Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Αποθορυβοποίηση Ανθρώπινης Ομιλίας Με Τεχνικές Βαθιάς Μηχανικής Μάθησης

Τελική Αναφορά

Αμοιρίδης Βασίλειος 8772 - vamoirid@auth.gr

Αναγνώστου Αθανάσιος 8774 - aanagnost@auth.gr

Εμμανουήλ Χρήστος 8804 - eachrist@auth.gr

Τσουκιάς Στέφανος 8936 - tsoukias@auth.gr

2020 - 2021

Περίληψη

Η αναφορά αυτή αποτελεί την προσπάθεια σχεδίασης ενός συστήματος βαθιάς μηχανικής μάθησης, ικανό να βελτιώσει την ποιότητα της ανθρώπινης ομιλίας (speech enhancement) με τεχνικές εξάλειψης θορύβου (noise cancellation). Αρχικά, αναφέρεται στην ανάγκη των σύγχρονων κοινωνιών για τέτοια συστήματα και το κίνητρο της ομάδας για την σχεδίαση ενός. Στην συνέχεια παρουσιάζονται συνοπτικά έννοιες και τεχνικές που εφαρμόζονται σε διαφορετικά στάδια του προβλήματος. Σκοπός είναι η δημιουργία μιας γενικότερης εικόνας για το θέμα και η παρουσίαση του συστήματος, που θα υλοποιήσει η συγκεκριμένη εργασία. Επιπλέον, γίνεται αναφορά στα δεδομένα, στα εργαλεία και τέλος στους τρόπους αξιολογήσεις.

Περιεχόμενα

Πίνακας Εικόνων

1. Εισαγωγή

Η εξάλειψη του περιβάλλοντος ήχου από την ομιλία του ανθρώπου έχει στόχο να μειώσει ή ιδανικά να καταστείλει ανεπιθύμητες ηχητικές διαταραχές. Βρίσκει εφαρμογή στην επεξεργασία ήχου, στην τηλεφωνία και στις διαδικτυακές πλατφόρμες επικοινωνίας [1]. Οι εντατικοποιημένοι ρυθμοί της κοινωνίας συχνά βρίσκουν τον κόσμο αντιμέτωπο με θορυβώδη περιβάλλοντα κατά τη διάρκεια μιας ηχητικής κλήσης είτε σε εσωτερικούς, είτε σε εξωτερικούς χώρους. Ωστόσο, θόρυβος μπορεί να εμφανίζεται και κατά τη διάρκεια μιας ηχογράφησης και να απαιτείται αποθορυβοποίηση στο πλαίσιο της μετεπεξεργασίας των δεδομένων ήχου. Έτσι, στο πλαίσιο του μαθήματος Τεχνολογία του Ήχου και της Εικόνας, η εργασία αυτή θέτει ως στόχο την επίλυση αυτού του προβλήματος.

1.1. Τυπογραφικές Παραδοχές Εγγράφου

Το παρόν έγγραφο έχει αναπτυχθεί με χρήση τη γραμματοσειράς “Arial”, σε μέγεθος 11pt για το βασικό κείμενο, μεγέθη 18pt, 16pt, 13pt, 12pt και 11pt για τον τίτλο, τις εκάστοτε ενότητες και υποενότητες και μέγεθος 9pt για τις περιγραφές των εικόνων. Ο τίτλος άλλα και όλες οι ενότητες και υποενότητες είναι σε αποχρώσεις του μπλε ενώ χρησιμοποιούνται παρενθέσεις στο έγγραφο με σκοπό την επεξήγηση όρων και την χρήση Αγγλικής ορολογίας. Το κείμενο ακολουθεί πλήρη στοίχιση ενώ οι εικόνες στοίχιση στο κέντρο. Σε όλο το έγγραφο χρησιμοποιείται διάστιχο 1.5 γραμμής. Οι εικόνες είναι αριθμημένες, όπως, και οι αντίστοιχες περιγραφές, ενώ οι βιβλιογραφικές αναφορές εμφανίζονται στο έγγραφο βάσει του προτύπου IEEE.

1.2. Δομή του Εγγράφου

Το συγκεκριμένο έγγραφο χωρίζεται σε έξι κεφάλαια. Το παρόν κεφάλαιο είναι εισαγωγικό, παρουσιάζει το πρόβλημα και το κίνητρο της εργασίας, ενώ δίνει και σημαντικές πληροφορίες για το έγγραφο. Στο δεύτερο κεφάλαιο αναφέρονται έννοιες και τεχνικές της βιβλιογραφίας, με στόχο την κατανόηση των παρακάτω κεφαλαίων. Στο τρίτο κεφάλαιο παρουσιάζεται η προσέγγιση που θα αναπτυχθεί, ερωτήματα που θα μελετηθούν, το σύνολο δεδομένων, το μοντέλο μηχανικής μάθησης, οι μετρικές αξιολόγησης καθώς και το σύνολο των εργαλείων. Στο τέταρτο κεφάλαιο παρουσιάζονται τα αποτελέσματα διάφορων πειραμάτων και του τελικού συστήματος, ενώ στο πέμπτο κεφάλαιο παρουσιάζονται κάποιες ιδέες για εξέλιξη του συστήματος και στο τελευταίο οι βιβλιογραφικές αναφορές.

2. Εισαγωγικές Έννοιες και Τεχνικές Αποθορυβοποίησης

Στο κεφάλαιο αυτό παρουσιάζονται διάφορες έννοιες και τεχνικές που βρέθηκαν στην βιβλιογραφία. Αυτές αφορούν τους τρόπους αναπαράστασης του ήχου, τις διάφορες μεθοδολογίες αποθορυβοποίησης αλλά και τις μεθόδους που ακολουθούνται για την αξιολόγηση συστημάτων.

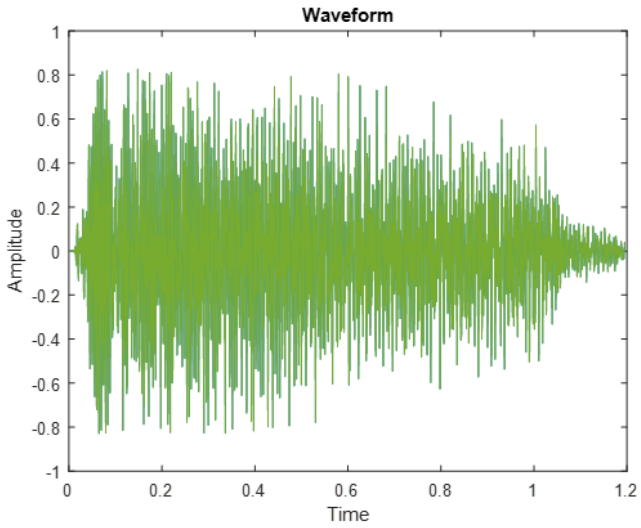
Σκοπός του συγκεκριμένου κεφαλαίου είναι η δημιουργία μιας εικόνας για το θέμα, η κατανόηση εννοιών και μεθοδολογιών που χρησιμοποιούνται σήμερα αλλά και η εξοικείωση του αναγνώστη με έννοιες απαραίτητες για την κατανόηση του εγγράφου.

2.1. Αναπαράσταση Ήχου

Ένα αντικείμενο που δονείται δημιουργεί ήχο. Οι δονήσεις προκαλούν εναλλαγές στην πίεση του αέρα δημιουργώντας ένα ακουστικό κύμα. Με χρήση εργαλείων (μικροφώνων), που μετρούν τις εναλλαγές της πίεσης, γίνεται η απεικόνιση του ήχου σε κυματομορφή.

2.1.1. Πεδίο Χρόνου

Για την αποθήκευση, επεξεργασία και απεικόνιση του ηχητικού σήματος από ένα ψηφιακό σύστημα είναι απαραίτητη η δειγματοληψία και ο κβαντισμός του. Μια κυματομορφή που προκύπτει από την παραπάνω διαδικασία φαίνεται παρακάτω (εικόνα 1) και μπορεί να είναι αρκετά σύνθετη.

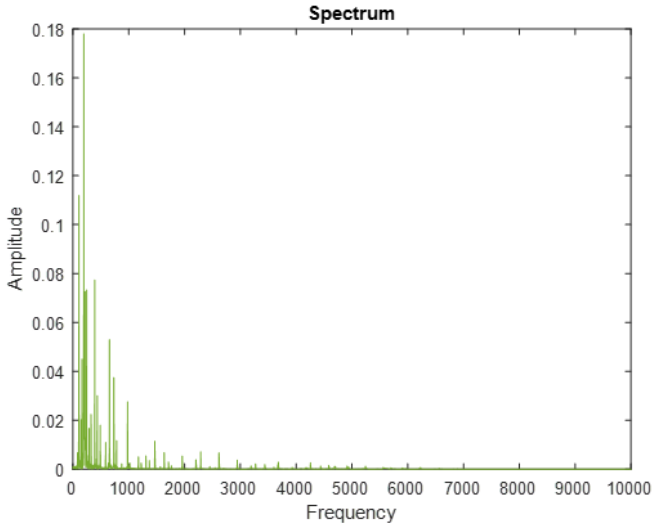


Εικόνα 1 - Κυματομορφή Ηχητικού Σήματος (Πεδίο Χρόνου)

Προκύπτουν λοιπόν ερωτήματα για το πλήθος πληροφορίας που δίνει μια αναπαράσταση στο πεδίο του χρόνου και αν υπάρχει διαφορετικός τρόπος απεικόνισης του ίδιου σήματος.

2.1.1. Πεδίο Συχνότητας

Απάντηση στο προηγούμενο ερώτημα έδωσε ο Fourier, που απέδειξε ότι ένα σήμα μπορεί να αναλυθεί και να γραφτεί ως ένα σύνολο σημάτων ημιτόνου ή συνημιτόνου. Ο μετασχηματισμός Fourier δίνει την δυνατότητα για απεικόνιση του σήματος στο πεδίο της συχνότητας. Συγκεκριμένα εξάγει το φάσμα του σήματος (εικόνα 2), δηλαδή ποιες συχνότητες και σε τι ποσοστό συνθέτουν το σήμα.



Εικόνα 2 - Φάσμα Ηχητικού Σήματος (Πεδίο Συχνότητας)

Ο μετασχηματισμός αυτός έχει αποδειχθεί πολύ σημαντικό εργαλείο για διάφορα συστήματα και τεχνικές επεξεργασίας σημάτων. Έχει αποτελέσει βάση και έμπνευση για άλλους μετασχηματισμούς, όπως ο Μετασχηματισμός Fourier Σύντομου Χρόνου που θα αναφερθεί στην συνέχεια. Παραλλαγές του συγκεκριμένου μετασχηματισμού, σαν τον Γρήγορο Μετασχηματισμό Fourier (FFT), αποτελούν μέρος αρκετών ψηφιακών συστημάτων σήμερα.

