Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Μηχανολόγων Μηχανικών

Ανάπτυξη μοντέλου αναγνώρισης φωνητικών εντολών με χρήση τεχνητών νευρωνικών δικτύων.

Διπλωματική Εργασία

Άννα Μαρία Ιατρίδη

Επιβλέπων καθηγητής Πανώριος Μπενάρδος

Αθήνα 2024

Περίληψη

Στην παρούσα εργασία μελετάται η ανάπτυξη μοντέλου αναγνώρισης φωνητικών εντολών με χρήση τεχνητών νευρωνικών δικτύων, για βιομηχανικές εφαρμογές. Η βιομηχανική εφαρμογή είναι ο ρομποτικός βραχίονας Staubli RX 90L που βρίσκεται στο εργαστήριο του Τομέα Κατεργασιών στο Εθνικό Μετσόβιο Πολυτεχνείο. Το νευρωνικό δίκτυο σχεδιάστηκε για να αναγνωρίζει μονολεκτικές φωνητικές εντολές και να τις μετατρέπει σε γραπτό κείμενο, με σκοπό τον προγραμματισμό του ρομπότ. Η ανάπτυξη του μοντέλου διεπαφής υπολογιστή-ρομπότ είναι εκτός του φάσματος της εργασίας, όμως η προοπτική συνεργασίας του μοντέλου με το ρομπότ καθορίζει σε μεγάλο βαθμό τις προδιαγραφές του ίδιου του μοντέλου.

Το πρώτο στάδιο της εργασίας είναι ο καθορισμός του λεξιλογίου προς αναγνώριση. Αυτό καθορίζεται από τις εντολές της V+ και από την διαθεσιμότητα εντολών από το Google's Speech Commands Dataset. Η βιβλιοθήκη εντολών, περιέχει εντολές στην αγγλική γλώσσα, όπως τα αριθμητικά ψηφία από το μηδέν μέχρι το εννιά και άλλες μικρές και απλές λέξεις, όπως "*on*", "*off*", "*stop*" κ.α.

Για να γίνει η αναγνώριση των εντολών είναι απαραίτητη η σωστή προεπεξεργασία των ηχητικών σημάτων και η εξαγωγή των χαρακτηριστικών μαθηματικών παραμέτρων, των οποίων ο συνδυασμός οδηγεί στην αναγνώριση της εντολής. Η προετοιμασία του σήματος, πριν την αναγνώριση, προσομοιώνει το τρόπο πρόσληψης και ανάλυσης των ηχητικών σημάτων του ανθρώπινου αυτιού και εγκεφάλου. Η προεπεξεργασία περιέχει πρώτον και κύριον το στάδιο της αποθορυβοποίησης, όπου χρησιμοποιούνται φίλτρα pre-emphasis για να καθαρίσουν το σήματα από περιττή και άχρηστη πληροφορία. Το φίλτρο αυτό αποτελεί μία μαθηματική συνάρτηση υπολογισμού της διαφοράς διαδοχικών σημείων του σήματος με έναν συντελεστή. Η συνάρτηση αυτή μειώνει την συνολική ένταση των σημάτων, λειτουργώντας σαν ένα είδος κανονικοποίησης. Με αυτόν τον τρόπο τα σήματα έχουν περισσότερη ομοιογένεια.

Την αποθορυβοποίηση διαδέχεται ο διαχωρισμός του σήματος σε επιμέρους τμήματα, πριν το στάδιο υπολογισμού του φάσματος. Ο λόγος για τον κατακερματισμό του σήματος είναι ότι ο υπολογισμός του φάσματος συχνοτήτων στο σύνολο του σήματος χάνει πληροφορία για την χρονική εξάρτηση της συχνότητας. Αντίθετα ο υπολογισμός του φάσματος σε μικρότερα τμήματα του σήματος, διασφαλίζει την χρονική εξάρτηση της συχνότητας, ως πληροφορία που θα συμβάλλει στην αναγνώριση των εντολών. Η διάσπαση του σήματος σε μικρότερα δημιουργεί ασυνέχειες, οι οποίες οδηγούν σε διαρροές φάσματος. Η διαρροή φάσματος είναι όταν εμφανίζονται συχνότητες, καθ όλο το εύρος, οι οποίες δεν αντιστοιχούν σε πραγματική πληροφορία, αλλά σε ασυνέχειες. Η απαλοιφή των ασυνεχειών έρχεται σε σύγκρουση με την διακριτότητα του σήματος. Η συνάρτηση Hamming window, εξασφαλίζει απουσία διαρροών και ταυτόχρονα καλή διακριτότητα. Τώρα το σήμα είναι έτοιμο για την εφαρμογή του διακριτού μετασχηματισμού Fourier (DFT).

Η εξαγωγή των φασματικών συντελεστών της κλίματας Mel (MFCCs), αποτελεί το πιο καίριο βήμα για την αναγνώριση εντολών. Αρχικά, το σήμα μετασχηματίζεται από την κλίμακα συχνοτήτων στην κλίμακα των Mel. Η κλίμακα Mel είναι μια αντιληπτική κλίμακα συχνοτήτων με ισαπέχοντα διαστήματα συχνοτήτων που αντιλαμβάνονται ως ισαπέχουσες απ' το ανθρώπινο αυτί. Ο άνθρωπος δεν έχει την ίδια ευαισθησία σε όλες τις συχνότητες¨στις χαμηλές μπορεί και αναγνωρίζει πολύ εύκολα ακόμα και πολύ μικρές μεταβολές, ενώ στις υψηλότερες η αντιληπτικοτητα του μειώνεται και διαφορετικές συχνότητες τις αντιλαμβάνεται ως ίδιες ή παρεμφερείς. Για κάθε ένα από τα τμήματα, υπολογίζονται οι 12 φασματικοί συντελεστές. Οι φασματικοί συντελεστές λειτουργούν ως ταυτότητα των διαφορετικών φωνημάτων και καθιστούν δυνατή την διαφοροποίηση των ηχητικών λέξεων. Αυτοί αποτελούν την είσοδο του νευρωνικού δικτύου, για την κατηγοριοποίηση άγνωστων εντολές, σε γνωστές κλάσεις.

Για την αναγνώριση των φωνητικών εντολών, γίνεται χρήση τεχνητών νευρωνικών δικτύων αναγνώρισης μοτίβων. Η επιλογή κατάλληλου μοντέλου μηχανικής μάθησης είναι καίρια για την επιτυχημένη αναγνώριση των εντολών. Κατά την εκπόνηση της διπλωματικής δόθηκε μεγάλη έμφαση στην εύρεσης της βέλτιστης αρχιτεκτονικής νευρωνικού δικτύου, προς την επίτευξη της μέγιστης απόδοσης. Το τελικό νευρωνικό δίκτυο επιλέχθηκε με 600 κρυμμένους νευρώνες στο πρώτο επίπεδο και 450 στο δεύτερο. Η σύγκριση πολυπλοκοτερο αρχιτεκτονικών δεν κρίθηκε απαραίτητη, αλλά θα αποτελούσε ενδιαφέρουσα διερεύνηση. Το τελικό μοντέλο αναγνωρίζει 18 φωνητικές εντολές με

ακρίβεια 80%, υπό προϋποθέσεις. Η μέγιστη ακρίβεια εμφανίζεται όταν οι άγνωστες, προ αναγνώριση, λέξεις ανήκουν στο σύνολο Google’s Speech Commands Dataset. Οι εντολές που δίνονται από ανεξάρτητους ομιλητές αναγνωρίζεται με ακρίβεια κοντα στο 60%. Αυτό δείχνει σημάδια υπερ+εκπαίδευσης και αδυναμία γενίκευσης προβλέψεων.

Σε κάθε περίπτωση, η χρήση τεχνητών νευρωνικών δικτύων θεωρείται ανταγωνιστική μέθοδος στο κομμάτι της αναγνώρισης εντολών και με μικρές διορθώσεις μπορεί να φτάσει καλύτερες επιδόσεις.

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντα καθηγητή Πανώριο Μπενάρδο για την εξαιρετική συνεργασία μας και που μου εμπιστευτηκε ένα τόσο ενδιαφέρον θέμα, που υπήρξε η αφορμή να εμβαθύνω τις γνώσεις μου στο πεδίο της μηχανικής μάθησης και να βελτιώσω τον τρόπο σκέψης μου ως μηχανικός. Η καθοδήγηση και η υπομονή του υπήρξαν καθοριστικές, καθόλη την διάρκεια της εργασίας.

Θα ήθελα επίσης να ευχαριστήσω τους φίλους και συμφοιτητές μου Χρήστο, Στάθη και Αναστασία για την συνεχόμενη στήριξη και αγάπη τους απ' την πρώτη μέρα. Φυσικά, δεν θα μπορούσα να παραλείψω τους «συνεργούς μου στο έγκλημα» Κωστή και Στεφανία που έκαναν τις εργασίες της σχολής παραγωγικές και ευχάριστες. Είμαι ευγνώμων που μοιραστήκατε μαζί αυτό το ταξίδι.

Τέλος, ένα τεράστιο ευχαριστώ στους γονείς και την αδερφή μου για την ανιδιοτελή αγάπη τους και που πάντα με σπρώχνουν πέρα απ' τα όριά μου και στηρίζουν τα όνειρά μου.

Άννα Μαρία Ιατρίδη
Αθήνα, Ιούλιος 2024

Πίνακας Περιεχομένων

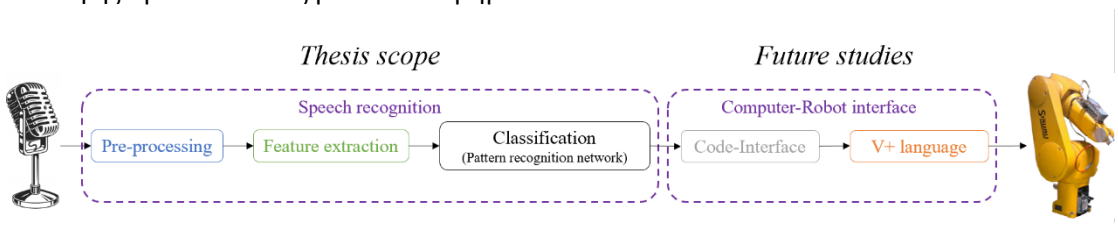
Κατάλογος Εικόνων

Κατάλογος Πινάκων

1. Εισαγωγή

1.1. Σκοπός της εργασίας

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη μοντέλου αναγνώρισης εντολών με χρήση τεχνητών νευρωνικών δικτύων στον βιομηχανικό τομέα. Η μελέτη εστιάζει στην εφαρμογή τεχνικών μηχανικής μάθησης για αναγνώριση μονολεκτικών εντολών. Οι εντολές επιλέγονται για εφαρμογή ρομποτικού βραχίονα, καθώς ο απώτερος σκοπός είναι η μελλοντική χρήση του μοντέλου σε συνεργασία με τον βραχίονα για την πραγματοποίηση βασικών εργασιών, όπως η προσέγγιση αντικειμένων και η μετακίνησή τους από ένα σημείου του χώρου σε άλλο κ.α.. Το ρομπότ είναι ο Staubli RX 90L, ένας αρθρωτός ρομποτικός βραχίονας έξι βαθμών ελευθερίας, που βρίσκεται στο εργαστήριο του Τομέα Κατεργασιών στο Εθνικό Μετσόβιο Πολυτεχνείο. Τέτοιου είδος ρομπτό είναι κατάλληλα για κατεργασίες συγκόλλησης, φινιρίσματος και διαδικασιών pick-and-place σε πολλές βιομηχανικές εφαρμογές. Το διάγραμμα ροής εργασίας παρουσιάζεται στην Εικόνα 1. Το κομμάτι αναγνώρισης ομιλίας αποτελεί αντικείμενο μελέτης της παρούσας εργασίας, ενώ το κομμάτι της διεπαφής προτείνεται ως μελλοντικό βήμα.



Εικόνα 1 – Διάγραμμα ροής εργασίας

Ο ρόλος του αντικειμένου αναγνώρισης ομιλίας είναι η μετατροπή των φωνητικών εντολών σε γραπτό κείμενο. Η είσοδος του συστήματος είναι το ηχητικό, καταγεγραμμένο σε πραγματικό χρόνο από τον χρήστη, ενώ η έξοδος είναι η αντίστοιχη εντολή γραπτή. Οι εντολές καταγράφονται με την χρήση απλού μικροφώνου, σαν αυτό του υπολογιστή ή του κινητού, επομένως δεν προαπαιτείται η χρήση ειδικού εξοπλισμού. Στην συνέχεια, το σήμα επεξεργάζεται μέσω φίλτρων και συναρτήσεων, μέχρι την εξαγωγή των φασματικών συντελεστών συχνότητας Mel (Mel Frequency Cepstral Coefficients – MFCCs). Οι συντελεστές MFCC περιέχουν σημαντικές πληροφορίες του ηχητικού σήματος για ταυτοποίηση και αποτελούν είσοδοι του τεχνητού νευρωνικού δικτύου, το οποίο αντιστοιχίζει την άγνωστη ηχητική καταγραφή με μία απ' τις γνωστές λέξεις/εντολές.

Συνοπτικά η μεθοδολογία είναι:

- Προ-επεξεργασία σήματος:
  - Αφαίρεση θορύβου και εξισορρόπηση συχνοτήτων
  - Διάσπαση του σήματος σε μικρότερα τμήματα
  - Υπολογισμός φάσματος συχνοτήτων
- Εξαγωγή χαρακτηριστικών:
  - Υπολογισμών των MFCCs
- Κατηγοριοποίηση:
  - Αναζήτηση βέλτιστου μοντέλου τεχνητής νοημοσύνης για ταυτοποίηση εντολών. (Εκπαίδευση)
  - Αναγνώριση εντολών με χρήση τεχνητών νευρωνικών δικτύων. (Αναγνώριση)

Το κριτήριο αποδοχής είναι το εκπαιδευμένο μοντέλο να έχει ακρίβεια μεγαλύτερη ή ίση με 80%. Στην σημερινή εποχή, τα υπερσύγχρον μοντέλα φτάνουν σφάλματα μικρότερα του 5%-10%. Επομένως, η ανάπτυξη μοντέλου ακρίβειας 80% είναι ένα καλό σημείο εκκίνησης, το οποίο μπορεί να βελτιωθεί περαιτέρω.

