchi, K. et al. (2018). "Bayesian multichannel speech enhancement with a deep speech prior". In: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE.

Wang, D. and Lim, J. (1982). "The unimportance of phase in speech enhancement". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 30.4, pp. 679–681.

Vincent, E., Gribonval, R., and Févotte, C. (2006). "Performance measurement in blind audio source separation". In: *IEEE transactions on audio, speech, and language processing* 14.4, pp. 1462–1469.

Rix, A. W. et al. (2001). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Le Roux, J., Wisdom, S., et al. (2019). "SDR–half-baked or well done?" In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Garofolo, J. S. et al. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. DOI: https://doi.org/10.35111/17gk-bn40.

Thiemann, J., Ito, N., and Vincent, E. (June 2013). *DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments*. Version 1.0. Supported by Inria under the Associate Team Program VERSAMUS. Zenodo. DOI: 10.5281/zenodo.1227121.

Rafii, Z. et al. (Dec. 2017). *The MUSDB18 corpus for music separation*. DOI: 10.5281/zenodo.1117372.

Venkataramani, S. (2020). "End-to-end non-negative auto-encoders: a deep neural alternative to non-negative audio modeling". PhD thesis. University of Illinois at Urbana-Champaign.

Venkataramani, S., Casebeer, J., and Smaragdis, P. (2017). "Adaptive front-ends for end-to-end source separation". In: *Workshop for Audio Signal Processing, NIPS*.

Παράρτημα Β

Λίστα με Ακρωνύμια

ΨΕΣ Ψηφιακή Επεξεργασία Σήματος

NMF Non Negative Matrix Factorization

NAE Non Negative Autoencoder

STFT Short-Time Fourier Transform

CTFT Continuous Time Fourier Transform

DTFT Discrete Time Fourier Transform

DFT Discrete Fourier Transform

WOLA Weighted Overlap Add

SGD Stochastic Gradient Descent

IBM Ideal Binary Mask

IRM Ideal Ratio Mask

RNN Recurrent Neural Network

MFCC Mel Frequency Cepstral Coefficient

LSTM Long Short Term Memory

FCN Fully Convolutional Network

TCNN Temporal Convolutional Neural Network

GAN Generative Adverserial Network

ICA Independent Component Analysis

ISA Independent Subspace Analysis

CASA Computational Auditory Scene Analysis

HMM Hidden Markov Model

GMM Gaussian Mixture Model

PIT Permutation Invariant Training

TCN Temporal Convolutional Network

PLCA Probabilistic Latent Component Analysis

VAE Variational Autoencoder

MSE Mean Square Error

SDR Signal-to-Distortion Ratio

SI-SDR Scale Invariant Signal-to-Distortion Ratio

SNR Signal-to-Noise Ratio