2.1.2. Πεδίο Χρόνου – Συχνότητας

Μειονέκτημα του απλού μετασχηματισμού Fourier είναι ότι δεν περιέχει πληροφορία για την χρονική εξέλιξη του σήματος. Είναι χρήσιμος για εξαγωγή πληροφορίας από στάσιμα σήματα ή σήματα που δεν αλλάζουν σημαντικά με τον χρόνο [2].

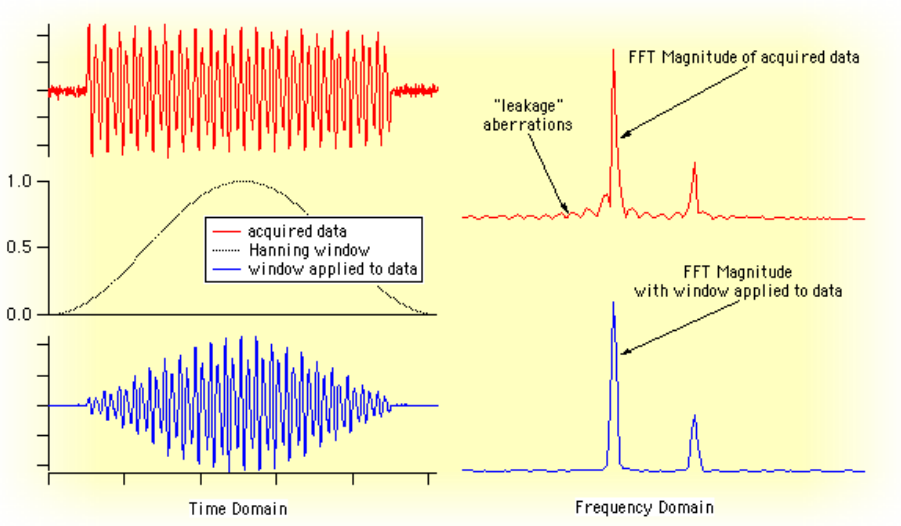
Η ανάπτυξη συστημάτων διαχείρισης πιο σύνθετων σημάτων δημιούργησε την ανάγκη για ταυτόχρονη απεικόνιση στο πεδίο του χρόνου και στο πεδίο της συχνότητας. Στην βιβλιογραφία χρησιμοποιούνται διάφορες τεχνικές (STFT, MFCCs, CQT, CWT) [3], ωστόσο στην αναφορά αυτή παρουσιάζεται μια περιγραφή του Μετασχηματισμού Fourier Σύντομου Χρόνου (STFT) που θα φανεί χρήσιμος σε αυτή την εργασία.

2.1.2.1. Μετασχηματισμός Fourier Σύντομου Χρόνου

Ο Μετασχηματισμού Fourier Σύντομου Χρόνου (STFT) είναι ο υπολογισμός μετασχηματισμού Fourier σε διαφορετικά διαστήματα. Πιο συγκεκριμένα, ο μετασχηματισμός αυτός περιέχει δύο στάδια.

Παραθυροποίηση (Windowing)

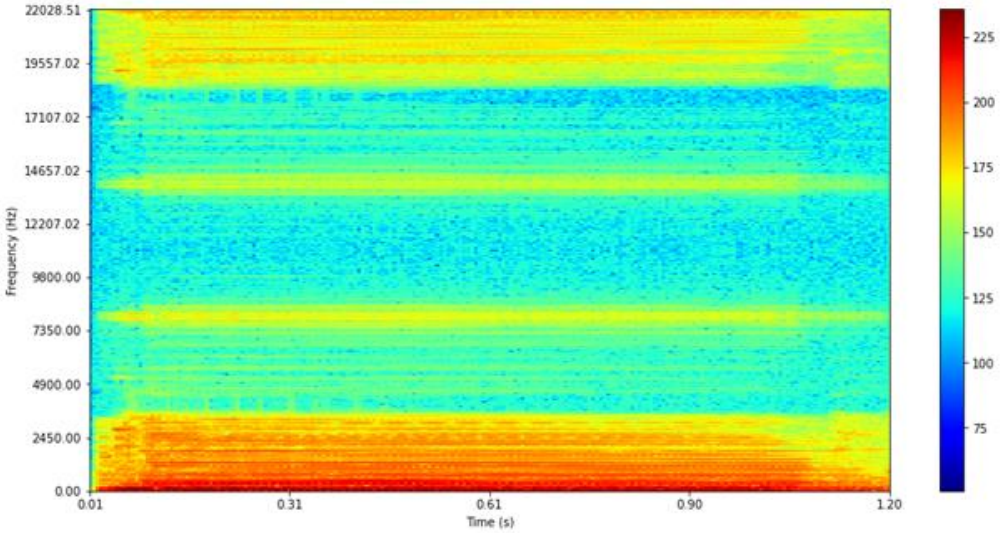
Αρχικά γίνεται η τμηματοποίηση του σήματος με χρήση παραθύρων. Επιλέγεται ένα μέγεθος παραθύρου και εφαρμόζεται στο σήμα, αρκετές φορές κατά το μήκος του. Ο μετασχηματισμός Fourier, που θα εφαρμοστεί στην συνέχεια, κάνει απαραίτητη την χρήση παραθύρων, για λόγους ελαχιστοποίησης της φασματικής διαρροής (spectral leakage) (εικόνα 3). Παρόλα αυτά η χρήση ενός παραθύρου εξασθενεί σημαντικά την αρχή και το τέλος ενός διαστήματος. Γι' αυτό το λόγο τα γειτονικά παράθυρα εμφανίζουν κάποιο ποσοστό επικάλυψης (window - overlapping), έτσι ώστε να ανακτηθεί ένα χαμένο κομμάτι του προηγούμενου διαστήματος.



Εικόνα 3 - Φασματική Διαρροή
Πηγή: [12]

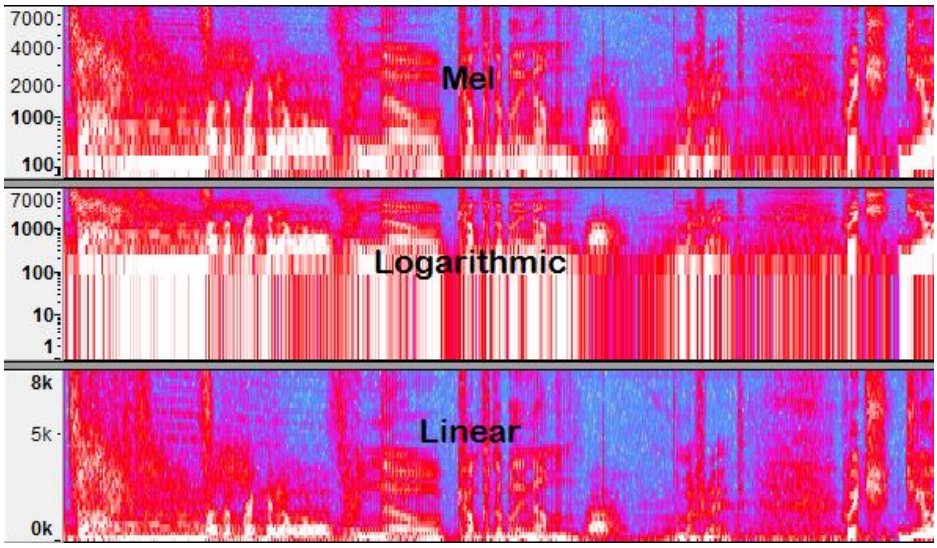
Μετασχηματισμός Fourier

Στην συνέχεια, με τον μετασχηματισμό Fourier υπολογίζεται το φάσμα κάθε διαστήματος που έχει περάσει από ένα παράθυρο και τέλος γίνεται η απεικόνιση κάθε φάσματος σε ένα τρισδιάστατο διάγραμμα όπου ο οριζόντιος άξονας αναπαριστά τον χρόνο, ο κατακόρυφος την συχνότητα και το βάθος-χρώμα το πλάτος των συχνοτήτων. Ένα τέτοιο διάγραμμα ονομάζεται φασματογράφημα (εικόνα 4).



Εικόνα 4 - Φασματογράφημα Ηχητικού Σήματος (Πεδίο Χρόνου – Συχνότητας)

Όπως φαίνεται, ο μετασχηματισμός αυτός δίνει πληροφορίες τόσο για το συχνοτικό περιεχόμενο όσο και την εξέλιξη του σήματος στον χρόνο. Ο ήχος όμως, είναι ένα ψυχο-ακουστικό φαινόμενο [4], έτσι κρίνεται αναγκαία η μελέτη για το πως ο άνθρωπος αντιλαμβάνεται τον ήχο. Μια προσέγγιση με περισσότερη πληροφορία και λιγότερη πολυπλοκότητα, όσον αφορά την επεξεργασία και ανάλυση της ανθρώπινης φωνής, είναι η κλίμακα Mel (Mel scale) [5]. Επειδή το ανθρώπινο αυτί αντιλαμβάνεται καλύτερα τις χαμηλές συχνότητες από ότι τις υψηλές, προτάθηκε η χρήση μιας κλίμακας η οποία χρησιμοποιεί γραμμική αναπαράσταση για τις χαμηλές συχνότητες και λογαριθμική για τις υψηλές (εικόνα 5). Αρκετές είναι οι έρευνες που χρησιμοποιούν την κλίμακα Mel και σημειώνουν την αποτελεσματικότητά της [3]-[6].



Εικόνα 5 - Σύγκριση Φασματογραφημάτων (Κλίμακα Mel, Λογαριθμική, Γραμμική)
Πηγή: [13]

2.2. Μεθοδολογία Εξάλειψης Θορύβου

Η βελτιστοποίηση της ανθρώπινης επικοινωνίας είναι μια ανάγκη που έχει δημιουργηθεί αρκετά χρόνια πριν. Πολλές εφαρμογές και συστήματα έχουν αναπτυχθεί με σκοπό να την κάνουν ποιοτικότερη και καθαρή από θόρυβο. Το είδος του θορύβου, ο χρόνος εκτέλεσης, το μέγεθος του συστήματος αλλά και η τεχνολογία που υπάρχει διαθέσιμη είναι παράμετροι που λαμβάνουν υπόψη τους οι ερευνητές, για την επιλογή της μεθοδολογίας αποθορυβοποίησης.