1.2. Κύριες προκλήσεις

Κατά την ανάπτυξη ενός μοντέλου αναγνώρισης ομιλίας, σε πολλά σημεία πρέπει να δοθεί ιδιαίτερη προσοχή. Το κλειδί για την σωστή και αποδοτική λειτουργία του συστήματος είναι η σωστή προεπεξεργασία των ηχητικών σημάτων. Είναι σημαντική η αναγνώριση και ταυτοποίηση της πληροφορίας των σημάτων που καταδεικνύει την προσφερόμενη λέξη, ώστε η αναγνώριση να είναι εφικτή και αποτελεσματική. Εάν το σήμα εμπεριέχει θόρυβο ή του λειπουν συγκεκριμένα χαρακτηριστικά, η απόδοση του μοντέλου αναγνώρισης ομιλίας θα είναι πολύ περιορισμένη. Είναι καίρια η εύρεση των χαρακτηριστών παραμέτρων που διαφοροποιούν τις λέξεις μεταξύ τους και η έκφρασή του ως μαθηματικές ποσότητες. Μόνο με κατάλληλη επεξεργασία και διαχείριση του λεξιλογίου είναι δυνατή και έγκυρη η διερεύνηση και ανάπτυξη του μοντέλου μηχανικής μάθησης. Θα πρέπει, επιπλέον, να ληφθεί υπόψιν, ότι ένα ηχητικό σήμα περιέχει πολλαπλές πληροφορίες, ως προς την ταυτότητα του ομιλητή, την συναισθηματική του κατάσταση, την προσφερόμενη λέξη κ.α.. Επομένως, η αναγνώριση πρέπει να εστιάζει σε συγκεκριμένες πληροφορίες, και να παραβλέπει τις υπόλοιπες λεπτομέρειες. Οι άνθρωποι, έχουν την δυνατότητα να αντιλαμβάνονται όλες τις πληροφορίες ταυτόχρονα, παραδείγματος χάρη ποιός μιλάει, τι λέει και με ποιόν τρόπο, ενώ οι εφαρμογές τεχνητής νοημοσύνης εστιάζουν σε ένα τασκ ανά μοντέλο.

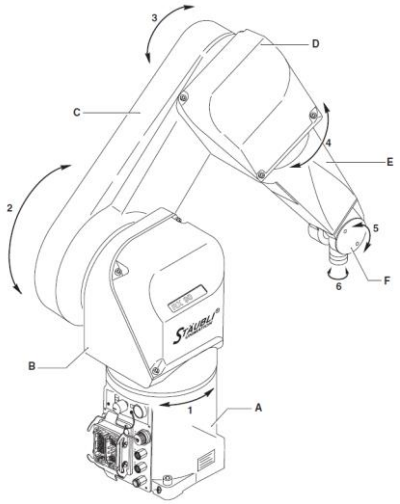
Μία πρόκληση κατά την αναγνώριση εντολών, είναι η ποικιλία και ανομοιομορφία της ανθρώπινης ομιλίας, και επιπλέων των καταγραφών. Οι άνθρωποι μιλούν σε διαφορετική ταχύτητα, με διαφορετική προφορά και διαφορετικά χαρακτηριστικά φωνής και ομιλίας. Αυτή η ποικιλομορφία πρέπει να αποτυπώνεται στο σετ δεδομένων εκπαίδευσης, ώστε το μοντέλο να επιτυγχάνει την αναγνώριση των λέξεων παρά τα διαφορετικά χαρακτηριστικά. Στις εφαρμογές αναγνώρισης μονολεκτικών εντολών όπου η κάθε λέξη καταγράφεται σε συγκεκριμένη χρονική διάρκεια τα ηχητικά σήματα μπορούν να διαφέρουν αρκετά. Το φώνημα μπορεί να αποτυπωθεί σε διαφορετικό χρονικό σημείο του συνολικού σήματος και να καταλαμβάνει διαφορετική διάρκεια ανάλογα με την ταχύτητα ομιλίας. Ειδικότερα, με την διάσπαση του σήματος σε επιμέρους μέρη, η πολυπλοκότητα αυξάνεται, εφόσον κάθε τμήμα μιας καταγραφής διαφέρει πολύ απ' το αντίστοιχο κάποιας άλλης. Το δίκτυο πρέπει να είναι εκπαιδευμένος να αναγνωρίζει τα μοτίβα, παρά τις διαφοροποιήσεις, και τα καταλήγει στο σωστό συμπέρασμα.

Για την πρόβλεψη με υψηλή ακρίβεια, εκτός από την επιλογή κατάλληλης μεθόδου επεξεργασίας, είναι καίρια η διαμόρφωση κατάλληλου σετ δεδομένων. Το «καλό» σετ σημαίνει αντιπροσωπευτικές καταγραφές και επαρκή αριθμό δεδομένων. Η απαίτηση για αντιπροσωπευτικές, όπως περιγράφηκε προηγουμένως, συνεπάγεται καταγραφές από ποικίλους ομιλητές, με διαφορετικές ταχύτητες ομιλίας και προσφορές. Ένα φτωχό σύνολο έχει ως αποτέλεσμα οι προβλέψεις να είναι πολύ ευαίσθητες σε λεπτομέρειες, μειώνοντας την δυνατότητα γενίκευσης σε διαφορετικά σύνολα. Ο αριθμός των καταγραφών που απαιτούνται, εξαρτάται από την φύση και πολυπλοκότητα του προβλήματος και από των αριθμό των εισόδων. Το νευρωνικό δίκτυο λειτουργεί όπως ένα σύστημα εξισώσεων, όπου χρειάζονται επαρκή δεδομένα για τον σωστό προσδιορισμό των άγνωστων μεταβλητών (εδώ: προσδιορισμός των βαρών). Η αναγνώριση ομιλίας είναι αρκετά πολύπλοκο πρόβλημα, επομένως απαιτεί εν γένει μεγάλα σύνολα για εκπαίδευση. Επιπλέον, πρέπει να επισημανθεί ότι οι είσοδοι του συστήματος είναι οι 12 συντελεστές MFCCs για κάθε ένα τμήμα του σήματος, από τα 98. Είναι σημαντική η εξασφάλιση ικανοποιητικού συνόλου, για την ομαλή λειτουργία του συστήματος.

2. Ρομποτικοί βραχίονες

2.1. Staubli RX 90L

Ο ρομποτικός βραχίονας Staubli RX 90L είναι ένα αρθρωτό ρομπότ, έξι βαθμών ελευθερίας, κατασκευασμένος από την ομώνυμη εταιρεία Stäubli. Κάθε ένας από τους συνδέσμους, λειτουργεί ως άξονας, γύρω από τον οποίο περιστρέφεται δύο μέλη. Η κίνηση των αρθρώσεων του ρομπότ ελέγχεται με χρήση κινητήρων χωρίς ψήκτρες, συζευγμένους με γωνιοαναλυτές, εξοπλισμένοι με φρένα. Το ρομπότ απαρτίζεται από κινητήρες, φρένα, μηχανισμούς ελέγχου της κίνησης και πνευματικά και ηλεκτρικά κυκλώματα. Η απόλυτη τοποθεσία στον χώρο καταγράφεται και δίνεται από μετριτικό σύστημα ανά πάσα στιγμή. Ο μηχανισμός είναι αξιόπιστος και στιβαρός, ευέλικτος και με δυνατότητα εκπόνησης πολλών καθηκόντων. Το ρομπότ χρησιμοποιείται συνήθως για φινίρισμα σε πολλές βιομηχανικές εφαρμογές, όπως στα πλαστικά και μεταλλικά μέρη μηχανών, σε ποδήλατα και σε αγροτικές εφαρμογές. Τα βασικά χαρακτηριστικά ενός ρομποτικού βραχίονα είναι εμπνευσμένα από το ανθρώπινο χέρι: (A), shoulder (B), arm (C), elbow (D), forearm (E) και wrist (F, όπως απεικονίζονται στην Εικόνα 2



Εικόνα 2 – Staubli RX 90L

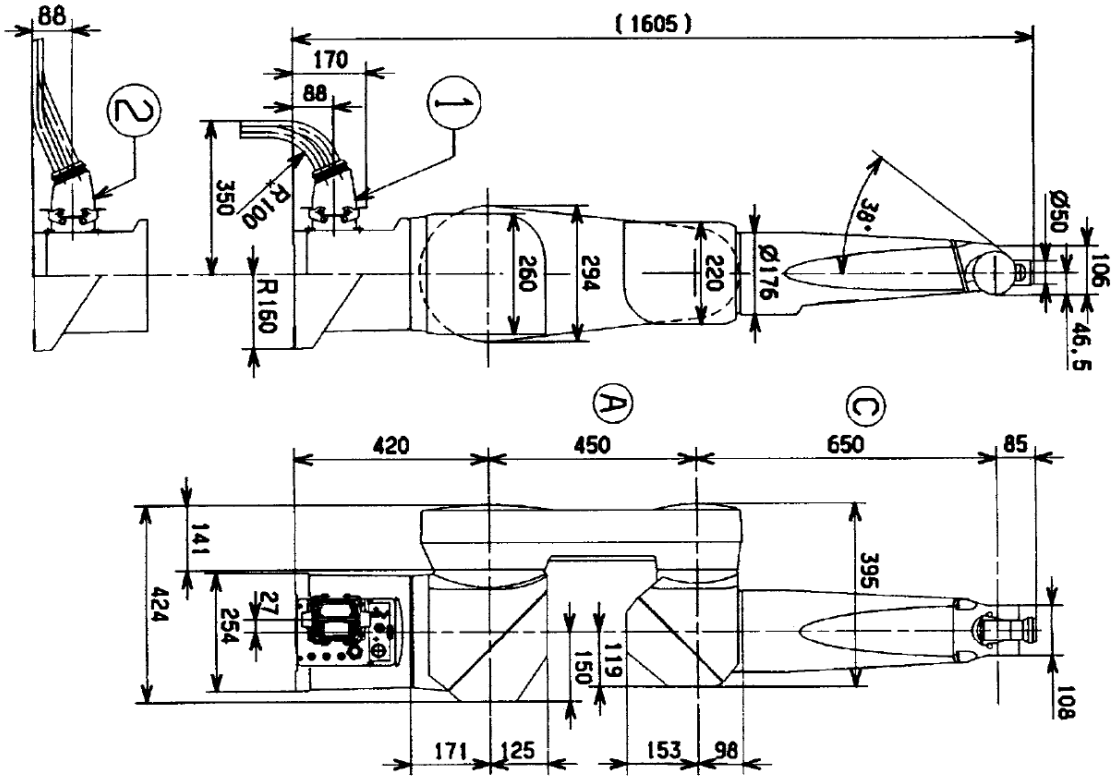
Όλα τα τεχνικά χαρακτηριστικά και οι αριθμητικές τιμές τους, παρουσιάζονται παρακάτω, όπως φαίνονται και στον οδηγό χρήσης απ' την Staubli [1].

Πίνακας 1– Χαρακτηριστικά Staubli RX 90L

Designation: RX 90 B L	Robot family	RX (changed to B)
	Maximum reach between 2nd and 5th axis[1]	9 dm
	Number of active axis (≡ DoF)	0 ≡ 6 (variation with 5 axis)
	Forearm version	Extended forearm (L)
General characteristics	Working temperature	+5℃ to +40℃
	Humidity	30% to 90%
	Altitude	2000 m
	Weight	113 kg
Performance	Maximum speed at load center of gravity	12.6 $^{m}/_{s}$
	Repeatability	$\pm$0.025 mm
Load capacity	At nominal speed	3.5 kg
	At reduced speed	6 kg

[1] *That is the reach for the original - not extended – version. With the longer forearm it becomes* 11 dm *instead of* 9 dm.

Στην συνέχεια, είναι ένα κατασκευαστικό σχέδιοτου ρομποτικού βραχίονα με όλες τις κύριες διαστάσεις.



Εικόνα 3 – Staubli RX 90L

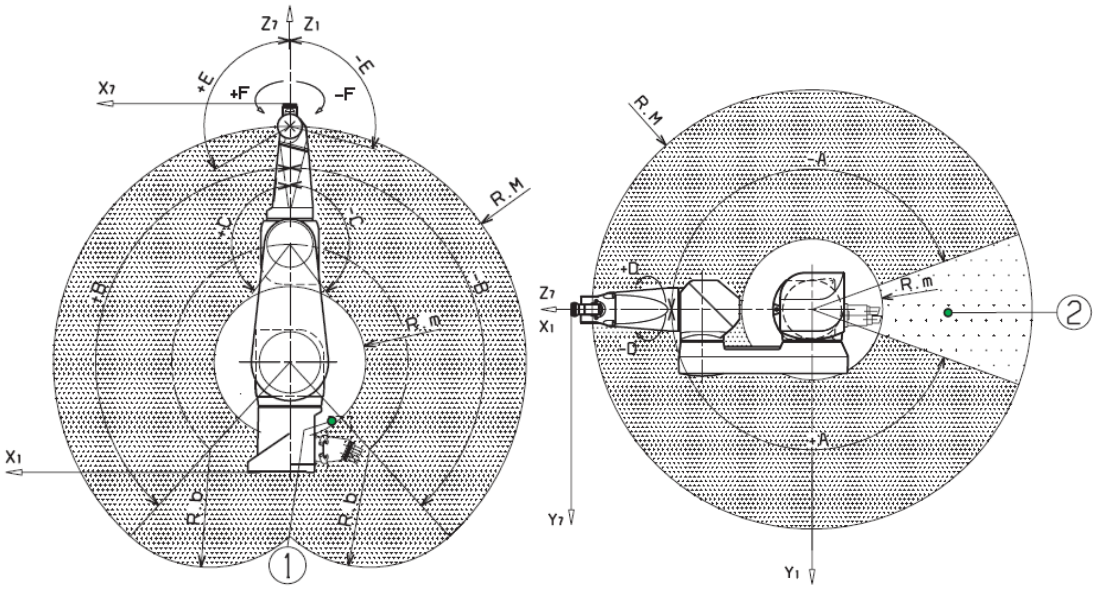
Το πεδίο λειτουργίας ορίζεται ως ο χώρος εργασίας, όπου η αρπάγη του ρομπότ μπορεί να βρεθεί με οποιοδήποτε προσανατολισμό και εξαρτάται από τις διαστάσεις των μελών. Οι βασικές παράμετροι συνοψίζονται στους πίνακες και καταδεικνύονται στις εικόνες.