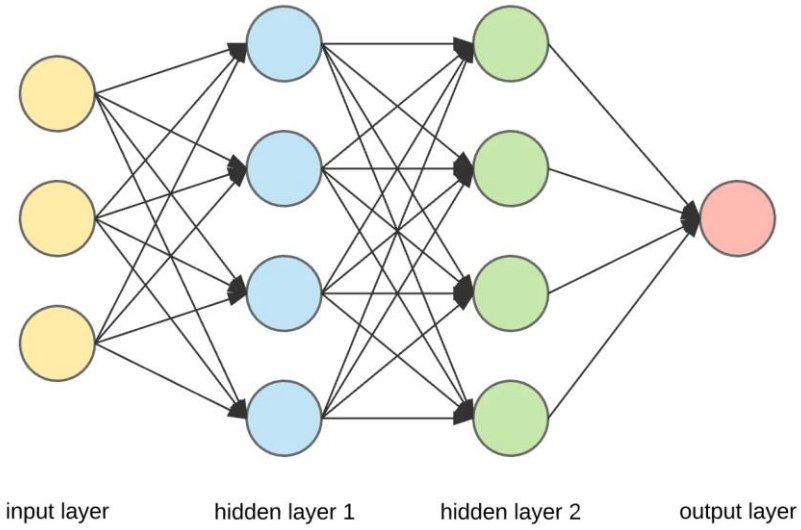
Η φασματική αφαίρεση (spectral subtraction) ήταν μια αρχική ιδέα για το πως θα γίνει η αφαίρεση του θορύβου, ωστόσο η απλότητά της οδηγούσε συχνά σε μη επιθυμητά αποτελέσματα [2]. Έπειτα, κάποια συστήματα καθιέρωσαν την χρήση προσαρμοστικών φίλτρων (adaptive filtering, Wiener filters, Kalman filters) τα οποία προσπαθούν να ελαχιστοποιήσουν κάποιο μαθηματικό σφάλμα με σκοπό να παράγουν το ελεύθερο από θόρυβο σήμα [2] αλλά και σε αυτή την περίπτωση η πολυπλοκότητα των φυσικών σημάτων εισάγει κάποιους προβληματισμούς.

2.2.1. Μηχανική Μάθηση (Machine Learning)

Αν και έχουν αναπτυχθεί αποτελεσματικά συστήματα βασισμένα στις παραπάνω μεθοδολογίες, η ανάγκη για διαχείριση μεγάλου σύνολού θορύβων, έχει οδηγήσει στην χρήση αλγορίθμων μηχανικής μάθησης. Οι αλγόριθμοι αυτοί παρουσιάζουν σημαντική ευελιξία, είναι χρήσιμοι σε πολλά στάδια και έτσι έχουν την δυνατότητα να παρέχουν βοήθεια ως μέρος κάποιου άλλου συστήματος [7], [8] ή να αποτελούν ολοκληρωμένα συστήματα.

2.2.1.1. Βαθιά Μάθηση (Deep Learning)

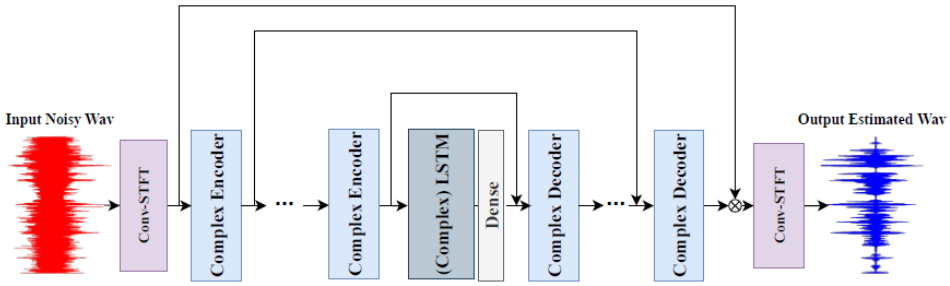
Η εξέλιξη των υπολογιστικών συστημάτων αλλά και των μεθοδολογιών κάνει όλο και συχνότερη την παρουσία μηχανικής μάθησης και συγκεκριμένα της βαθιάς μάθησης σε διάφορες εφαρμογές. Η βαθιά μάθηση είναι ένας κλάδος της μηχανικής μάθησης που στηρίζεται στα βαθιά νευρωνικά δίκτυα (εικόνα 6). Στην περίπτωση της αποθορυβοποίησης δίνουν την δυνατότητα σχεδιασμού end-to-end συστημάτων, ικανών να εντοπίσουν τα χαρακτηριστικά που χρειάζονται με αποτέλεσμα την επιτυχημένη εξάλειψη θορύβου. Συγκεκριμένα, τα βαθιά νευρωνικά δίκτυα αποτελούνται από πολλά επίπεδα, τα επίπεδα αποτελούνται από ένα σύνολο κόμβων, οι οποίοι εκτελούν βασικές πράξεις στα δεδομένα. Ανάλογα τον τύπο πράξεων που εκτελούν οι κόμβοι ενός επιπέδου, χαρακτηρίζεται και το είδος του επιπέδου ή και του δικτύου (πυκνό, συνελικτικό, επαναλαμβανόμενο).



Εικόνα 6 - Απλό παράδειγμα νευρωνικού δικτύου

Σκοπός ενός τέτοιου δικτύου είναι η δημιουργεία κατάλληλων συσχετίσεων μεταξύ δεδομένων εισόδου και εξόδου. Σημαντικό ρόλο σε αυτό παίζει το κομμάτι της εκπαίδευσης του δικτύου, όπου γίνεται η αυτόματη ρύθμιση των διάφορων παραμέτρων του δικτύου.

Στην βιβλιογραφία έχουν αναπτυχθεί πολλά μοντέλα, με σύνθετες αλλά και απλές αρχιτεκτονικές, ικανά για εφαρμογές πραγματικού χρόνου και μη [4], [6], [9], [10]. Αξίζει να αναφερθεί ότι μεθοδολογίες που κερδίζουν μια θέση στο βάθρο στηρίζονται σε DCCR δίκτυα (Deep Complex Convolution Recurrent) [9], δηλαδή, συστήματα που αποτελούνται από νευρωνικά πολλών επιπέδων που λαμβάνουν υπόψη τους την φάση των συχνοτήτων, εφαρμόζουν πράξεις συνέλιξής στα δεδομένα και περιέχουν επίπεδα επαναλαμβανόμενων δομών (εικόνα 7).



Εικόνα 7 - Deep Complex Convolution Recurrent Δίκτυο

Πηγή: [10]

2.3 Μετρικές Αξιολογήσεις

Η δημιουργία τόσων συστημάτων απαιτεί και την αξιολόγησή τους. Με χρήση διάφορων μετρικών μπορεί να γίνει η αξιολόγηση ενός συστήματος, να μελετηθεί η συμπεριφορά του,

να βρεθούν πιθανές αδυναμίες του αλλά και να γίνει σύγκριση με άλλα συστήματα. Στην βιβλιογραφία χρησιμοποιούνται διάφορες μετρικές, κάθε ερευνητής μπορεί να επιλέξει ανάλογα με το σύστημα του και την συμπεριφορά που θέλει να μελετήσει. Στο κομμάτι της αποθορυβοποίησης ομιλίας παρατηρήθηκε ότι χρησιμοποιούνται μετρικές που αξιολογούν ένα σήμα βάση της ποιότητας του (Quality) αλλά και μετρικές που αξιολογούν το κατά πόσο η ομιλία μπορεί να γίνει κατανοητή (Intelligibility). Επιπλέον σε πολλές περιπτώσεις διαμορφώνονται και σύνολα οδηγιών (ITU-T - P.800, ITU-T - P.808), ώστε η αξιολόγηση διαφορετικών συστημάτων να γίνεται υπό τις ίδιες συνθήκες.

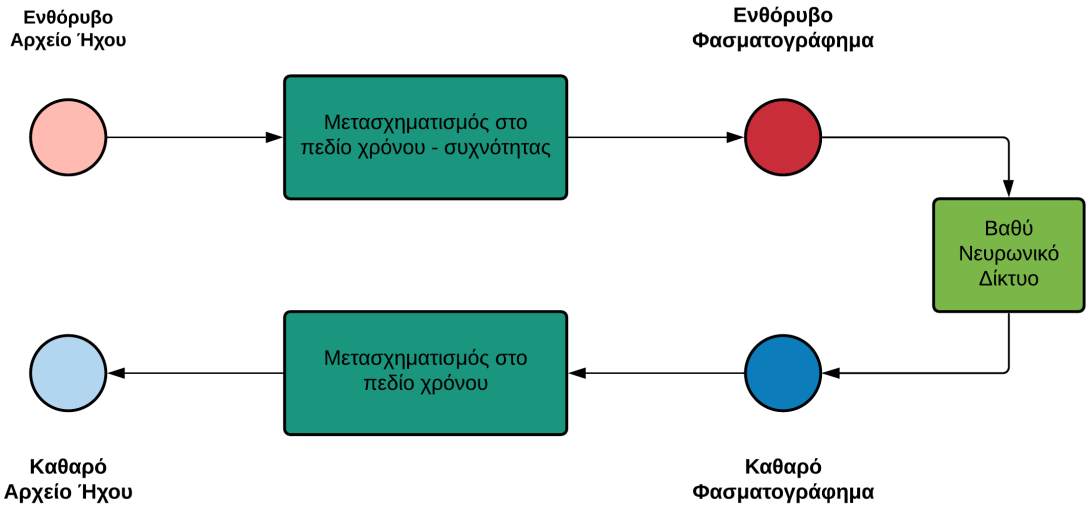
3. Παρουσίαση Συστήματος

Σε αυτό το κεφάλαιο παρουσιάζονται μια σύντομή περιγραφή του προβλήματος που θα προσπαθήσει να αντιμετωπίσει το αναπτυσσόμενο σύστημα αλλά και οι επιλογές που έγιναν στα διάφορα βήματα της υλοποίησης. Γίνεται αναφορά στο σύνολο δεδομένων που επιλέχθηκε, στους τρόπους αναπαράστασης του ήχου, στο νευρωνικό μοντέλο που θα χρησιμοποιηθεί και με ποιόν τρόπο αλλά και στις μετρικές αξιολόγησης που θα χρησιμοποιηθούν.

3.1. Συνοπτική Περιγραφή Προσέγγισης

Το πρόβλημα που καλείται να αντιμετωπίσει η ομάδα, είναι η αφαίρεση θορύβου σε αρχεία ήχου. Για το λόγο αυτό αποφασίστηκε η σχεδίαση ενός συστήματος βαθιάς μηχανικής μάθησης, με σχετικά χαμηλή πολυπλοκότητα ώστε να μπορεί να τρέξει σε έναν κοινό επεξεργαστή ή μια απλή κάρτα γραφικών.

Το σύστημα θα δέχεται στην είσοδό του ένα αρχείο ήχου, θα κάνει την απαραίτητη προεπεξεργασία, θα κάνει την απεικόνιση στο πεδίο χρόνου-συχνότητας και θα προωθεί το αποτέλεσμα σε ένα βαθύ νευρωνικό δίκτυο που θα παράγει μια νέα αναπαράσταση στο ίδιο πεδίο. Στην συνέχεια θα εφαρμόζει έναν αντίστροφο μετασχηματισμό για την απεικόνιση του καινούριου σήματος στο πεδίο του χρόνου, με σκοπό την κατασκευή, αναπαραγωγή και αποθήκευση της ελεύθερης πλέον από θόρυβο τελικής κυματομορφής. Ένα μπλοκ διάγραμμα που παρουσιάζει τις διαδικασίες ενός τέτοιου συστήματος φαίνεται στην παρακάτω (εικόνα 8).