Πίνακας 2 – Staubli RX 90L πεδίο εργασίας

Parameter	Symbol	Value
Maximum reach between 2nd and 5th axis	$R.M.$	$1100\ mm$
Minimum reach between 2nd and 5th axis	$R.m.$	$401\ mm$
Reach between 3rd and 5th axis	$R.b.$	$650\ mm$

Πίνακας 3 – Staubli RX 90L εύρος, πλάτος, ταχύτητα

Axis	1	2	3	4	5	6
Amplitude (°)	320	275	285	540	225	540
Working range (°)	A ± 160	B ± 137.5	C ± 142.5	D ± 270	E + 120 − 105	F ± 270
Nominal speed $(^{\circ}/_{s})$	236	200	286	401	800	1125
Maximum speed $(^{\circ}/_{s})$	356	356	296	409	800	1125
Resolution $(^{\circ} \cdot 10^{-3})$	0.87	0.87	0.72	1	1.95	2.75

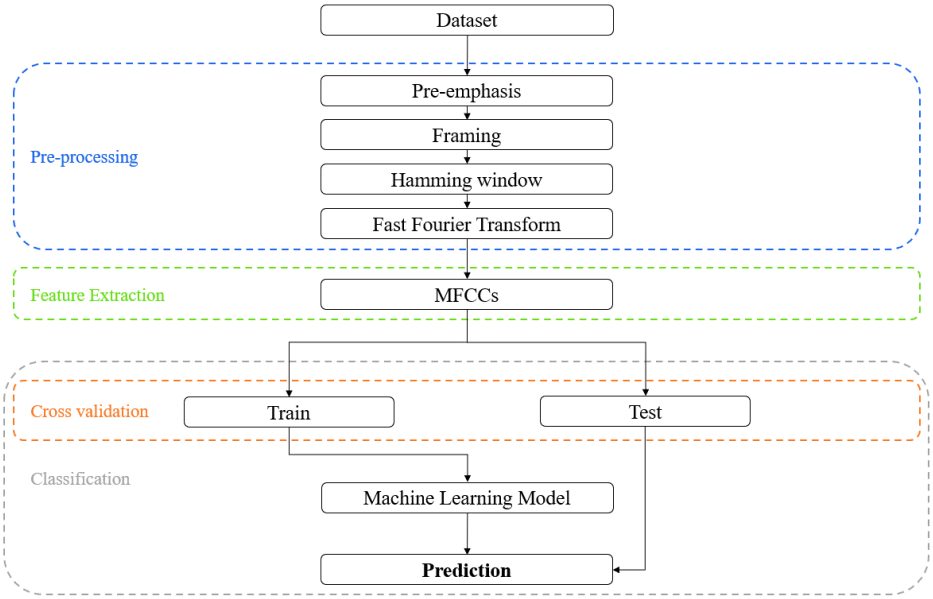


Εικόνα 4 – Staubli RX 90L χώρος εργασίας

3. Methodology Speech Recognition

Η παρούσα εργασία αφορά την ανάπτυξη ενός μοντέλου αναγνώρισης ομιλίας, για την κατανόηση προφορικών εντολών και τη μετάφρασή τους σε γραπτό κείμενο. Οι εντολές εισόδου είναι ηχογραφήσεις μεμονωμένων λέξεων και όχι συνεχής φυσική ομιλία. Δεδομένου ότι η εφαρμογή προορίζεται για βιομηχανικό περιβάλλον, είναι σημαντικό τα χαρακτηριστικά που εξάγονται να μην είναι ευαίσθητα στο περιβάλλον, στο θόρυβο του περιβάλλοντος ή στις αναντιστοιχίες των μικροφώνων.

Η διαδικασία αναγνώρισης ομιλίας περιλαμβάνει όλα τα βήματα από την καταγραφή του φωνητικού σήματος μέχρι την ταξινόμηση των εντολών με βάση το δεδομένο σύνολο δεδομένων/λεκτολόγιο. Η προεπεξεργασία είναι το πρώτο βήμα, όπου το σήμα απομονώνεται από το θόρυβο ή τις περιττές πληροφορίες και μειώνεται το μέγεθός του. Η εξαγωγή χαρακτηριστικών είναι το στάδιο όπου από το προ+επεξεργασμένο σήμα εξάγονται συγκεκριμένες παράμετροι, ενδεικτικές για το περιεχόμενο του σήματος. Αυτές οι παράμετροι αποτελούν τα σημαντικά χαρακτηριστικά των ηχητικών σημάτων και χρησιμοποιούνται από το μοντέλο μηχανικής μάθησης, στο στάδιο της ταξινόμησης, για την αντιστοίχιση των ηχογραφήσεων στις αντίστοιχες εντολές. Η αλληλουχία των βημάτων παρουσιάζεται στην Εικόνα 5 .



Εικόνα 5 – Μεθοδολογία εργασίας [2]

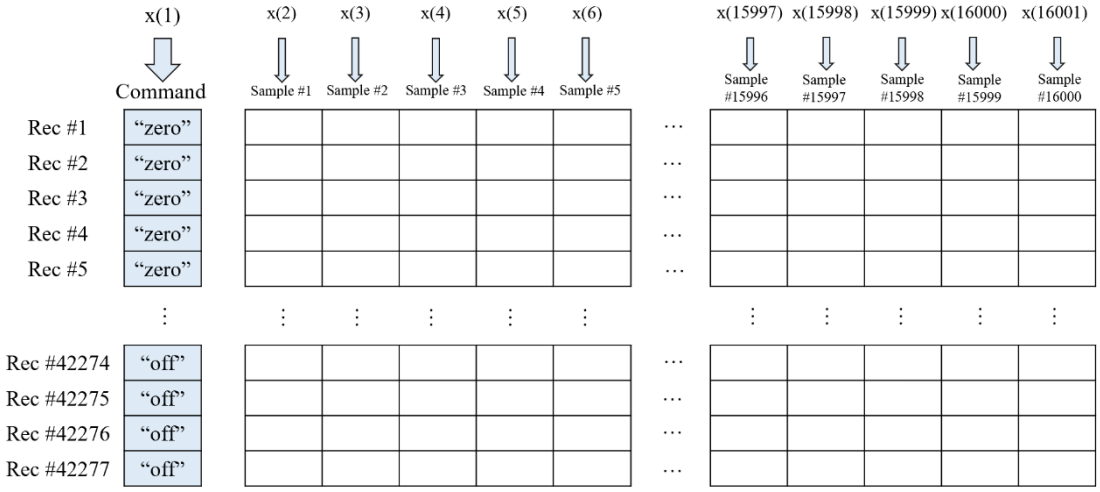
4. Αποτελέσματα και Ανάλυση

Στο προηγούμενο κεφάλαιο εξηγήθηκε και αναλύθηκε λεπτομερώς η μεθοδολογία. Στο παρόν κεφάλαιο παρουσιάζονται τα πρακτικά στοιχεία της παρούσας μελέτης, συμπεριλαμβανομένης της εφαρμογής της μεθόδου και της ανάλυσης των αποτελεσμάτων. Εδώ συζητούνται τα ευρήματα από τη διερεύνηση του ANN και την αξιολόγηση του τελικού δικτύου. Επιπλέον, περιλαμβάνονται αποτελέσματα από συμπληρωματικές μελέτες, ώστε να εξεταστεί η επίδραση διαφορετικών τεχνικών προεπεξεργασίας. Για παράδειγμα, αξιολογείται η αναγνώριση χωρίς προ-έμφαση και με διαφορετικές συναρτήσεις window.

4.1. Εκπαίδευση

4.1.1. Προ-επεξεργασία

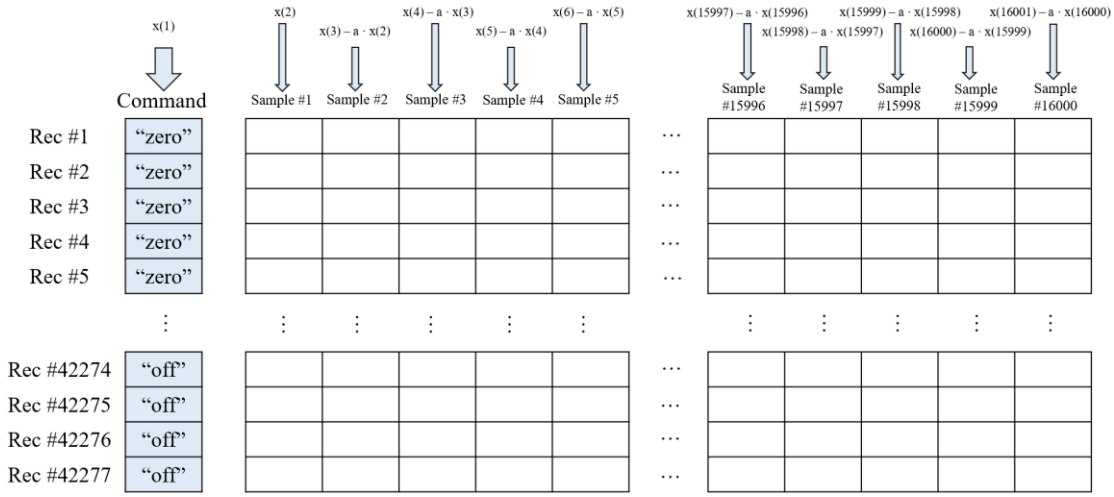
Το σύνολο των καταγραφών, αποθηκεύεται σε πίνακα 42277x16000, όπου ο αριθμός τον σειρών είναι ο αριθμός όλων των παρατηρήσεων, ενώ ο αριθμός των στηλών είναι ίσς με τον αριθμό των πλαισίων.Η προ-επεξεργασία έγινε σε Python.



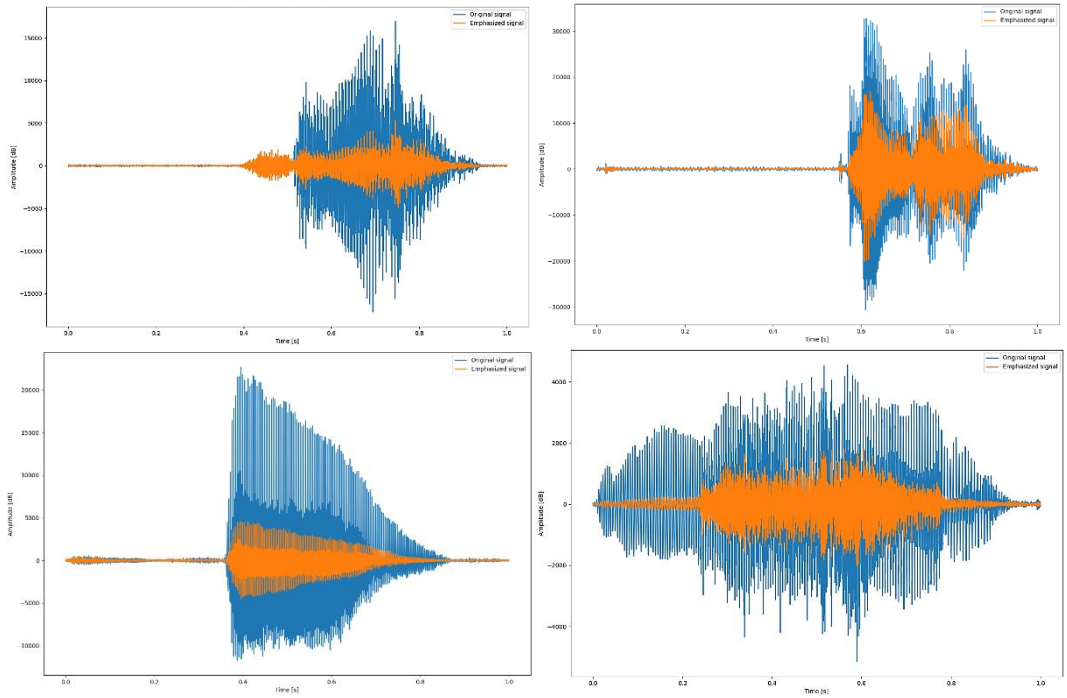
Εικόνα 6 – Αρχικός πίνακας [42277x16001]

4.1.1.1. Pre-emphasis

$$y(n) = x(n) - a \cdot x(n-1), \qquad n \subseteq [1, N]$$

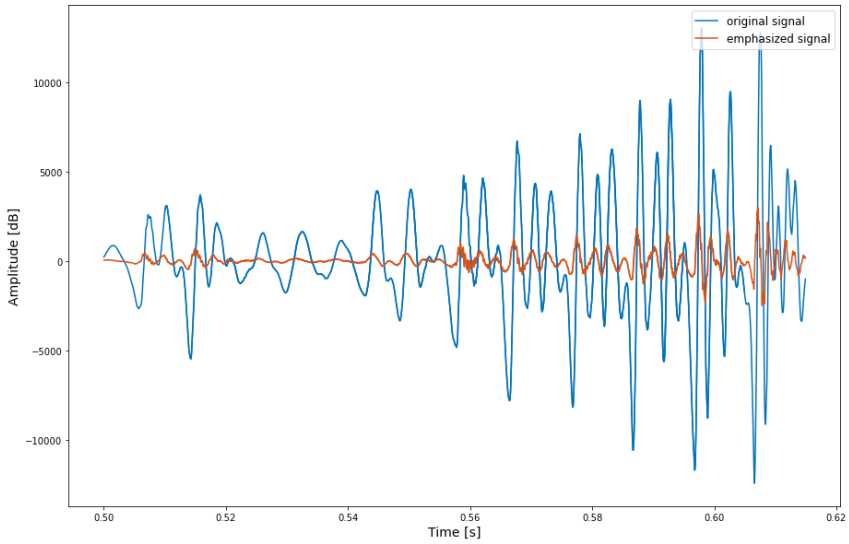


Εικόνα 7 – Πίνακας μετά το pre-emphasis



Εικόνα 8 – Φίλτρο pre-emphasis

Το φίλτρο προ-έμφασης εξισορροπεί το φάσμα συχνοτήτων, ενισχύοντας περισσότερο τις υψηλότερες συχνότητες και αποδυναμώνοντας τις χαμηλότερες. Ταυτόχρονα, απομακρύνει αποτελεσματικά το θόρυβο υποβάθρου, ειδικά όταν αυτός παράγεται από μια σταθερή πηγή. Όπως αναφέρθηκε προηγουμένως, το πλάτος του σήματος μετά την προ-έμφαση μειώνεται περίπου κατά το ήμισυ. Στην πραγματικότητα, το φίλτρο έχει αποτέλεσμα παρόμοιο με την κανονικοποίηση, φέρνοντας τα σήματα στο ίδιο εύρος πλάτους. Με αυτόν τον τρόπο το μοντέλο γίνεται λιγότερο ευαίσθητο στην ένταση της ομιλίας. Εάν μια λέξη προφέρεται πιο δυνατά στην πλειονότητα των ηχογραφήσεων, το μοντέλο θα συνδέσει την ένταση με αυτή τη λέξη και όταν λαμβάνει ακουστικά υψηλής έντασης θα τείνει να τα ταξινομεί ως αυτή τη λέξη. Το φίλτρο προ-έμφασης μειώνει αυτόν τον κίνδυνο, βοηθώντας το μοντέλο να εστιάσει στα σωστά χαρακτηριστικά της ομιλίας για να ταξινομήσει τις λέξεις.