Εικόνα 8 - Ενδεικτικό Μπλοκ Διάγραμμα Συστήματος

3.2. Σύνολο Δεδομένων

Το σύνολο των δεδομένων που αποφασίστηκε να αξιοποιηθεί προσφέρεται από την Microsoft (Microsoft Scalable Noisy Speech Database, MS-SNSD) [16]. Αποτελείται από δύο διαφορετικά μέρη, το 1ο μέρος αφορά την εκπαίδευση του νευρονικού και περιέχει καθαρές ομιλίες στα αγγλικά χωρίς καθόλου θόρυβο από ένα μεγάλο πλήθος ομιλητών αλλά και ένα μεγάλο αριθμό ηχητικών θορύβων. Το 2ο μέρος αποσκοπεί στην αξιολόγηση ενός μοντέλου, περιέχει επίσης δεδομένα καθαρής ομιλίας και θορύβου ξεχωριστά αλλά σε μικρότερο πλήθος σε σχέση με το 1ο. Και στις δύο περιπτώσεις επιλέχθηκαν συγκεκριμένοι τύποι θορύβων, ανάλογα με το αν υπήρχαν επαρκή δεδομένα και έγινε η μίξη τους με τα σήματα καθαρής ομιλίας σε διάφορες αναλογίες SNR (0 dB, 5 dB, 10 dB, 15 dB, 20 dB) με σκοπό την καλύτερη εκπαίδευση του νευρωνικού και την καλύτερη εποπτεία του συστήματος όσον αφορά τη δυνατότητα του για αποθορυβοποίηση σε διαφορετικά ποσοστά θορύβου. Έτσι δημιουργήθηκε ένα σύνολο τεσσάρων ωρών για εκπαίδευση και μισής ώρας για αξιολόγηση. Σημειώνεται ότι μαζί με το σύνολο δεδομένων δινόντουσαν και χρησιμοποιήθηκαν συναρτήσεις που αφορούν το διάβασμα των ηχητικών αρχείων την κανονικοποίηση των σημάτων αλλά και την μίξη καθαρής ομιλίας με τον θόρυβο.

3.3. Μετασχηματισμός στο Πεδίο Χρόνου-Συχνότητας

Για την απεικόνιση του σήματος στο πεδίο χρόνου-συχνότητας θα εφαρμοστεί ο μετασχηματισμός Fourier σύντομου χρόνου (STFT). Η επιλογή αυτού έγινε καθώς στην βιβλιογραφία σημειώνεται η συνεισφορά του στην συνολική αποτελεσματικότητα [3]. Αφού δοκιμάστηκαν φασματογραφήματα πλάτους, ισχύος λογαριθμικά ή μη, αποφασίστηκε να

γίνει χρήση φασματογραφημάτων πλάτους καθώς παρουσίαζαν καλύτερα αποτελέσματα από τα υπόλοιπα.

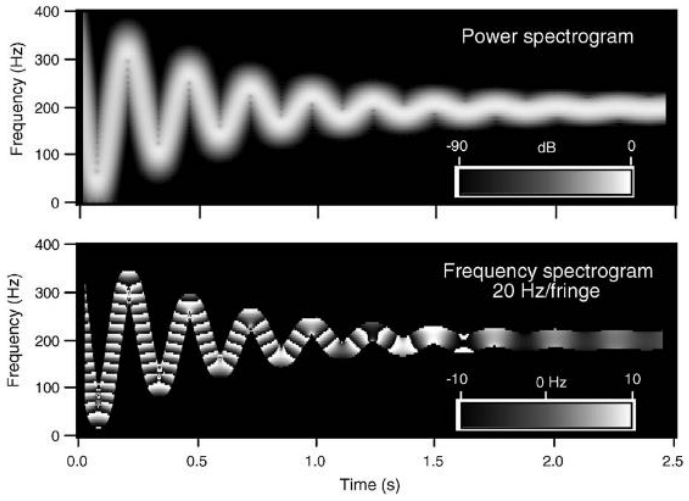
Επιπλέον, στο μετασχηματισμό αυτό το μέγεθος του παραθύρου αλλά και το ποσοστό επικάλυψης των γειτονικών παραθύρων είναι παράμετροι που μπορούν να μεταβάλλονται ώστε να βελτιώνεται η αποτελεσματικότητα του συστήματος ή ακόμα και να μελετηθεί η επιρροή τους στο συνολικό αποτέλεσμα. Για την συγκεκριμένη εργασία έγινε επιλογή μεγέθους παραθύρου στα 512 δείγματα με ποσοστό επικάλυψης τα 128 δείγματα.

3.4. Μετασχηματισμός στο Πεδίο Χρόνου

Αντίστροφη διαδικασία της παραπάνω αποτελεί ο μετασχηματισμός στο πεδίο του χρόνου, που θα γίνει με την εφαρμογή του αντίστροφου μετασχηματισμού Fourier σύντομου χρόνου (iSTFT).

Σημειώνεται ότι το μέγεθος του παραθύρου αλλά και το ποσοστό επικάλυψης των γειτονικών παραθύρων πρέπει να είναι ίδια και στους δύο μετασχηματισμούς, με αποτέλεσμα να χρειάζεται ιδιαίτερη προσοχή.

Επιπλέον σε αυτό το σημείο αναφέρεται ότι ο μετασχηματισμός Fourier σύντομου χρόνου έχει σαν αποτέλεσμα τόσο το φασματογράφημα ισχύος, όσο και το φασματογράφημα φάσης των συχνοτήτων (εικόνα 9). Στο έγγραφο η χρήση του όρου φασματογράφημα γινόταν υποδηλώνοντας το φασματογράφημα ισχύος ή πλάτους. Υπάρχουν διάφορες έρευνες που μελετούν την επιρροή της φάσης στο τελικό αποτέλεσμα, συστήματα που δεν την χρησιμοποιούν καθόλου[5] και άλλα που την χρησιμοποιούν και είναι πολύ αποτελεσματικά [10].



Εικόνα 9 - Φασματογράφημα Ισχύος, Φασματογράφημα Φάσης

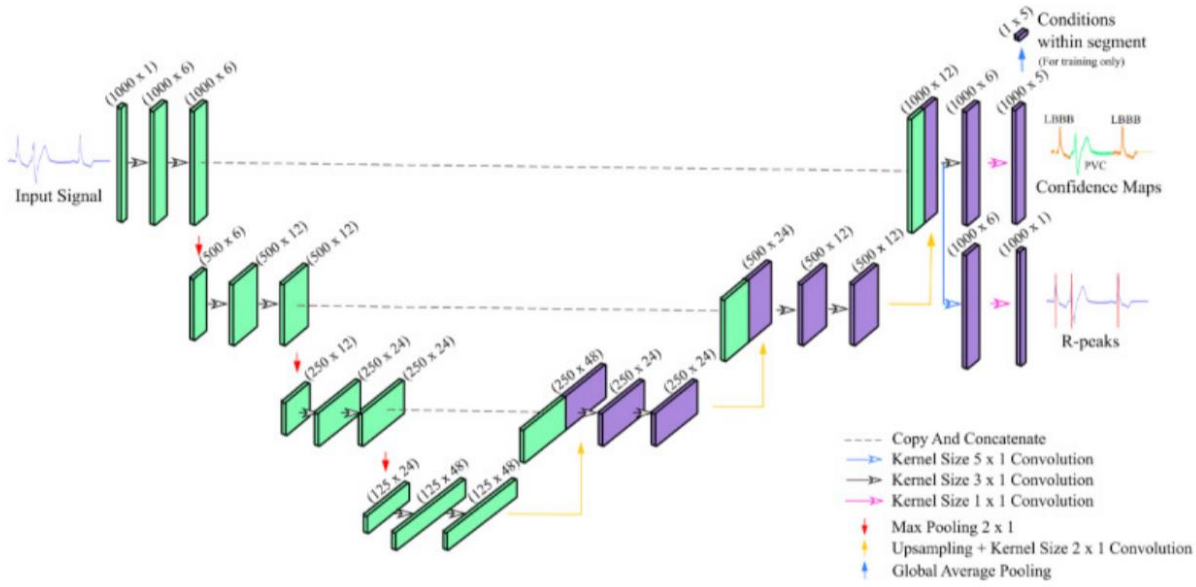
Πηγή: [11]

Στην εργασία χρησιμοποιήθηκαν τα φασματογραφήματα φάσης αποκλιστηκά για την επιστροφή του σήματος από το πεδίο της συχνότητας στο πεδίο του χρόνου και όχι για την αποθορυβοποίηση τους.

3.5. Βαθύ Νευρωνικό Δίκτυο

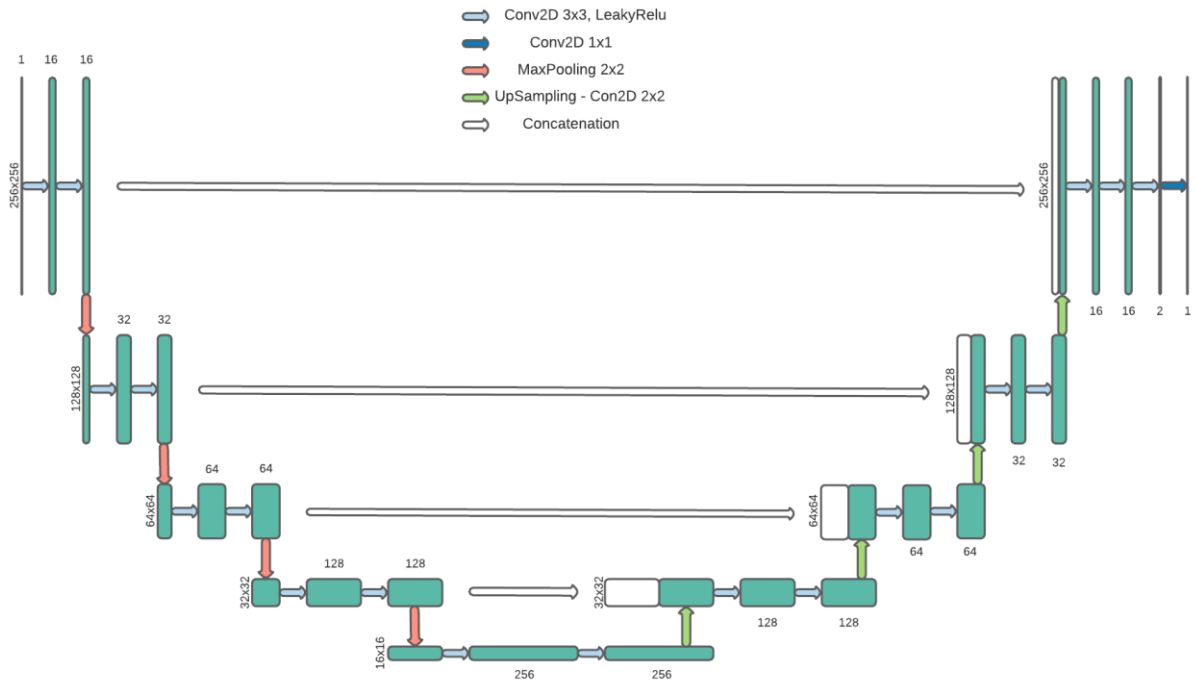
Όπως ήδη αναφέρθηκε, το νευρωνικό δέχεται στην είσοδο του διάφορα είδη φασματογραφημάτων. Αυτό κάνει δυνατή την χρήση αρχιτεκτονικών που εφαρμόζονται στην διαχείριση και επεξεργασία εικόνων [5]. Κάποια μοντέλα, που χρησιμοποιούνται σε τέτοιες εφαρμογές, αποτελούνται από συνελικτικά νευρωνικά δίκτυα (CNNs) και παίρνουν το ρόλο κωδικοποιητή - αποκωδικοποιητή.