Εικόνα 9 – Εφέ pre-emphasis

4.1.1.2. Framing

Οι παράμετροι διάσπασης του σήματος είναι το μήκος πλαισίου και το μήκος επικάλυψης. Το πλαίσιο πρέπει να είναι αρκετά σύντομο ώστε να θεωρείται χρονικά σταθερό, αλλά και αρκετά μεγάλο ώστε να καταγράφει όλα τα σημαντικά χαρακτηριστικά της ομιλίας. Η επιλογή λίγων μεγάλων πλαισίων ελαχιστοποιεί τα οφέλη της κατάτμησης, αλλά απαιτεί λιγότερο χώρο και υπολογιστική ισχύ. Από την άλλη πλευρά, η επιλογή πολύ μικρών πλαισίων αυξάνει την ευαισθησία σε μικρές ηχητικές διακυμάνσεις που μπορεί να είναι ακόμη και τεχνητές και απαιτεί πολύ περισσότερο χώρο. Ένας βραχνάς για την επιλογή πλαισίων είναι ο λόγος δειγματοληψίας. Η ανάλυση του σήματος, η ποσότητα των σημείων δεδομένων, εξαρτάται από τη συχνότητα δειγματοληψίας. Θα πρέπει να είναι σαφές ότι η τμηματοποίηση σε μικρότερα πλαίσια δεν αυξάνει την ανάλυση της ηχογράφησης. Τα επικαλυπτόμενα καρέ μειώνουν την απώλεια πληροφοριών για τη διάσπαση του σήματος και καλύπτουν τα χαρακτηριστικά που διαστρεβλώνονται ή χάνονται στα τελικά σημεία του καρέ. Εάν η επικάλυψη μεταξύ των καρέ είναι πολύ μεγάλη, τότε τα περισσότερα σημεία του δείγματος θα ληφθούν υπόψη πάρα πολλές φορές. Αυτό πιθανότατα δεν θα θέσει σε κίνδυνο τη συνολική ακρίβεια, αλλά θα τροφοδοτήσει το δίκτυο με πολλά περιττά δεδομένα. Από τη βιβλιογραφική ανασκόπηση, οι συνιστώμενες τιμές για το μήκος πλαισίου και την επικάλυψη είναι μεταξύ 15-25 ms και 10-15 ms, αντίστοιχα. Το μήκος πλαισίου επιλέγεται να είναι 25 ms, για να καταλαμβάνει λιγότερη μνήμη, και η επικάλυψη 15 ms.

Για συχνότητα δειγματοληψίας $f_s = 16000\,Hz$ μήκος πλαισίου και επικάλυψης είναι:

$$frame\_length = frame\_size \cdot sample\_rate \Rightarrow frame\_length = 400\ δείγματα$$

$$frame\_step = frame\_stride \cdot sample\_rate \Rightarrow frame\_step = 160\ δείγματα$$

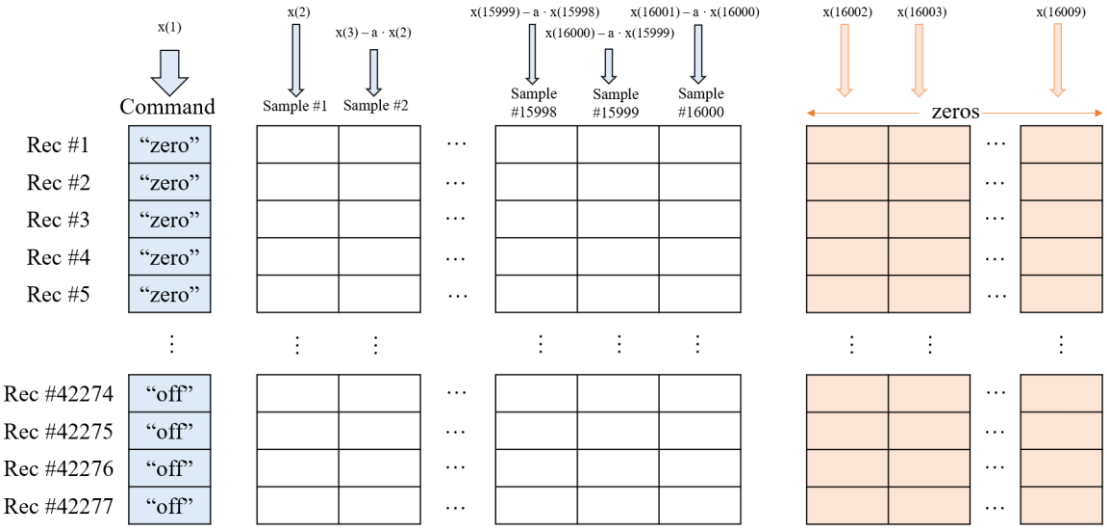
Ο αριθμός πλαισίων ανά σήμα είναι:

$$num\_frames = round\left(\frac{signal\_length - frame\_length}{frame\_step}\right) = 98\ πλαίσια$$

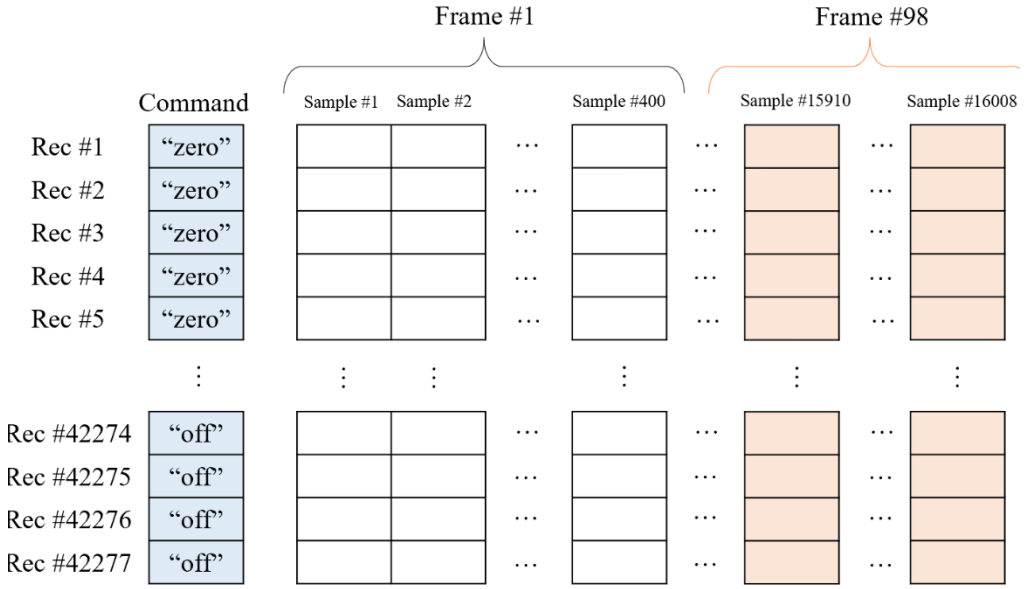
Σε περίπτωση που το ηχητικό σήμα δεν μπορεί να διαιρεθεί σε ακέραιο αριθμό πλαισίων, το τελευταίο πλαίσιο θα έχει μικρότερο μήκος ή θα πρέπει να προστεθούν μηδενικές τιμές στο αρχικό σήμα. Για να έχουμε τέλεια ακέραια διαίρεση, το σήμα πρέπει να έχει μήκος ίσο με:

$$pad\_signal = num\_frames \cdot frame\_step + frame\_length \Rightarrow$$
$$\Rightarrow pad\_signal = 16080\ δείγματα$$

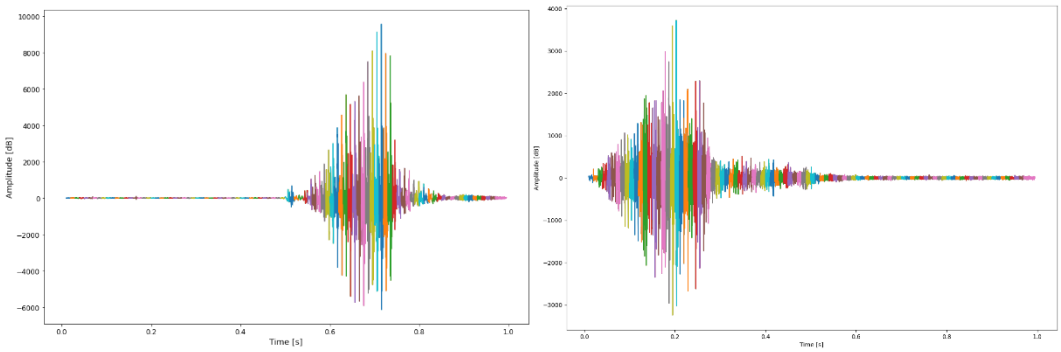
80 σειρές με μηδενικές τιμες προτίθενται στο τέλος του σήματος.



Εικόνα 10 – Pad signal



Εικόνα 11 – Σύνολο δεδομένων μετά την τμηματοποίηση



Εικόνα 12 – Διαχωρισμός σε πλαίσα

Είναι προφανές ότι για τις διάφορες εγγραφές τα καρέ αντιπροσωπεύουν διαφορετικό μέρος της εντολής. Για παράδειγμα, το καρέ με αριθμό 30 στην αριστερή εικόνα δεν περιλαμβάνει κανένα μέρος του φωνήματος, ενώ στη δεξιά εικόνα βρίσκεται στη μέση του.

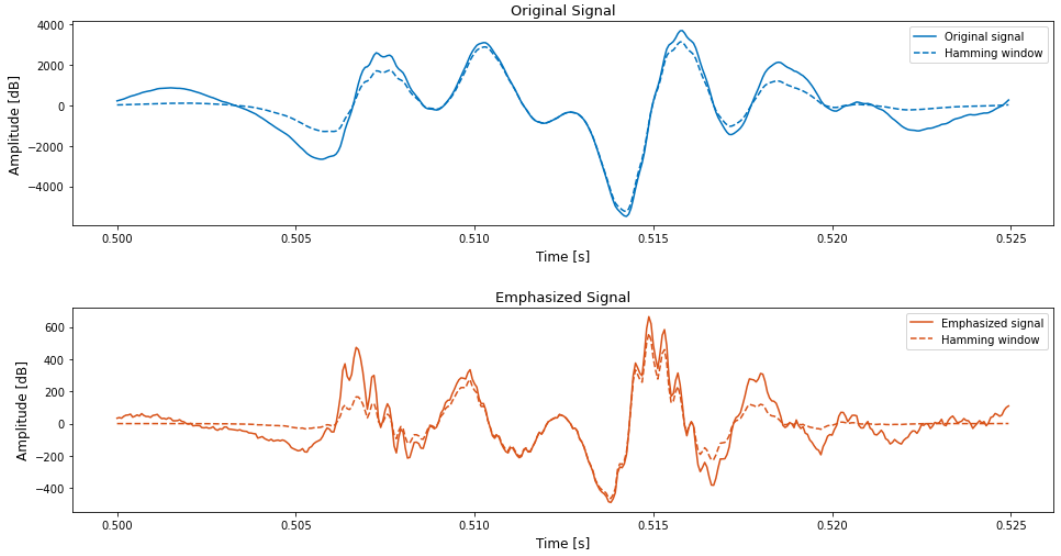
4.1.1.3. Hamming Window

Η συνάρτηση παραθύρου Hamming εφαρμόζεται σε κάθε πλαίσιο για να βελτιωθεί η μετάβαση μεταξύ των πλαισίων, να μειωθούν οι διαρροές φάσματος και να βελτιωθεί η ανάλυση συχνότητας. Η συνάρτηση Hamming window είναι η ακόλουθη:

$$w(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2 \cdot \pi \cdot n}{N-1}\right), 0 \leq n \leq N-1$$

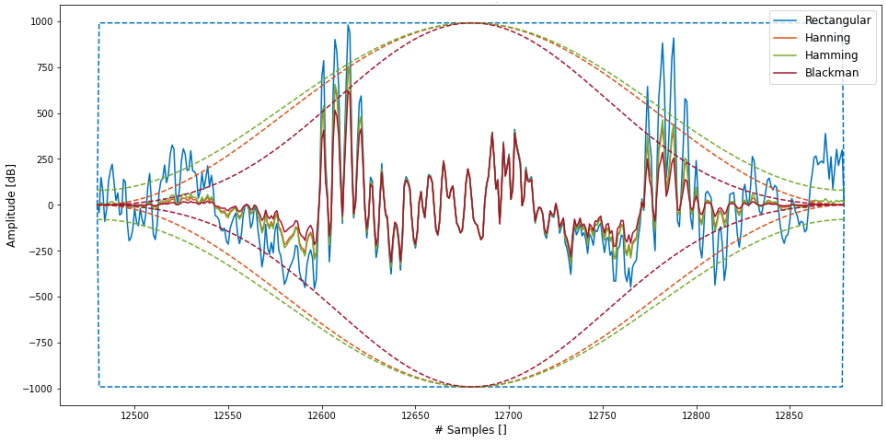
Μετά την εφαρμογή της συνάρτησης, ο κύριος στόχος είναι να έχουμε την ίδια τιμή στα άκρα κάθε καρέ. Αυτό είναι σημαντικό διότι τα πλαίσια αντιμετωπίζονται από τον FFT ως περιοδικά και εάν τα άκρα δεν ταυτίζονται τότε αυξάνεται η ασυνέχεια, η οποία οδηγεί σε διαρροές φάσματος. Οδηγώντας τις τιμές των τελικών σημείων στο μηδέν, η πληροφορία που περιέχεται σε αυτό το τμήμα του σήματος χάνεται, αλλά χάρη στα επικαλυπτόμενα τμήματα περιλαμβάνεται στο επόμενο ή στο προηγούμενο πλαίσιο. Στην ακόλουθη εικόνα (Σχήμα 12) παρουσιάζεται η επίδραση του παραθύρου που εφαρμόζεται σε ένα πλαίσιο α. του αρχικού και β. του τονισμένου σήματος.

included in the next or previous frame. In the following image (**Error! Reference source not found.**) is shown the effect of window applied on a frame a. of the original and b. of the emphasized signal.



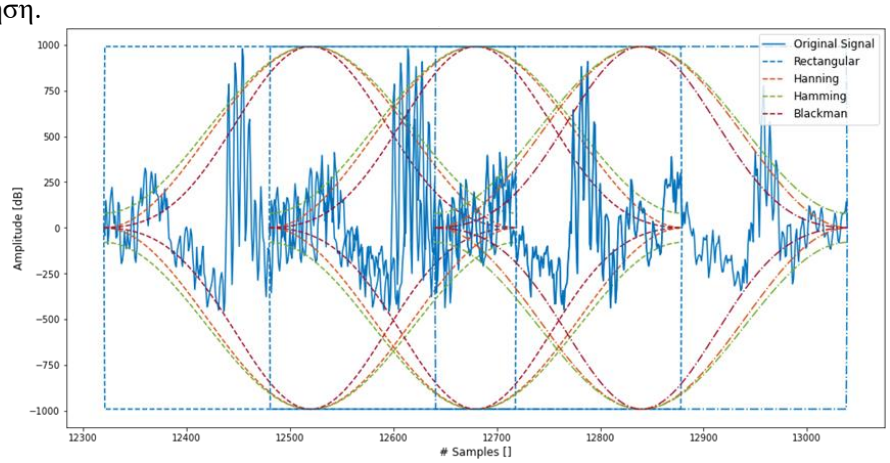
Εικόνα 13 – Συναρτήση Hamming window σε pre-emphasized και μησήμα.

Είναι ενδιαφέρον να βρεθεί πώς άλλες λειτουργίες Window τροποποιούν το ηχητικό σήμα. Στην Εικόνα 13 παρουσιάζεται η αναπαράσταση των διαφόρων συναρτήσεων και η τελική μορφή του σήματος. Όλες, εκτός από το παράθυρο Hamming, έχουν μηδενικό πλάτος στα άκρα του πλαισίου και μέγιστο στη μέση.



Εικόνα 14 – Συναρτήσεις Windowing

Στο Εικόνα 14παρουσιάζεται το φίλτρο παραθύρου που εφαρμόζεται σε τρία διαδοχικά καρέ. Είναι ορατό ότι τα τμήματα του σήματος, των οποίων το πλάτος ελαχιστοποιείται σε ένα καρέ, μεγιστοποιούνται στο προηγούμενο ή στο επόμενο, διατηρώντας όλες τις απαραίτητες πληροφορίες για την ταξινόμηση.