Σε αυτή την εργασία χρησιμοποιήθηκαν μοντέλα U-Net [15] (εικόνα 10), που αποτελούνται από αμιγώς συνελικτικά νευρωνικά δίκτυα (CNNs) [9] και ο τρόπος με τον οποίο ιεραρχούνται οι είσοδοι και οι έξοδοι της κάθε βαθμίδας σχηματίζουν το γράμμα «U» το οποίο αποτελεί και την προέλευση της ονομασίας τους.



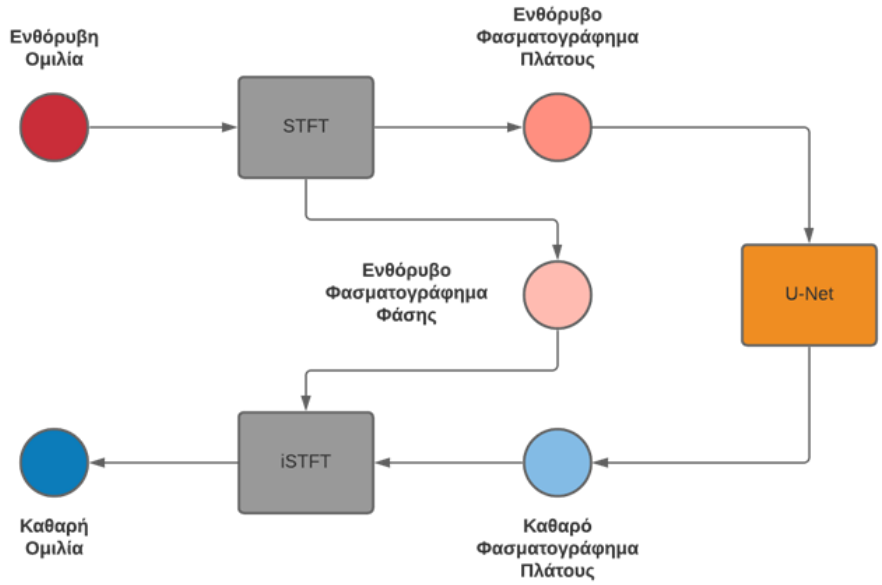
Εικόνα 10 - Μοντέλο U-Net
Πηγή: [15]

Το μοντέλο που διαμορφώθηκε βάση των αναγκών του συστήματος που μελετάτε φαίνεται παρακάτω (εικόνα 11). Από το σχήμα γίνονται διακριτά όλα τα ενδιάμεσα επίπεδα από τα οποία θα προσέλθει κάθε φασματογράφημα καθώς και οι πράξεις που θα εκτελούνται σε αυτό.

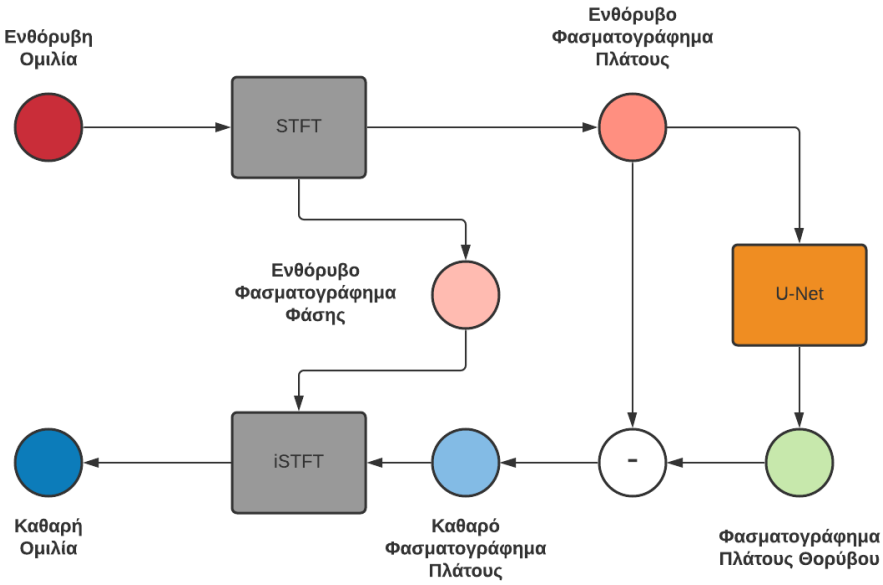


Εικόνα 11 - Μοντέλο U-NET Συστήματος

Σε αυτό το σημείο αναφέρεται ότι θα εξεταστεί η αποτελεσματικότητα του μοντέλου σε δύο περιπτώσεις. Στην πρώτη το νευρωνικό σύστημα προβλέπει απευθείας τον καθαρό ήχο (εικόνα 12) ενώ στη δεύτερη περίπτωση το νευρωνικό σύστημα καλείται να προβλέψει τον θόρυβο και στη συνέχεια με μια απλή αφαίρεση από το αρχικό ενθόρυβο φασματογράφημα να υπολογιστεί το καθαρό (εικόνα 13).



Εικόνα 12 - Διάγραμμα Συστήματος Πρόβλεψης Καθαρής Ομιλίας



Εικόνα 13 – Διάγραμμα Συστήματος Πρόβλεψης Θορύβου

3.6. Αξιολόγηση Μοντέλου

Η αξιολόγηση του συστήματος θα γίνει με μετρικές που αναφέρονται στην βιβλιογραφία (STOI, Segmental SNR, Frequency-Weighted SNR).

Ο τμηματικός λόγος σήματος προς θόρυβο (Segmental SNR), αντί να λειτουργεί σε ολόκληρο το σήμα, υπολογίζει τον μέσο όρο των τιμών SNR των μικρών τμημάτων (15 έως 20ms).

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\left(\sum_{i=Nm}^{Nm+N-1}\left(\frac{\sum_{i=1}^{N}x^2(i)}{\sum_{i=1}^{N}\big(x(i)-y(i)\big)^2}\right)\right)$$

Το μέτρο fwSNRseg σχεδιάστηκε αρχικά ως μια αντικειμενική μετρική για την πρόβλεψη της ποιότητας ομιλίας, και έχει αποδειχθεί ότι συσχετίζεται καλά με την κατανοησιμότητά της.

Το μέτρο fwSNRseg υπολογίζεται ως:

$$fwSNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K}W(j,m)\log_{10}\frac{X(j,m)^2}{(X(j,m)-\widehat{X(j,m)})^2}}{\sum_{j=1}^{K}W(j,m)}$$

‘Οπου W(j,m) είναι το βάρος που αντιστοιχεί στην j_{th} ζώνη συχνοτήτων στο m_{th} καρέ, το K είναι ο αριθμός των ζωνών, το M είναι ο συνολικός αριθμός καρέ στο σήμα. Το X(j,m) είναι το μέγεθος κρίσιμης ζώνης του καθαρού σήματος στην j_{th} ζώνη συχνοτήτων στο m_{th} καρέ, και $\dot{X}$(j,m) είναι το αντίστοιχο φασματικό το μέγεθος του επεξεργασμένου σήματος στην ίδια ζώνη συχνοτήτων και ίδιο πλαίσιο.

Σημειώνεται ότι οι μετρικές που αναφέρονται χρησιμοποιήθηκαν από βιβλιοθήκες που βρέθηκαν στο διαδίκτυο [19], [20].

3.7. Εργαλεία Ανάπτυξης

Σημαντικό για την ανάπτυξη αλγορίθμων μηχανική μάθησης αποτελούν τα εργαλεία προεπεξεργασίας των δεδομένων, της εξαγωγής χαρακτηριστικών και της ανάπτυξης κώδικα.

Τα παρακάτω εργαλεία χρησιμοποιήθηκαν για την ανάπτυξη της εργασίας:

1. Python για την ανάπτυξη της εφαρμογής αποθορυβοποίησης.

2. Keras API για την ανάπτυξη αλγορίθμων μηχανικής μάθησης.

3. STOI, Segmental SNR, frequency weighted SNR για την αξιολόγηση του αλγορίθμου.

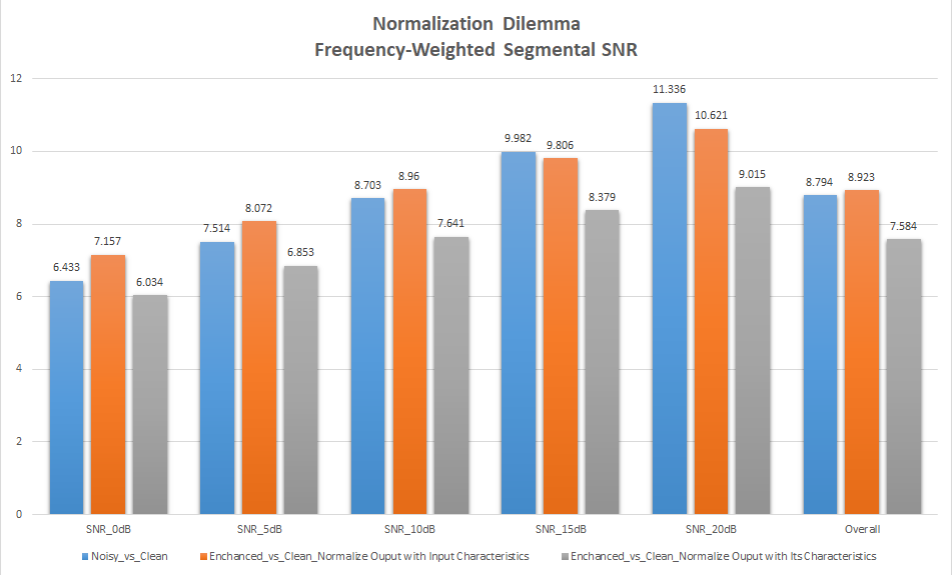
4. Συμπεράσματα & Τελική Υλοποίηση

Στο παρόν κεφάλαιο παρουσιάζονται τα αποτελέσματα από διάφορα πειράματα τα οποία έγιναν με σκοπό να μελετηθεί η καταλληλότερη δομή του συστήματος.