Εικόνα 15 – Συναρτήσεις Windowing

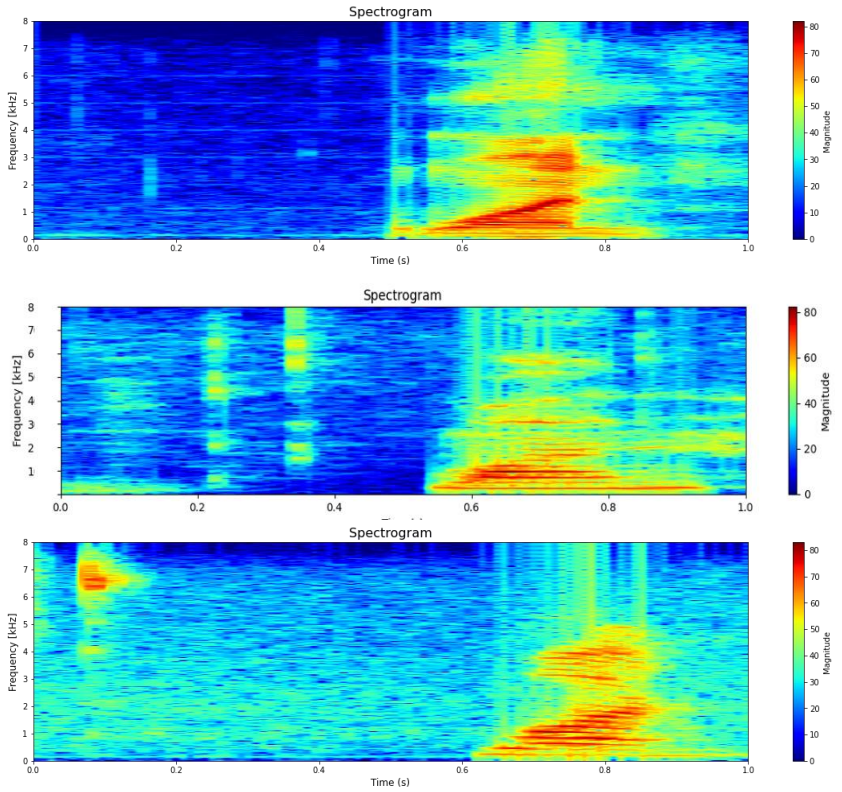
4.1.1.4. Fast Fourier Transform (FFT)

Η ολοκλήρωση της προ-επεξεργασίας τελειώνει με την εφαρμογή του και υπολογίζεται το φάσμα συχνοτήτων. Ο αριθμός των σημείων για τον μηχανισμό από την βιβλιοραφία είναι 256 ή 512; Εδώ επιλέγεται 512.

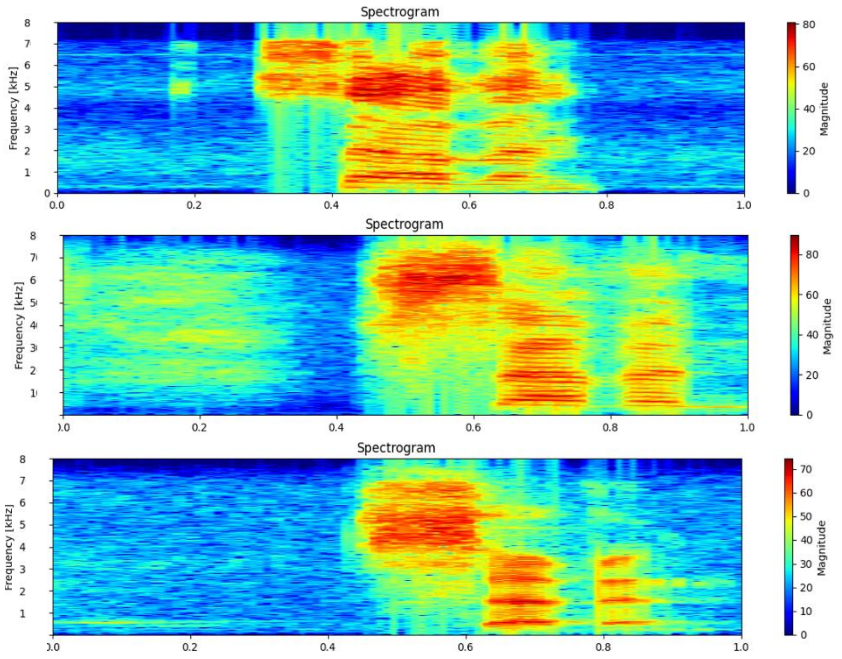
$$X[n] = \sum_{i=0}^{N-1} x[n] \cdot e^{-j \cdot \frac{2\pi}{N} \cdot k \cdot n}$$

Στα σχήματα που ακολουθούν παρουσιάζονται μερικά παραδείγματα φάσματος για διάφορες εγγραφές. Στον κατακόρυφο άξονα είναι η συχνότητα [kHertz], στον οριζόντιο είναι ο χρόνος [sec] ή ο αριθμός του δείγματος και το χρώμα αντιπροσωπεύει διαφορετικά μεγέθη. Η θέση σε σχέση με τον

οριζόντιο άξονα δείχνει τη χρονική στιγμή που εκφωνήθηκε η λέξη. Το μοτίβο του χρωματικού διαγράμματος δείχνει την κατανομή της ενέργειας μεταξύ διαφορετικών συχνοτήτων.



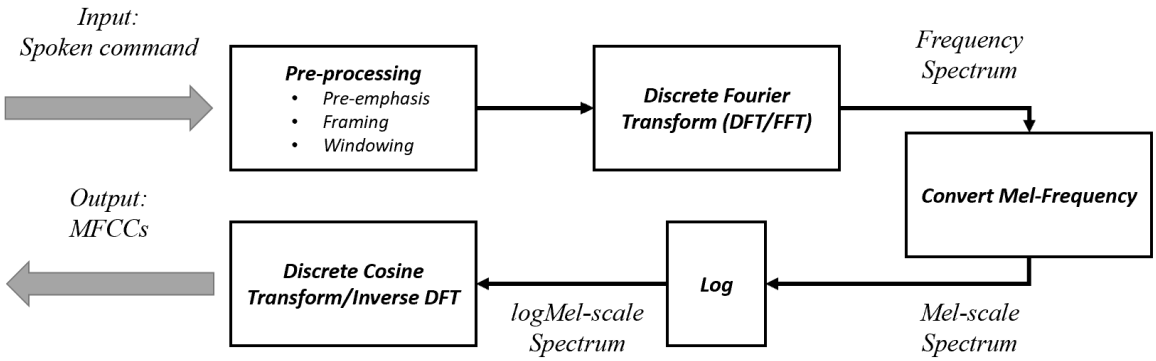
Εικόνα 16 – Φάσμα συχνοτήτων "one"



Εικόνα 17 – Φάσμα συχνοτήτων "seven"

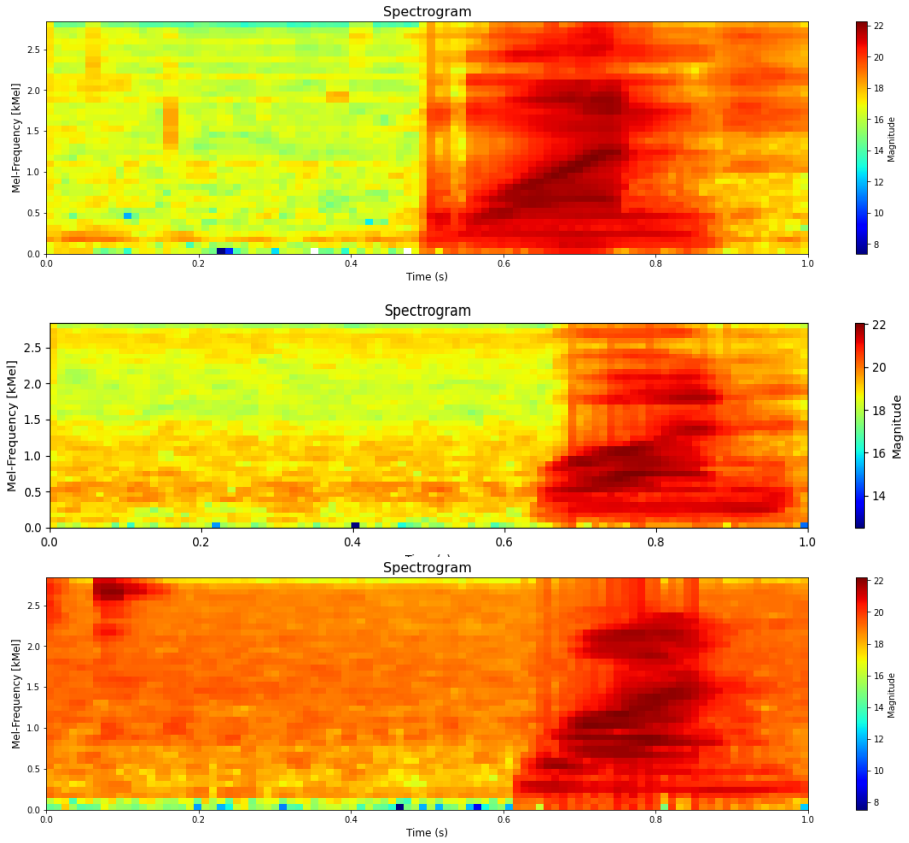
4.1.2. Εξαγωγή χαρακτηριστικών

Η διαδικασία εξαγωγής των MFCCs, παρουσιάζεται στην Εικόνα 18.

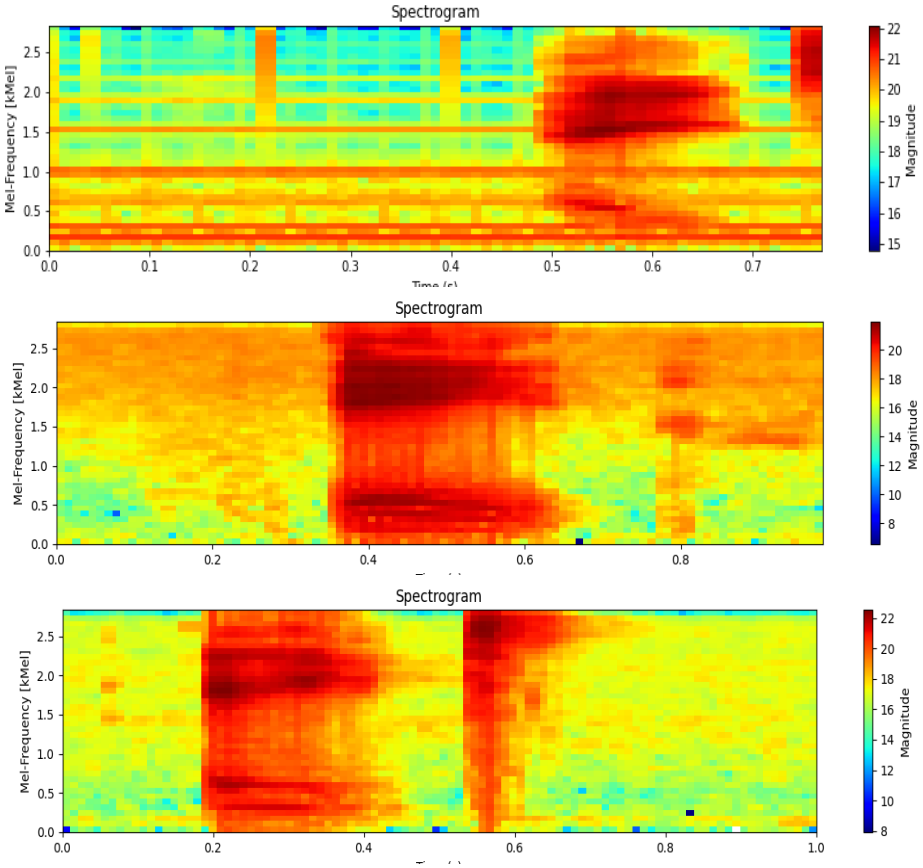


Εικόνα 18 – Διαδικασία υπολογισμού των MFCCs [3]

Αρχικά, η συχνότητα μετατρέπεται σε κλίμακα Mel. Όπως φαίνεται στις παρακάτω εικόνες, οι διαφορετικές καταγραφές της ίδιας εντολής έχουν παρόμοια απεικόνιση.

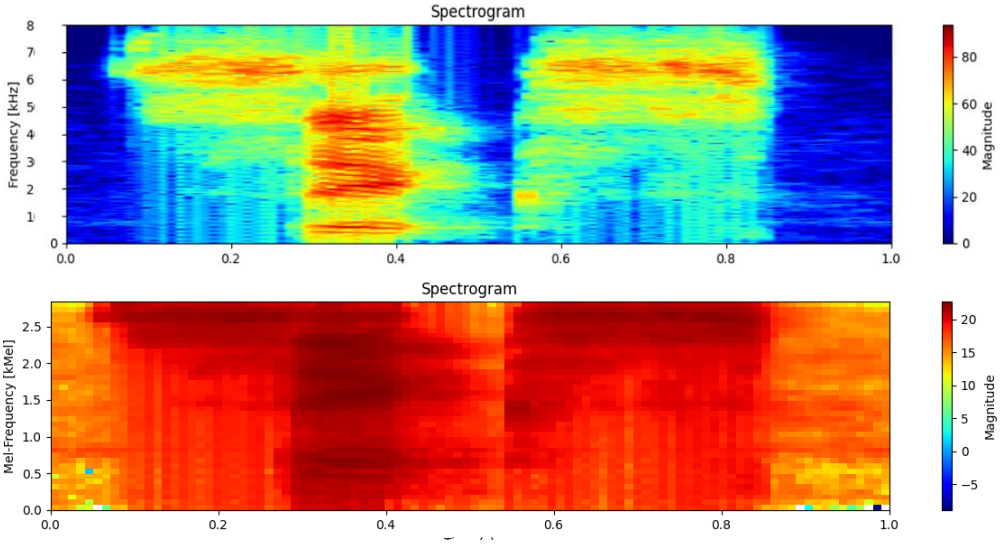


Εικόνα 19 – Φάσμα κλίμακας Mel- Εντολή "one"



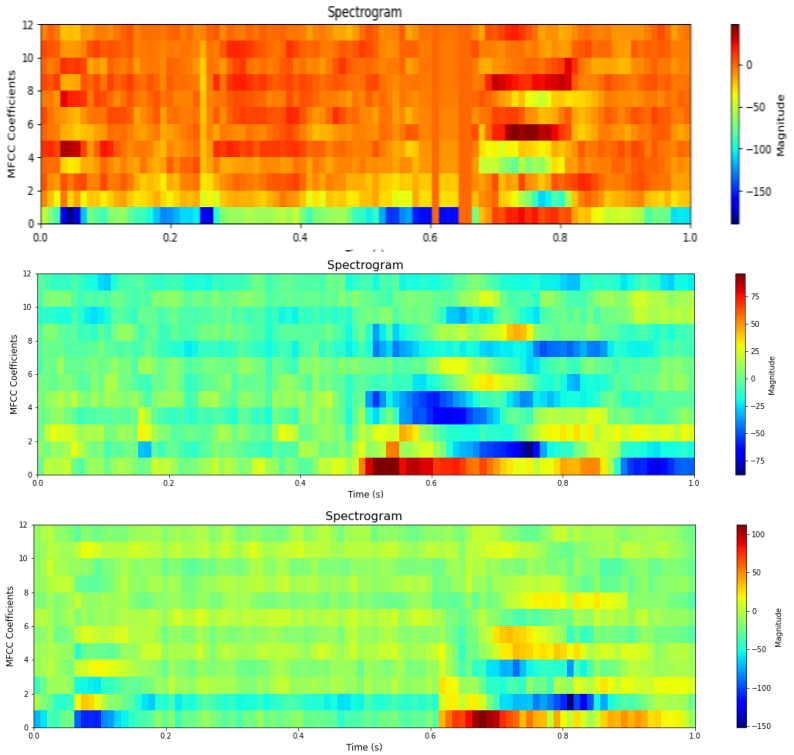
Εικόνα 20 – Φάσμα κλίμακας Mel- Εντολή "eight"

Στην Εικόνα 21 φαίνεται η σύγκριση του φάσματος συχνοτήτων και Mel για την ίδια καταγραφή της εντολής "*six*".

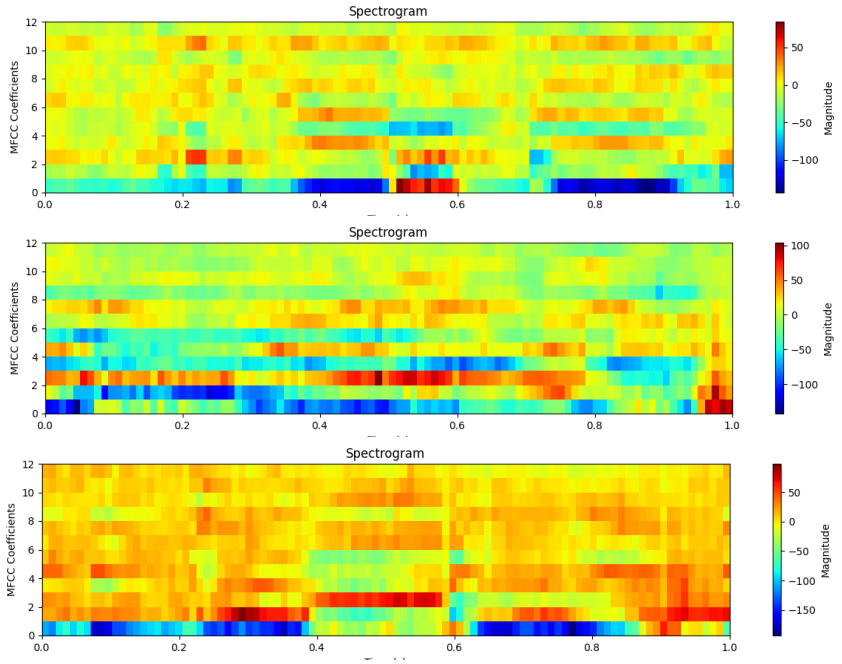


Εικόνα 21 – Σύγκριση Φασματικής κλίμακας Mel- Εντολή "six"

Τα φάσματα έχουν πολύ παρόμοια απεικόνιση, η κύρια διαφορά είναι η τάξη μεγέθους, όμως φαίνεται ότι σε υψηλές συχνότητες, η περιοχές υψηλής ενέργειας είναι διογκωμένες, καθώς η διάκριση των διαφορών στην ένταση για παραπλήσιες συχνότητες είναι δυσκολότερη σε σχέση με τις χαμηλές.



Εικόνα 22 –Φασμ συντελεστές MFCCs "one"



Εικόνα 23 –Φασμ συντελεστές MFCCs "six"

4.1.3. Κατηγοριοποίηση

Μετά την εξαγωγή των χαρακτηριστικών, οι συντελεστές MFCC χρησιμοποιούνται ως είσοδοι του νευρωνικού δικτύου για εκπαίδευση. Αυτή είναι «η στιγμή της αλήθειας», καθώς η απόδοση του νευρωνικού υποδεικνύει και την ποιότητα της προ-επεξεργασίας μέχρι τώρα. Με το κατάλληλο νευρωνικό δίκτυο, είναι εφικτή η επίτευξη απόδοσης 80%-85% στο υποσύνολο ελέγχου. Η κατηγοριοποίηση γίνεται σε περιβάλλον MATLAB R2023b με την χρήση τεχνητού νευρωνικού δικτύου κατηγοριοποίησης (patternnet).

4.1.3.1. k-Fold Cross Validation

Η μέθοδος the k-fold cross validation χρησιμοποιείται για τον αντικειμενικό τερο υπολογισμό της ακρίβειας των νευρωνικών. Όταν ένα νευρωνικό δίκτυο είναι «προκατειλλημένο» σημαίνει ότι τείνει να αναγνωρίζει συγκεκριμένα χαρακτηριστικά στα οποία έχει μεγαλύτερη ευαισθησία και να κατευθύνει τις προβλέψεις προς λίγες κλάσεις. Από την βιβλιογραφία οι προτεινόμενες τιμές είναι 5 με 10 για τον αριθμό των folds. Για k=5, το υποσύνολο εκπαίδευσης αποτελείται από 33822 παρατηρήσεις και το υποσύνολο ελέγχου από 8455. Οι είσοδοι του νευρωνικού είναι 12 συντελεστές για κάθε τμήμα σήματος, επί 98 τμήματα συνολικά.

4.1.3.2. Παράμετροι τεχνητού νευρωνικού δικτύου

Η διερεύνηση που πραγματοποιείται είναι ως προς την αρχιτεκτονική του δικτύου. Οι άλλες παράμετροι του νευρωνικού συνοψίζονται στον παρακάτω πίνακα. Η συνάρτηση εκπαίδευσης που χρησιμοποιείται είναι ο αλγόριθμος Scaled Conjugate Gradient (SCG). Η μέθοδος SCG αποτελεί εξέλιξη των συζευγμένων μεθόδων, αλλά αντί να χρησιμοποιεί την αναζήτηση γραμμής για τον καθορισμό του βέλτιστου βήματος, υπολογίζει το διάστημα με βάση έναν μηχανισμό κλιμάκωσης, βελτιώνοντας την αποτελεσματικότητα της μεθόδου. Επιπλέον, χρησιμοποιεί τον πινακα Hessian με πληροφορίες δεύτερης τάξης, σε αντίθεση με τη μέθοδο απλής κλίσης πρώτης τάξης, επιταχύνοντας τη σύγκλιση. Ο μέγιστος αριθμός εποχών ορίζεται σε 500 για να υπάρχει μια καλή ισορροπία μεταξύ υψηλής ακρίβειας και χαμηλού υπολογιστικού χρόνου. Η έξοδος από το δίκτυο προώθησης είναι η δυνατότητα, η παρατήρηση να ανήκει σε κάθε μία από τις 18 κλάσεις, αλλά η επιθυμητή έξοδος είναι η γραπτή εντολή, άρα η ετικέτα της κλάσης. Η συνάρτηση μεταφοράς «SoftMax», κάνει αυτή τη δουλειά, επιλέγοντας την κλάση με την υψηλότερη πιθανότητα

Πίνακας 4 – Παράμετροι ΤΝΔ

Type	Pattern network
k-fold cross validation	k = 5
Hidden layers	2
Training function	Scaled Conjugate Gradient ('trainscg')
Train ratio	80%
Test ratio	20%
Transfer function	SoftMax function ('softmax')
Maximum number of epochs	500 epochs

4.1.3.3. Διερεύνηση αρχιτεκτονικής

Η διερεύνηση γίνεται με την μέθοδο της υποβολής (brute force), δηλαδή χειροκίνητα εξετάζονται όλοι οι πιθανοί συνδυασμοί. Το βασικό κριτήριο είναι η ακρίβεια μεγαλύτερη του 80%, αλλά για το τελικό δίκτυο υπολογίζονται όλα τα κριτήρια αξιολόγησης. Εάν δύο νευρωνικά έχουν την ίδια

απόφαση, προφανώς θα επιλεχθεί αυτό με την μικρότερη αρχιτεκτονική. Αλλά ας ξεκινήσει η διερεύνηση απ' την αρχή.

Σε πρώτη φάση, η έρευνα έγινε για τα ψηφία μόνο (10 εντολές), αρχίζοντας από μικρά μεγέθη (30x30) και καταλήγοντας σε μεγαλύτερα (300x300).

Πίνακας 5 – Μόνο ψηφία (10 εντολές)

k = 5	*30x30*	*40x40*	*50x50*	*60x60*	*80x80*	*100x100*
Fold No 1	69.5%	73.5%	75.8%	78.3%	82.0%	83.7%
Fold No 2	61.5%	73.9%	75.4%	79.5%	80.2%	81.6%
Fold No 3	69.2%	73.3%	76.7%	78.0%	80.8%	75.3%
Fold No 4	69.6%	75.6%	76.9%	79.0%	80.6%	82.2%
Fold No 5	70.6%	73.5%	76.0%	78.7%	80.1%	82.0%
Average	***68.1%***	***74.0%***	***76.1%***	***78.7%***	***80.7%***	***80.9%***

Στο μικρότερο νευρωνικό δίκτυο, 30x30, η ακρίβεια είναι ήδη σχετικά καλή (~70%) και αυξάνοντας τους νευρώνες στο 100x100 φτάνει ήδη τον στόχο/ Σε όλες τις περιπτώσεις .η ακρίβεια στο υποσύνολο ελέγχου είναι κοντά στο 99%. Με την περαιτέρω αύξηση της αρχιτεκτονικής φτάνει η ακρίβεια στο 87%.

Πίνακας 6 – Μόνο ψηφία (10 εντολές)

k = 5	*120x120*	*140x140*	*180x180*	*250x250*	*300x300*
Fold No 1	84.3%	85.2%	86.3%	86.4%	88.7%
Fold No 2	84.5%	84.7%	86.4%	87.9%	88.0%
Fold No 3	83.8%	84.9%	85.2%	86.6%	86.9%
Fold No 4	84.3%	85.8%	86.3%	87.0%	86.8%
Fold No 5	83.8%	84.8%	85.4%	86.0%	87.2%
Average	***84.1%***	***85.1%***	***85.9%***	***86.8%***	***87.5%***

Το επόμενο στάδιο είναι η αύξηση του αριθμού των εντολών, με προσθήκες των λέξεων, "*Right*", "*Stop*", "*On*" και "*Off*". Η απόδοση πέφτει περίπου 10% για τις ίδιες αρχιτεκτονικές με αύξηση των εντολών από 10 σε 17.

Πίνακας 7 – 17 εντολές

k = 5	*80x80*	*100x100*	*120x120*	*140x140*	*180x180*	*250x250*
Fold No 1	*67.5%*	*70.7%*	*74.2%*	*73.4%*	*77.5%*	*78.6%*
Fold No 2	*69.3%*	*71.4%*	*73.3%*	*74.3%*	*76.4%*	*78.9%*
Fold No 3	*68.0%*	*72.6%*	*75.2%*	*75.9%*	*76.8%*	*81.1%*
Fold No 4	*67.8%*	*71.7%*	*74.4%*	*76.1%*	*76.6%*	*74.3%*
Fold No 5	*67.2%*	*69.8%*	*73.4%*	*74.6%*	*77.8%*	*79.4%*
Average	***68.0%***	***71.2%***	***74.1%***	***74.9%***	***77.0%***	***78.5%***

Η εντολή "Go" προστίθεται στο λεξιλόγιο και ταα νευρωνικά δίκτυα, επανεκπαιδεύονται. Για την ίδια αρχιτεκτονική, σε σχέση με το σύνολο 17 εντολών, η απόδοση μειώνεται κατά 1%-2%, αλλά ακόμα επιτυγχάνεται ο στόχος (80%), με μοντέλο 300x300.

Πίνακας 8 – Τελικό σετ: 18 εντολές

k = 5	*120x120*	*140x140*	*180x180*	*250x250*	*300x300*
Fold No 1	*73.3%*	*74.1%*	*74.6%*	*77.8%*	*80.3%*
Fold No 2	*70.5%*	*74.3%*	*74.5%*	*76.3%*	*81.1%*
Fold No 3	*73.5%*	*74.5%*	*73.6%*	*74.9%*	*80.4%*
Fold No 4	*71.0%*	*74.0%*	*76.1%*	*80.2%*	*81.0%*
Fold No 5	*73.4%*	*74.8%*	*76.5%*	*77.0%*	*81.2%*
Average	***72.4%***	***74.3%***	***75.1%***	***77.2%***	***80.8%***

Τα αποτελέσματα αυτής της αρχικής μελέτης θέτουν τις βάσεις και τα όρια για τα επόμενα βήματα της έρευνας. Με την αύξηση της πολυπλοκότητας της δομής του δικτύου, η βελτίωση της ακρίβειας των δοκιμών εξακολουθεί να είναι αξιοσημείωτη, γεγονός που δείχνει ότι η κατεύθυνση είναι σωστή και δεν έχει επιτευχθεί ακόμη η απόλυτα καλύτερη απόδοση. Βέβαια, σε μικρότερες αρχιτεκτονικές το όφελος στην ακρίβεια είναι μεγαλύτερο με την ίδια αύξηση του αριθμού των νευρώνων, σε σύγκριση με το όφελος σε μεγαλύτερα δίκτυα. Είναι επίσης σαφές, ότι δεν έχει νόημα η χρήση δικτύων με λιγότερους από 100 νευρώνες ανά στρώμα, καθώς η ακρίβειά τους είναι μικρότερη από 70%. Τα δίκτυα, μέχρι στιγμής, είναι τετραγωνικής διάταξης, που σημαίνει ότι και τα δύο στρώματα έχουν τον ίδιο αριθμό κρυφών νευρώνων. Για το επόμενο βήμα, εφαρμόζεται η μέθοδος της ωμής βίας για αρχιτεκτονικές μεταξύ 100x100 και 850x850, με διάστημα 50 νευρώνων.