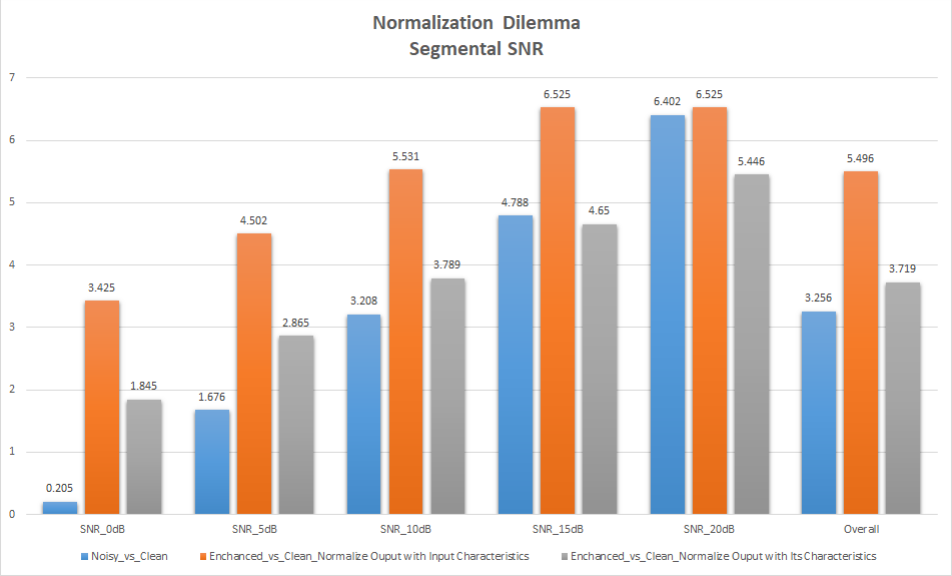
4.1. Ερώτημα Κανονικοποίησης

Κατά την συγγραφεί του απαραίτητου κώδικα για την υλοποίηση του συστήματος δημιουργήθηκε ο προβληματισμός για τον τρόπος με τον οποίο θα γίνεται η κανονικοποίηση των δεδομένων εισόδου για την εκπαίδευση του μοντέλου. Οι προτάσεις ήταν να χρησιμοποιούνται τα χαρακτηριστικά του σήματος εισόδου και στο ίδιο το σήμα εξόδου, ή να χρησιμοποιηθούν στο σήμα εξόδου τα δικά του χαρακτηριστικά.

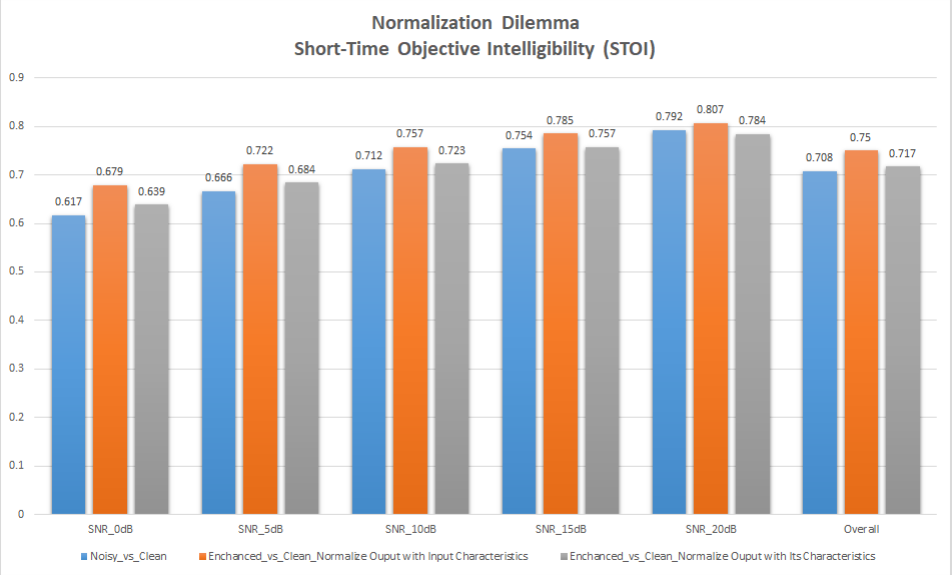
Για τον λόγο αυτό έγινε η αξιολόγηση του συστήματος, με τις μετρικές που αναφέρθηκαν στο προηγούμενο κεφάλαιο, και στις δύο περιπτώσεις. Τα αποτελέσματα είναι τα παρακάτω (εικόνα 14, εικόνα 15, εικόνα 16). Με μπλε χρώμα απεικονίζονται οι μετρικές για το αρχικό ενθόρυβο σήμα, ενώ με πορτοκαλί και γκρι τα αποθορυβοποιημένα σήματα για τις δύο περιπτώσεις αντίστοιχα.



Εικόνα 14 – Ερώτημα Κανονικοποίησης - Frequency-Weighted Segmental SNR



Εικόνα 15 - Ερώτημα Κανονικοποίησης - Segmental SNR

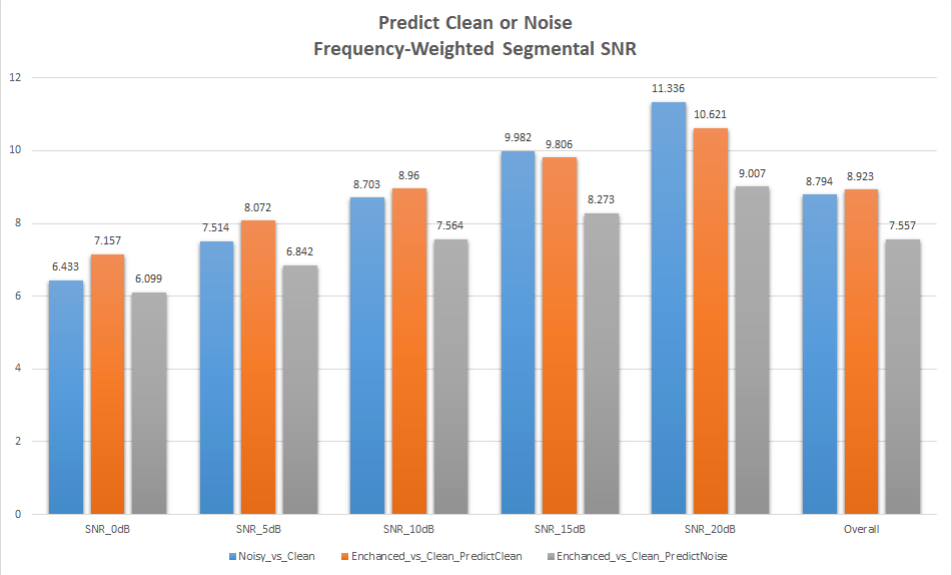


Εικόνα 16 - Ερώτημα Κανονικοποίησης - STOI

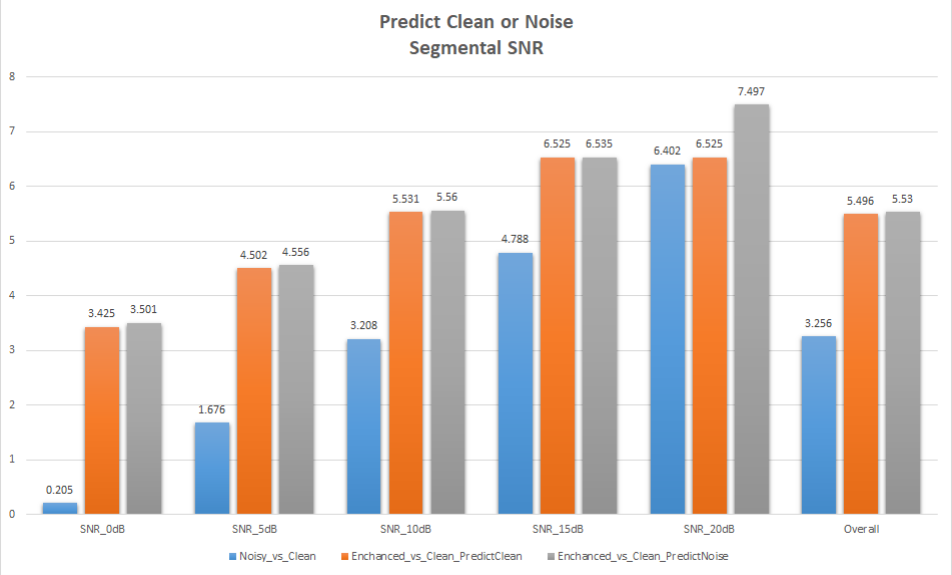
Στην προκειμένη περίπτωση σε όλες τις μετρικές φαίνεται ότι η κανονικοποίηση με βάση τα χαρακτηριστικά της εισόδου φαίνεται να αποδίδουν πολύ καλύτερα σε σχέση με τη δεύτερη περίπτωση. Αυτό είναι ιδιαίτερα βολικό διότι στην περίπτωση που θέλουμε να αποθορυβοποιήσουμε ένα άγνωστο σήμα, δεν υπάρχουν από πριν τα χαρακτηριστικά της εξόδου οπότε με αυτό τον τρόπο πρέπει να γίνεται και η εκπαίδευση. Επίσης αξίζει να σημειωθεί ότι στην Frequency-Weighted Segmental SNR μετρική στα 15 dB και 20 dB βλέπουμε ότι και στην περίπτωση που γίνεται κανονικοποίηση της εξόδου με τα χαρακτηριστικά της εισόδου (πορτοκαλί χρώμα), η μετρική έχει χαμηλότερη τιμή. Αυτό δείχνει ότι το σύστημα, στις περιπτώσεις που ο θόρυβος υπάρχει σε χαμηλότερο επίπεδο, αλλοιώνει την ποιότητα του σήματος περισσότερο.