Πίνακας 9:Διερεύνιση ακρίβειας 1

		Hidden Layer #2									
		100	***150***	***200***	***250***	***300***	***350***	***400***	***450***	***500***	***Mean***
Hidden Layer #1	***100***	70.70%	71.03%	72.59%	71.78%	72.99%	71.73%	73.98%	73.40%	74.96%	***72.57%***
	150	72.32%	74.49%	75.03%	74.96%	75.86%	75.27%	75.76%	75.34%	76.84%	***75.10%***
	200	75.63%	75.81%	76.73%	77.73%	76.17%	76.00%	78.16%	76.78%	77.02%	***76.67%***
	250	77.29%	77.31%	77.56%	77.98%	77.93%	78.92%	78.23%	79.13%	77.12%	***77.94%***
	300	78.02%	77.85%	77.16%	77.53%	78.13%	77.92%	77.41%	79.39%	79.34%	***78.08%***
	350	78.17%	78.58%	77.76%	78.80%	78.88%	75.57%	79.13%	78.25%	79.70%	***78.32%***
	400	78.89%	78.83%	79.28%	80.23%	80.16%	79.60%	79.57%	79.87%	78.61%	***79.45%***
	450	79.17%	80.15%	78.70%	79.35%	79.26%	79.43%	80.16%	80.07%	80.60%	***79.65%***
	500	79.63%	79.81%	80.57%	79.37%	80.00%	79.96%	79.77%	78.06%	79.77%	***79.66%***
	550	79.9%	80.2%	80.2%	80.4%	78.8%	79.6%	80.5%	79.9%	80.1%	***79.96%***
	600	79.9%	79.7%	80.2%	79.1%	79.1%	79.1%	80.9%	79.5%	80.6%	***79.79%***
	650	80.2%	80.6%	80.7%	79.4%	80.2%	80.2%	80.7%	81.4%	78.0%	***80.14%***
	Mean	***77.49%***	***77.86%***	***78.04%***	***78.05%***	***78.12%***	***77.77%***	***78.68%***	***78.43%***	***78.56%***	

Στον πίνακα 9 συνοψίζονται τα αποτελέσματα για διάφορους συνδυασμούς αρχιτεκτονικής. Ο αριθμός των νευρώνων στα στρώματα 1 ορίζεται ως n_1 και n_2 είναι ο αριθμός των νευρώνων στο δεύτερο στρώμα. Οι στήλες είναι σταθερού n_2 και διαφορετικού n_1 και οι γραμμές είναι σταθερού n_1 και διαφορετικού n_2. Στον πρώτο πίνακα είναι τα αποτελέσματα για $n\_1 \in [100,650]$ και $n\_2 \in [100,400]$. Για κάθε στήλη και γραμμή υπολογίζεται η μέση ακρίβεια. Τα συμπεράσματα από αυτά τα αποτελέσματα είναι τα εξής:

- Η ελάχιστη ακρίβεια, 70%, εμφανίζεται για την αρχιτεκτονική 100x100.
- Η μέγιστη ακρίβεια είναι γύρω στο 80%-81% και υπάρχει για διαφορετικούς συνδυασμούς.
- Η αύξηση του μεγέθους του στρώματος 1 είναι πιο αποτελεσματική από την ίδια αύξηση του στρώματος 2. Παρατηρώντας το μέσο όρο ανά γραμμή, η ακρίβεια βελτιώνεται, με την προοδευτική αύξηση του n_1, περισσότερο στα πρώτα βήματα (μεταξύ 100 και 300 σχεδόν 2% ανά αύξηση), λιγότερο στη μέση (0,5% μέχρι 500 νευρώνες) και στο τέλος η διαφορά

είναι σχεδόν μηδενική. Η μέση τιμή ανά στήλη, δεν ακολουθεί την ίδια τάση, και είναι σχεδόν η ίδια για όλα τα διαφορετικά μεγέθη του στρώματος 2 και ίση με 78%.

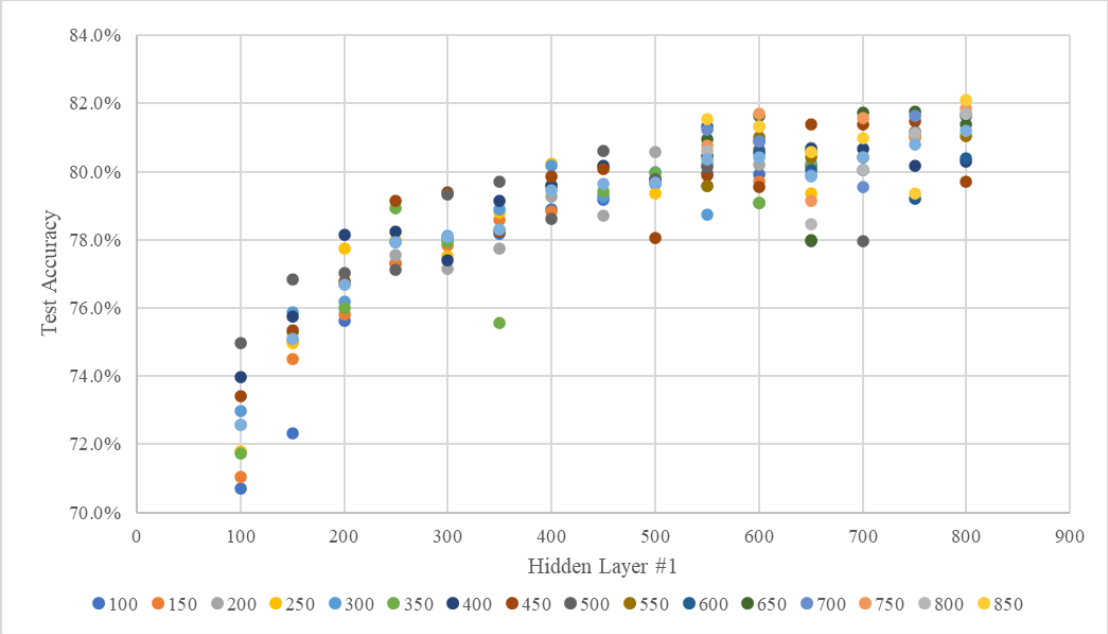
Η ίδια αλλαγή (π.χ. αύξηση του στρώματος 1 κατά 50 νευρώνες) είναι πιο επωφελής σε μικρότερες αρχιτεκτονικές, παρά σε πιο σύνθετες. Αυτό είναι λογικό, όταν φτάνουμε πιο κοντά στη μέγιστη απόδοση οι δυνατότητες είναι μικρότερες.

Η καλύτερη ακρίβεια μπορεί να επιτευχθεί με διάφορους συνδυασμούς.

- n_1=400, θα πρέπει να είναι n_2≥300
- n_1=500 ή n_1=600, θα πρέπει να είναι n_2≥200

Πίνακας 10:Διερεύνιση ακρίβειας 1

		Hidden Layer #2										
		400	***450***	***500***	***550***	***600***	***650***	***700***	***750***	***800***	***800***	***Mean***
Hidden Layer #1	***550***	80.5%	79.9%	80.1%	79.6%	81.3%	81.0%	81.2%	80.8%	80.6%	80.6%	***80.65%***
	600	80.9%	79.5%	80.6%	81.0%	80.5%	81.6%	80.9%	81.7%	81.3%	81.3%	***80.95%***
	650	80.7%	81.4%	78.0%	80.4%	80.1%	78.0%	79.9%	79.1%	78.5%	78.5%	***79.65%***
	700	80.7%	81.4%	78.0%	80.4%	80.1%	81.7%	79.5%	81.6%	80.0%	80.0%	***80.43%***
	750	80.2%	81.5%	81.2%	81.0%	79.2%	81.8%	81.6%	81.0%	81.1%	81.1%	***80.79%***
	800	80.3%	79.7%	81.7%	81.1%	80.4%	81.4%	81.7%	81.9%	81.7%	82.1%	***81.01%***
	Mean	***80.30%***	***80.32%***	***79.77%***	***80.59%***	***80.26%***	***80.91%***	***80.81%***	***81.01%***	***80.32%***	***80.76%***	



Εικόνα 24 – Επιρροή δεύτερου κρυμμένου επιπέδου

Εικόνα 25 – Επιρροή δεύτερου κρυμμένου επιπέδου

Η Εικόνα 25 δείχνει τη σχέση μεταξύ της ακρίβειας (κάθετος άξονας) και του μεγέθους του πρώτου στρώματος (οριζόντιος άξονας). Τα διαφορετικά χρώματα αντιπροσωπεύουν διαφορετικό αριθμό νευρώνων στο στρώμα #2. Στο Σχήμα 24 απεικονίζεται η ακρίβεια του δικτύου για παραλλαγές του μεγέθους του στρώματος #2. Εδώ τα διαφορετικά χρώματα αντιπροσωπεύουν το διαφορετικό μέγεθος του στρώματος #1. Συγκρίνοντας τα δύο σχήματα είναι προφανές ότι η σχέση με τη συνολική ακρίβεια είναι πολύ διαφορετική: το δεύτερο κρυφό στρώμα δεν έχει την ίδια επιρροή με το πρώτο. Σε γενικές γραμμές, είναι λογικό να υποθέσουμε ότι η δομή του πρώτου στρώματος καθορίζει το επίπεδο ακρίβειας και η αύξηση του αριθμού των νευρώνων στο δεύτερο επίπεδο κάνει τη ρύθμιση.

Από την έρευνα, δεν είναι σαφές ποια αρχιτεκτονική είναι η καλύτερη, αφού η απαίτηση για ακρίβεια 80%, μπορεί να επιτευχθεί με διαφορετικούς συνδυασμούς. Αξίζει να εξεταστούν περαιτέρω οι επιλογές και να χρησιμοποιηθεί ο πίνακας αξιολόγησης για τη λήψη της τελικής απόφασης.

4.1.3.4. Evaluation Metrics – ANN 600x450

Ακολουθούν τα αποτελέσματα ταξινόμησης για το δίκτυο 600x450. Η συνολική ακρίβεια είναι 80,7%, η οποία παραμένει στο ίδιο επίπεδο και στις πέντε αναδιπλώσεις. Η απόδοση είναι πολύ ισορροπημένη, καθώς όλες οι μετρικές αξιολόγησης βρίσκονται πολύ κοντά η μία στην άλλη. Στον πίνακα 11 παρουσιάζονται όλες οι μετρικές για αυτό το δίκτυο.

Πίνακας 11 – Test Accuracy 600x450

	Fold No 1	*Fold No 2*	*Fold No 3*	*Fold No 4*	*Fold No 5*	***Average***
Accuracy	*81.4%*	*79.5%*	*80.5%*	*80.9%*	*81.1%*	***80.7%***
Precision	*80.1%*	*78.5%*	*80.3%*	*81.0%*	*81.1%*	***80.7%***
Recall	*79.7%*	*80.3%*	*78.5%*	*79.9%*	*79.4%*	***80.7%***
F1-score	*79.8%*	*79.3%*	*79.4%*	*80.4%*	*81.2%*	***80.7%***

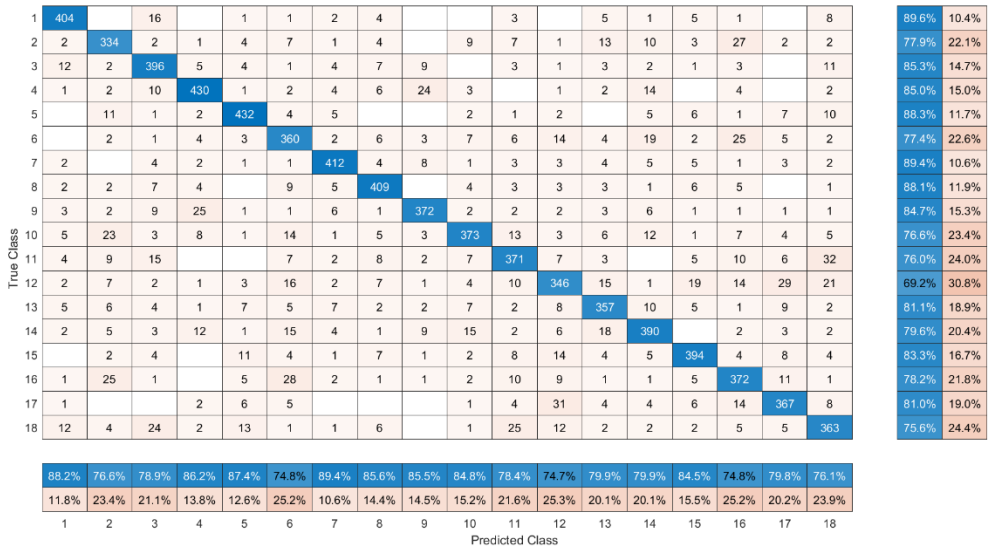


Figure 1 – Confusion matrix k-fold1 – 600x450

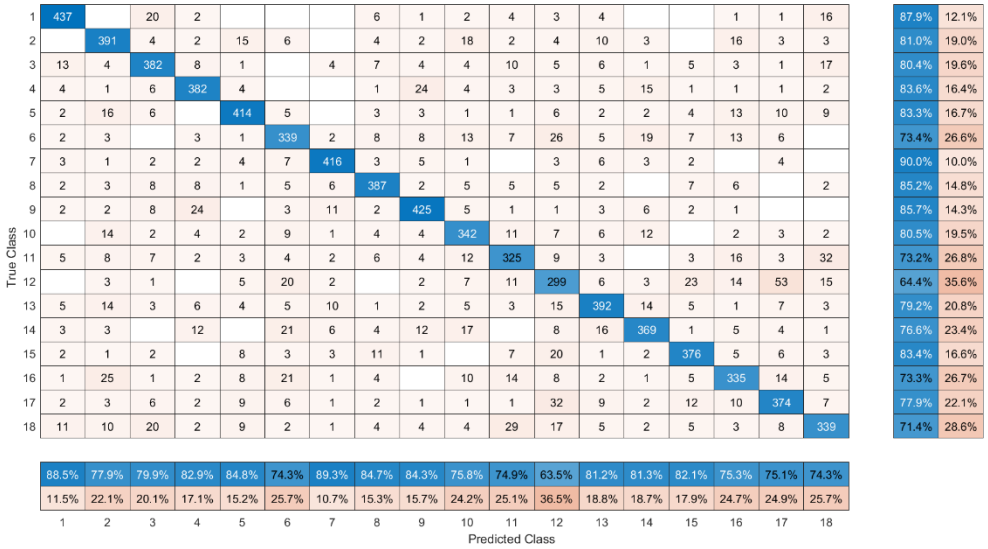


Figure 2 – Confusion matrix k-fold2 – 600x450

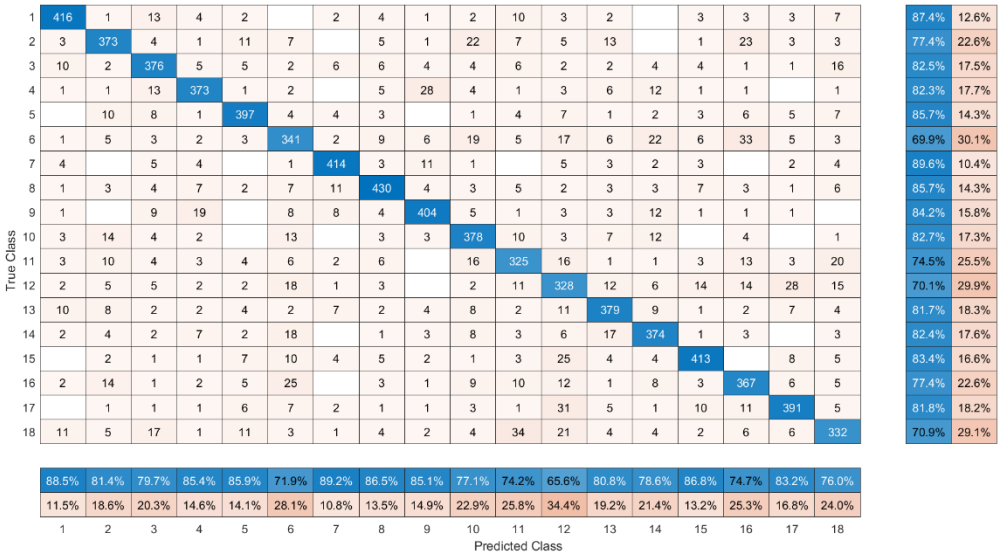


Figure 3 – Confusion matrix k-fold3 – 600x450

Οι επιμέρους κλάσεις έχουν απόκλιση 10% στην ακρίβειά τους. Οι ακρίβειες είναι γενικά πάνω από 75%, αλλά σε όλες τις διαφορετικές αναδιπλώσεις, υπάρχουν τρεις, περίπου, κλάσεις που η ακρίβειά τους είναι γύρω στο 70%. Οι λανθασμένες προβλέψεις φαίνεται να είναι τυχαίες και να μην ακολουθούν κάποιο μοτίβο. Για παράδειγμα, θα ήταν αναμενόμενο να συγχέονται οι εντολές «ένα» (No2) και «on» (No16) ή «on» (No16) και «off» (No17), καθώς ο ήχος μπορεί να είναι παρόμοιος, αλλά το μοντέλο δεν παρουσιάζει αυτού του είδους την ευαισθησία. Σε όλες τις περιπτώσεις, η λέξη «μηδέν» (No1) είναι μία από τις καλύτερες προβλεπόμενες κλάσεις, ενώ η λέξη «κάτω» (No12) είναι μία από τις χειρότερες. Οι εντολές «δύο» και «τρία» έχουν πολύ καλά ποσοστά, πολύ υψηλότερα από το «πέντε» ή το «κάτω». Όταν ο χρήστης χρησιμοποιεί το μοντέλο, αυτή η διαφορά στην ακρίβεια θα αυξηθεί και σε ορισμένες λέξεις η απόδοση ταξινόμησης θα είναι χαμηλότερη από ό,τι σε άλλες.