4.2. Ερώτημα Δομής Συστήματος

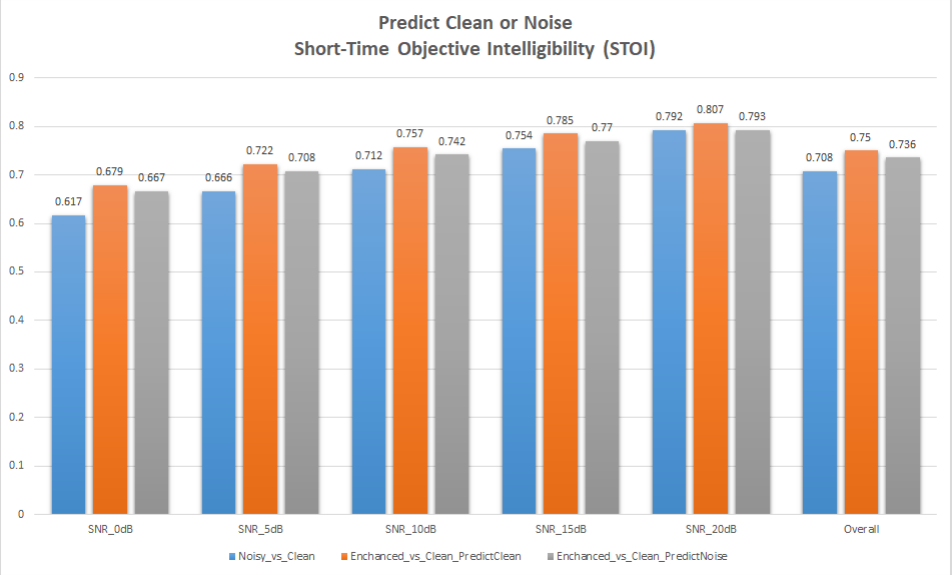
Επόμενο ερώτημα που εξετάστηκε ήταν αυτό της δομής του συστήματος, όπως αυτό αναφέρθηκε στο προηγούμενο κεφάλαιο. Στην πρώτη περίπτωση, που απεικονίζεται με πορτοκαλί χρώμα, το νευρωνικό υπολογίζει απευθείας το φασματογράφημα της καθαρής φωνής, ενώ η δεύτερη που απεικονίζεται με γκρι, υπολογίζει το φασματογράφημα του θορύβου το οποίο αφαιρείται από το αρχικό φασματογράφημα για να παραχθεί το αποθορυβοποιημένο. Τα αποτελέσματα παρουσιάζονται παρακάτω (εικόνα 17, εικόνα 18, εικόνα 19).



Εικόνα 17 - Ερώτημα Δομής - Frequency-Weighted Segmental SNR



Εικόνα 18 - Ερώτημα Δομής - Segmental SNR

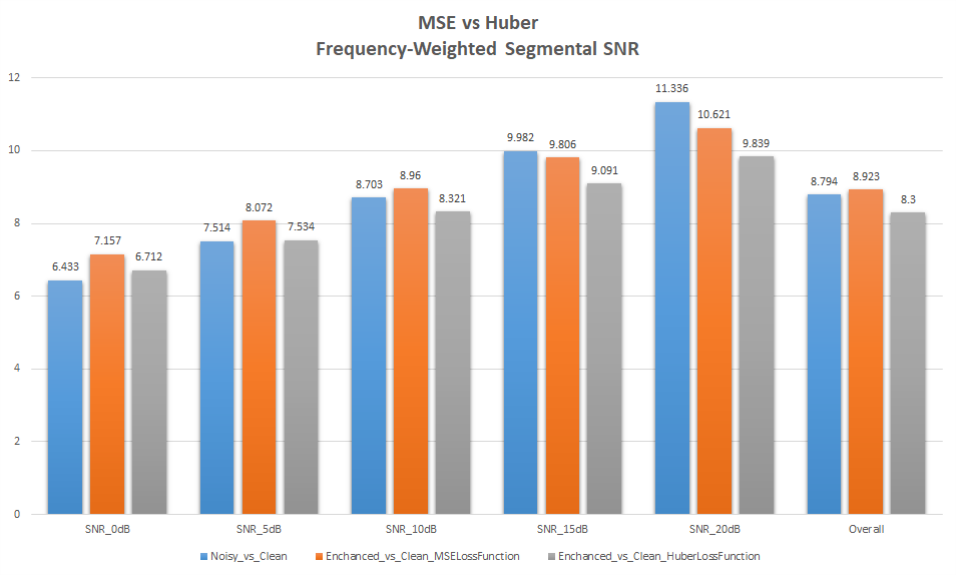


Εικόνα 19 - Ερώτημα Δομής – STOI

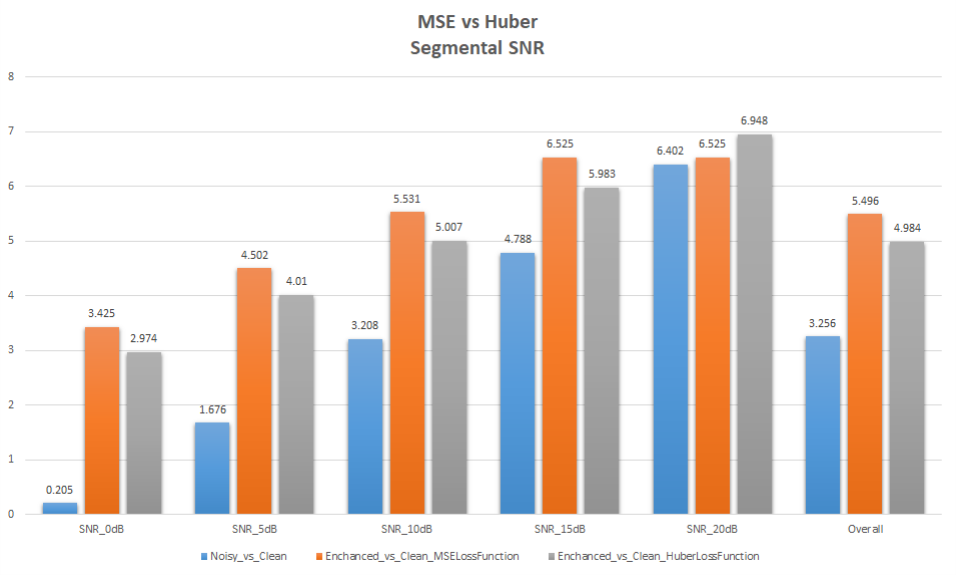
Από τις παραπάνω μετρικές φαίνεται ότι ενώ στη STOI και στη Segmental SNR είναι αρκετά ίδια τα αποτελέσματα των 2 διαφορετικών υλοποιήσεων, στη Frequency-Weighted Segmental SNR είναι προφανές ότι η πρόβλεψη με βάση την καθαρή φωνή είναι αρκετά καλύτερη. Γι' αυτό το λόγο αποφασίστηκε να ακολουθηθεί και στα υπόλοιπα βήματα της εργασίας η συγκεκριμένη μέθοδος.

4.3 Συνάρτηση Κόστους Νευρωνικού Μοντέλου

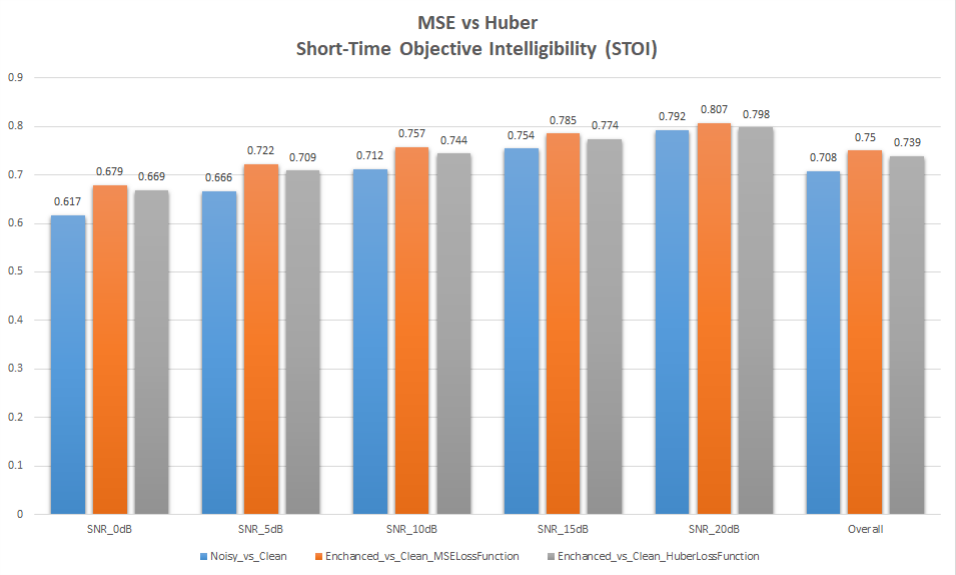
Ένα ακόμα ερώτημα που εξετάστηκε είναι το ποια θα έπρεπε να είναι η συνάρτηση κόστους σε ένα νευρωνικό το οποίο καλείται να δώσει λύση στο πρόβλημα της αποθορυβοποίησης. Έγιναν δοκιμές για συναρτήσεις κόστους που βρέθηκαν στην βιβλιογραφία και σε παραδείγματα του διαδικτύου. Ωστόσο παρακάτω (εικόνα 20, εικόνα 21, εικόνα 22). παρουσιάζονται τα αποτελέσματα για την συνάρτηση μέσου τετραγωνικού σφάλματος, με πορτοκαλί χρώμα και της συνάρτησης Huber με γκρι χρώμα. Η απεικόνιση των δύο αυτών συναρτήσεων γίνεται καθώς αυτές παρουσίασαν καλύτερα αποτελέσματα σε σχέση με τις υπόλοιπες που δοκιμάστηκαν.



Εικόνα 20 – Ερώτημα Συνάρτησης Κόστους - Frequency-Weighted Segmental SNR



Εικόνα 21 – Ερώτημα Συνάρτησης Κόστους - Segmental SNR



Εικόνα 22 – Ερώτημα Συνάρτησης Κόστους - STOI

Στην πλειοψηφία των αποτελεσμάτων η συνάρτηση κόστους μέσου τετραγωνικού σφάλματος (MSE) φέρνει καλύτερα αποτελέσματα από την Huber. Η συνάρτηση μέσου τετραγωνικού σφάλματος είναι μια μετρική η οποία σύμφωνα με τη βιβλιογραφία συμπεριφέρεται πολύ καλά σε εικόνες και από τη στιγμή που η είσοδος στο νευρωνικό είναι φασματογράφημα το οποίο αναπαρίσταται όπως και μία εικόνα, είναι λογική η σχετική αποτελεσματικότητα της στο σύστημα.

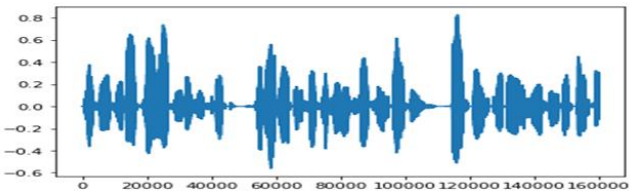
4.4. Τελική Υλοποίηση

Με βάση τα παραπάνω αποτελέσματα διαμορφώθηκε ένα τελικό σύστημα, το οποίο αφού δεχτεί ένα ηχητικό σήμα ενθόρυβης ομιλίας, το διασπάσει σε τμήματα κατάλληλου μεγέθους και χρησιμοποιεί το είδη εκπαιδευμένο μοντέλο με σκοπό να καταστείλει τον θόρυβο σε κάθε τμήμα και συνεπώς σε όλο το ηχητικό σήμα.

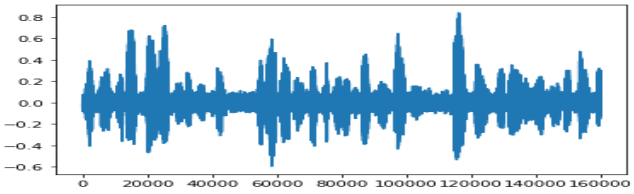
Το τελικό σύστημα που διαμορφώθηκε χρησιμοποιεί το νευρωνικό μοντέλο που εκπαιδεύτηκε ώστε να παράγει κατευθείαν την καθαρή ομιλία (εικόνα 12) και χρησιμοποιεί συνάρτηση μέσου τετραγωνικού σφάλματος. Παρακάτω παρουσιάζονται οι κυματομορφές ενθόρυβου, καθαρού και αποθορυβοποιημένου σήματος για δύο παραδείγματα ομιλίας.

Παράδειγμα 1

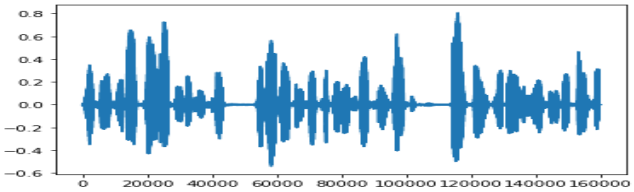
Καθαρή Ομιλία



Ενθόρυβη Ομιλία



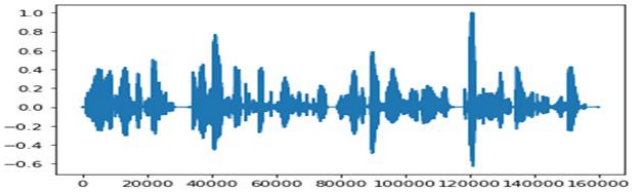
Αποθορυβοποιημένη Ομιλία



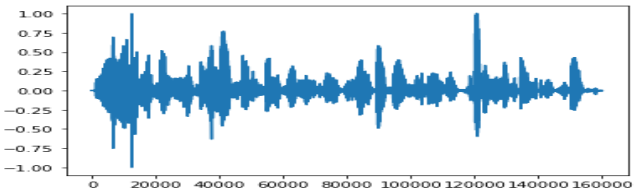
Εικόνα 23 - Παράδειγμα Κυματομορφών 1

Παράδειγμα 2

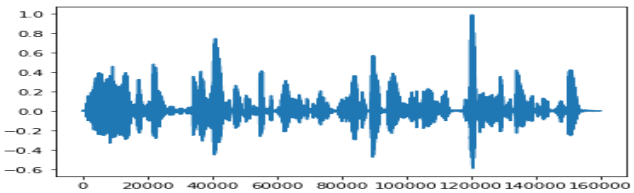
Καθαρή Ομιλία



Ενθόρυβη Ομιλία



Αποθορυβοποιημένη Ομιλία



Εικόνα 24 - Παράδειγμα Κυματομορφών 2

5. Τρόποι Βελτίωσης

Στο κεφάλαιο αυτό αναφέρεται ότι υπάρχουν διάφοροι τρόποι που μπορούν να βελτιώσουν την αποτελεσματικότητα του συστήματος αλλά και να αξιολογήσουν την ποιότητα του.

Κάποιες βελτιώσεις θα μπορούσαν να είναι η ανάπτυξη και η εκπαίδευση του μοντέλου σε μεγαλύτερο πλήθος δεδομένων, η ανάπτυξη νευρονικού δικτύου μεγαλύτερου βάθους, η χρήση πιο εξειδικευμένων συναρτήσεων κόστους αλλά και η χρήση της φάσης και στο κομμάτι της αποθορυβοποίησης.

Επιπλέον μπορεί να γίνει επιπλέον έρευνα για το πώς το σύστημα ανταποκρίνεται σε άλλες μετρικές αλλά και σε σύνολα αξιολόγησης με σκοπό την σύγκρισή του με ακόμα περισσότερα μοντέλα που κυριαρχούν στο χώρο. Επιπρόσθετα μπορεί να γίνει πιο αναλυτική μελέτη για το πως το σύστημα συμπεριφέρεται σε διάφορα επίπεδα θορύβου, σε διαφορετικά είδη θορύβου αλλά και σε διαφορετικές γλώσσες ομιλίας.

6. Βιβλιογραφία

[1] C. Cole, M. Karam, and H. Aglan, “Spectral subtraction of noise in speech pro-cessing applications,” in2008 40th Southeastern Symposium on System Theory(SSST). IEEE, 2008, pp. 50–53.

[2] P. C. Loizou,Speech enhancement: theory and practice. CRC press, 2013.

[3] M. Huzaifah, “Comparison of time-frequency representations for environmen-tal sound classification using convolutional neural networks,”arXiv preprintarXiv:1706.07156, 2017.

[4] S. R. Park and J. Lee, “A fully convolutional neural network for speech enhance-ment,”arXiv preprint arXiv:1609.07132, 2016.

[5] K. Tan and D. Wang, “A convolutional recurrent neural network for real-timespeech enhancement.” inInterspeech, 2018, pp. 3229–3233.

[6] V. Belz, “Speech-enhancement with deep learning,” https://towardsdatascience.com/speech-enhancement-with-deep-learning-36a1991d3d8d,2020, accessed on 10.10.2020.

[7] J.-M. Valin, “A hybrid dsp/deep learning approach to real-time full-band speechenhancement,” in2018 IEEE 20th international workshop on multimedia signalprocessing (MMSP). IEEE, 2018, pp. 1–5.

[8] B. Xia and C. Bao, “Wiener filtering based speech enhancement with weighteddenoising auto-encoder and noise classification,”Speech Communication,vol. 60, pp. 13–29, 2014.

[9] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks forbiomedical image segmentation,” inInternational Conference on Medical imagecomputing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[10] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie,“Dccrn: Deep complex convolution recurrent network for phase-aware speechenhancement,”arXiv preprint arXiv:2008.00264, 2020.

[11] F. Léonard, “Phase spectrogram and frequency spectrogram as new diagnostictools,”Mechanical Systems and Signal Processing, vol. 21, no. 1, pp. 125–137,2007.

[12] F. J. Harris, “On the use of windows for harmonic analysis with the discretefourier transform,”Proceedings of the IEEE, vol. 66, no. 1, pp. 51–83, 1978.

[13] audacityteam.org, “Spectrogram view,” https://manual.audacityteam.org/man/spectrogram_view.html, 2020, accessed on 10.10.2020.

[14] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for seman-tic segmentation,” inProceedings of the IEEE conference on computer vision andpattern recognition, 2015, pp. 3431–3440.

[15] S. L. Oh, E. Y. Ng, R. San Tan, and U. R. Acharya, “Automated beat-wisearrhythmia diagnosis using modified u-net on extended electrocardiographicrecordings with heterogeneous arrhythmia types,”Computers in biology andmedicine, vol. 105, pp. 92–101, 2019.

[16] C. K. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matu-sevych, R. Aichner, A. Aazami, S. Braunet al., “The interspeech 2020 deepnoise suppression challenge: Datasets, subjective speech quality and testingframework,”arXiv preprint arXiv:2001.08662, 2020.

[17] D. Ellis, “Sound examples,” https://www.ee.columbia.edu/~dpwe/sounds/?fbclid=IwAR0oK6fez2hHpN7HvKBWY3PfunAU-c2S8XzPZoHlYARWVMhYdqvciiVJwas,2003, accessed on 10.10.2020.

[18] Z. Liu, H. T. Ma, and F. Chen, “A new data-driven band-weighting function forpredicting the intelligibility of noise-suppressed speech,” in2017 Asia-PacificSignal and Information Processing Association Annual Summit and Conference(APSIPA ASC). IEEE, 2017, pp. 492–496.

[19] P. Manuel, “Python implementation of stoi,” https://github.com/mpariente/pystoi, 2020.

[20] schmiph2, “Python speech enhancement performance measures (quality andintelligibility),” https://github.com/schmiph2/pysepm, 2020.

Παράρτημα – Αρμοδιότητες Μελών Ομάδας

Στο παράρτημα αυτό αναφέρετε ο τρόπος με τον οποίο λειτούργησε η ομάδα καθώς και οι αρμοδιότητες που ανέλαβε κάθε μέλος. Σημειώνεται ότι λόγο των ιδιαίτερων συνθήκων, η συνεννόηση των μελών έγινε σε μεγάλο ποσοστό από απόσταση. Ωστόσο αυτό δεν αποτέλεσε εμπόδιο ώστε η συγκεκριμένη εργασία να αποτελεί μια συλλογική προσπάθεια. Κοινοί στόχοι των μελών ήταν η ενασχόληση με έννοιες που εξετάστηκαν στα προηγούμενα κεφάλαια αλλά και η παρουσίαση μιας ολοκληρωμένης εργασίας, έτσι δεν ήταν δύσκολος ο καταμερισμός των επιμέρους εργασιών και η ανάπτυξη σχέσεων εμπιστοσύνης.

Πιο συγκεκριμένα παρακάτω αναφέρονται οι αρμοδιότητες που ανέλαβαν τα μέλη για κάθε στάδιο:

Ο **Αμοιρίδης Βασίλειος** συνέβαλε στην αναζήτηση και επιλογή τρόπου προσέγγισης του προβλήματος, στην αναζήτηση και διαχείριση του συνόλου δεδομένων, στην αναζήτηση και επιλογή κατάλληλου μετασχηματισμού του ηχητικού σήματος, στην αναζήτηση και εκπαίδευση του νευρονικού μοντέλου, στην συγγραφή τόσο των αναφορών και των παρουσιάσεων όσο και του κώδικα.

Ο **Αθανάσιος Αναγνώστου** συνέβαλε στην αναζήτηση και επιλογή τρόπου προσέγγισης του προβλήματος, στην αναζήτηση του συνόλου δεδομένων, στην αναζήτηση και επιλογή μετρικών αξιολόγησης, στην αναζήτηση και αξιολόγηση του μοντέλου, στην συγγραφή τόσο των αναφορών και των παρουσιάσεων όσο και του κώδικα.

Ο **Εμμανουήλ Χρήστος** συνέβαλε στην οργάνωση της ομάδας, στην αναζήτηση και επιλογή τρόπου προσέγγισης του προβλήματος, στην αναζήτηση και διαχείριση του συνόλου δεδομένων, στην αναζήτηση και επιλογή κατάλληλου μετασχηματισμού του ηχητικού σήματος, στην αναζήτηση και επιλογή μετρικών αξιολόγησης, στην αναζήτηση, εκπαίδευση και αξιολόγηση του νευρονικού μοντέλου, στην εκτέλεση πειραμάτων και καταγραφή αποτελεσμάτων, στην οργάνωση του υλικού, στην συγγραφή τόσο των αναφορών και των παρουσιάσεων όσο και του κώδικα.

Ο **Στέφανος Τσουκιάς** συνέβαλε στην οργάνωση της ομάδας, στην αναζήτηση και επιλογή τρόπου προσέγγισης του προβλήματος, στην αναζήτηση συνόλου δεδομένων, στην αναζήτηση του νευρονικού μοντέλου, στην οργάνωση του υλικού, στην συγγραφή τόσο των αναφορών και των παρουσιάσεων όσο και του κώδικα.