5. Συμπεράσματα

Η παρούσα εργασία εστιάζει στην ανάπτυξη μοντέλου αναγνώρισης φωνητικών εντολών με χρήση τεχνητών νευρωνικών δικτύων και χαρακτηριστικών συντελεστών MFCCs. Κύριο μέρος της διπλωματικής είναι η εύρεση της κατάλληλης μεθόδου για την προ-επεξεργασία του σήματος και την εξαγωγή τω κατάλληλων χαρακτηριστικών. Το δεύτερο μεγάλο κομμάτι είναι η εκπαίδευση τεχνητού νευρωνικού δικτύου, για κατηγοριοποίηση 18 εντολών. Οι εντολές επιλέχθηκαν από το Speech Commands Dataset της Google, για την χρήση σε σύζευξη με βιομηχανικό βραχίονα, καθώς υπάρχει η προοπτική χρήσης του μοντέλου σε βιομηχανικές εφαρμογές ρομπότ. Η επεξεργασία του σήματος αποτελείται από την αφαίρεση θορύβου, την εξισορρόπηση του φάσματος συχνοτήτων, την τμηματοποίηση του σήματος – ώστε να μπορεί να θεωρεί σταθερό ως προς τον χρόνο – και την εξαγωγή των συντελεστών MFCCs, εφαρμόζοντας τον διακριτό μετασχηματισμό Fourier, μετατρέποντας το αποτέλεσμα σε κλίμακα Mel και εφαρμόζοντας τον ανάστροφο μετασχηματισμό Fourier. Για το κομμάτι της τεχνητής νοημοσύνης, το αντικείμενο που επιλέχθηκε είναι ένα τεχνητό νευρωνικό δίκτυο αναγνώρισης μοτίβων. Το κύριο κομμάτι της έρευνας είναι η αναζήτηση της βέλτιστης αρχιτεκτονικής για επίτευξη ακρίβειας 80%.

Το πρότζεκτ αυτό είναι μία καλή αρχή για την ανάπτυξη τεχνικών τεχνητής νοημοσύνης για αναγνώριση εντολών σε βιομηχανικό επίπεδο. Η μεθοδολογία που αποφασίστηκε και ακολουθήθηκε καλύπτει όλες τις βασικές αρχές της αναγνώρισης ομιλίας. Είναι ιδιαίτερα απαιτητική η κατανόηση και ανάλυση της παραγωγής ομιλίας και της πρόσληψης ήχων από τον άνθρωπο, στην συνέχεια, η αντιστοίχισή τους σε μαθηματικές ποσότητες και συναρτήσεις και, τελικά, η αναγνώρισή τους.

Τα βασικά ευρήματα της μελέτης συνοψίζονται παρακάτω.

- Η ανθρώπινη ομιλία περιέχει πολλή πληροφορία για την ταυτότητα του ομιλητή, την συναισθηματική του κατάσταση, για την προφορά και, επομένως, για την καταγωγή του και φυσικά για τις λέξεις που εκφράζονται. Οι άνθρωποι μπορούν να αντιληφθούν όλες τις παραπάνω πληροφορίες ταυτόχρονα, ενώ τα μοντέλα μηχανικής μάθησης συνήθως εστιάζουν σε μία αναγνώριση την φορά. Η σωστή προεπεξεργασία του σήματος, περιέχει τον υπολογισμό των κατάλληλων μαθηματικών μεγεθών για την αναγνώριση συγκεκριμένης πληροφορίας, ανάλογα με τις απαιτήσεις της εκάστοτε εφαρμογής.
- Ο διαχωρισμός του σήματος σε μικρότερα τμήματα, πληρεί τις προϋποθέσεις, ώστε κάθε τμήμα να θεωρείται χρονικά αμετάβλητο. Η τμηματοποίηση δίνει την απαιτούμενη σημασία σε μικρές λεπτομέρειες του σήματος, που θα χάνοντας υπό άλλες συνθήκες. Κατά το στάδιο αυτό, είναι σημαντική η χρήση επικαλυπτόμενων διαστημάτων, και όχι σειριακών, ώστε να διατηρηθεί όλη η σημαντική πληροφορία και να μην χαθεί στα άκρα των τμημάτων. Η χρήση windows ελαχιστοποιεί την διαρροή ενέργειας στο φάσμα των συχνοτήτων.
- Κατά το στάδιο της κατηγοριοποίησης, υπάρχουν πολλές μέθοδοι που μπορούν να χρησιμοποιηθούν, αλλά όποια και αν επιλεχθεί είναι σημαντικό να βελτιστοποιηθεί για την παρούσα εφαρμογή. Στην περίπτωση των τεχνητών νευρωνικών δικτύων εμπρόσθιας τροφοδότησης, η επιλογη κατάλληλης αρχιτεκτονικής είναι βασικός παράγοντας επιλογής. Για την επιλογή πολλών κρυμμένων επιπέδων, το πρώτο επίπεδο είναι πιο σημαντικό για το τελικό αποτέλεσμα και καθορίζει σε μεγάλο βαθμό την συνολική ακρίβεια του μοντέλου, ενώ τα υπόλοιπα επίπεδα χρησιμοποιούνται για τελειοποίηση της απόδοσης.
- Η χρήση απλών δομών, μικρού μεγέθους, οδηγεί σε ελαφριά και γρήγορα μοντέλα, αλλά περιορίζει την απόδοσή τους, καθώς υπάρχει υπο-εκπαίδευση. Ειδικά σε προβλήματα ιδιαίτερα απαιτητικά και πολύπλοκα, όπως η αναγνώριση φωνητικών εντολών. Αντίθετα, η χρήση μεγάλων και πολύπλοκων δομών, αυξάνει την ακρίβεια με επιπλέον υπολογιστικό κόστος και απαιτήσεις σε μνήμα. Η χρήση υπερβολικά μεγάλων μοντέλων, για την συγκεκριμένη εφαρμογή, δημιουργεί τον κίνδυνο υπερ-εκπαίδευσης στο γνωστό υποσύνολο και περιορίζει την ικανότητα γενίκευσης προς άγνωτες εντολές.

5.1. Μελλοντικά Βήματα

5.1.1. Προτάσεις Βελτίωσης

Η παρούσα εργασία αποτελεί μία πρώτη απόπειρα ανάπτυξης τεχνητών νευρωνικών δικτύων προς αναγνώριση φωνητικών εντολών. Τα αποτελέσματα της μελέτης είναι ικανοποιητικά κα υποσχόμενα, όμως επιδέχονται επιπλέον βελτιώσεις και προσθήκες. Αναφορικά με το λεξιλόγιο, προς το παρόν δεν είναι αρκετό για την ολοκληρωμένη διατύπωση των εντολών προς υλοποίηση από τον βραχίονα. Δεν ήταν εφικτή η εύρεση ενός πλήρους συνόλου δεδομένων, που να περιέχει όλες τις εντολές της V+ για την περιγραφή των κινήσεων του βραχίονα. Προτείνεται η δημιουργία ειδικού λεξιλογίου, που να περιέχει όλες τις απαιτούμενες εντολές για την αποδεκτή λειτουργία του βραχίονα. Κάτι τέτοιο δεν υπήρχε ο χρόνος να πραγματοποιηθεί κατά την παρούσα μελέτη, θα ήταν, όμως, αρκετά χρήσιμο να γίνει, καθώς ο έλεγχος κατά την δημιουργία του θα ήταν μεγαλύτερος, εξασφαλίζοντας ικανοποιητική ποιότητα, και θα έδινε μεγαλύτερη ελευθερία επιλογής των κατάλληλων λέξεων.

Όσον αφορά το μοντέλο κατηγοριοποίησης, διερευνήθηκαν μόνο τεχνητά νευρωνικά δίκτυα με μόνο δύο κρυμμένα επίπεδα. Θα ήταν ενδιαφέρουσα, η διερεύνηση δικτύων μεγαλύτερης πολυπλοκότητας, ειδικότερα για την χρήση μεγαλύτερου λεξιλογίου. Επιπλέον, θα ήταν χρήσιμη η σύγκριση του συγκεκριμένου τύπου δικτύου με άλλα μοντέλα κατηγοριοποίησης, όπως Naïve Bayes (NB), Random Forest (RF) και Nearest Neighbors (k-NN) ή μοντέλα βαθιάς μάθησης, όπως τα συνελικτικά νευρωνικά δίκτυα. Τα τελευταία θα μπορούσαν να χρησιμοποιηθούν με δύο τρόπος, είτε προς αναγνώριση των μαθηματικών συντελεστών MFCCs είτε προς αναγνώριση εικόνας, με είσοδο τα οπτικοποιημένα φάσματα των MFCCs (Cepstrums).

5.1.2. Επιπλέον Μελέτες

Το μοντέλο αναγνώρισης εντολών προορίζεται για σύζευξη με βιομηχανικό ρομποτικό βραχίονα, προς έλεγχο των διεργασιων του. Στην παρούσα μελέτη, αυτο χρησιμοποιήθηκε μόνο για τον καθορισμό κατάλληλου λεξιλογίου, που απαιτείται για τον πλήρη έλεγχο των κινήσεων του ρομπότ. Μία πρόταση για επιπλέον μελέτη, είναι η ανάπτυξη της διεπαφής προγράμματος-ρομπότ, έτσι ώστε να μπορεί να αξιολογηθεί και πρακτικά η ποιότητα και χρησιμότητα της εφαρμογής. Μέσω του προγράμματος διεπαφής, όλες οι φωνητικές εντολές, και η αντίστοιχη γραπτή αναπαράστασή τους, πρέπει να αντιστοιχούν σε προκαθορισμένες εντολές της γλώσσας V+ και στην συνέχεια να διαμορφώνεται το πρόγραμμα V+ με σωστή σύνταξη και δομή. Παραδείγματος χάριν, η αναγνώριση της λέξης ”*move*” θα πρέπει να αντιστοιχεί σε εντολή “*MOVE*”, ή ακόμα καλύτερη, η σειριακή αναγνώριση των εντολών “*MOVE*”, “*five*”, “*zero*”, “*slash*”, “*two*”, “*zero*”, “*slash*”, “*two*”, “*zero*”, θα πρέπει να συνεπάγεται “*MOVE (50, 20, 30)”*.

6. Βιβλιογραφία

[1] Staubli, Arm - RX series 90B family, 2008.

[2] A. T. Ashraf, A. S. Hasanen και F. N. Mohammad, «Voice recognition system using machine learning techniques,» *Elsevier,* April 2021.

[3] V. Tiwari, «MFCC and its applications in speaker recognition,» *International Journal on Emerging Technologies,* p. 4, February 2010.

[4] B. Copeland, «Artificial Intelligence definition,» Encyclopaedia Britannica, 2024. [Ηλεκτρονικό].

[5] J. Holdsworth και M. Scapicchio, «Deep learning vs. machine learning,» IBM American multinational technology corporation, 2017. [Ηλεκτρονικό]. Available: https://www.ibm.com/topics/deep-learning.

[6] A. L. Samuel, «Some Studies in Machine Learning Using the Game of Checkers,» *IBM Journal of Research and Development,* p. 21, 1959.

[7] R. Karjian, «History and evolution of maching learning: A timeline,» TechTarget, 13 June 2024. [Ηλεκτρονικό]. Available: https://www.techtarget.com/whatis/A-Timeline-of-Machine-Learning-History.

[8] W. Pitts και W. McCulloch, «A logical calculus of the ideas immanent in nervous activity,» *Bulletin of Mathematical Biology,* p. 17, 1943.

[9] D. Hebb, The Organization of Behavior: A neuropsychological Theory, 1949.

[10] A. Turing, Computing Machinery and Intelligence, Mind, 1950.

[11] Y. LeCun , Y. Bengio και P. Haffner, Backpropagation Applied to Handwritten Zip Code Recognition, MIT Press, 1989.

[12] S. Albahli, F. Alhassan, W. Albattah και R. U. Khan, Handwritten Digit Recognition: Hyperparameters-Based Analysis, MDPI Applied Science, 2020.

[13] M. Pinola, Speech Recognition Through the Decades: How we ended up with Siri, 2011.

[14] D. Spicer, «AUDREY, Alexa and more: A history of automatic speech recognition,» 2021. [Ηλεκτρονικό]. Available: https://computerhistory.org/blog/audrey-alexa-hal-and-more/.

[15] H. Kumari, J. Biji και K. A. Navas, «A Novel Objective Audio Quality Measure,» *10th National Conference on Technological Trends,* 2009.

[16] UNIVERSAL ROBOTS, "Best Applications of Robotic Arms," 2022.

[17] UNIVERSAL ROBOTS, «Types of Robotic Arms,» 2022.

[18] E. M. Rosales και Q. Gan, «Forward and Inverses Kinematics Models for a 5-dof Pioneer 2 Robot Arm,» University of Essex - Department of Computer Science, 2002.

[19] P. Makrylakis, «Industrial robot programming through voice commands,» National Technical University of Athens, Athens, 2023.

[20] A. Techhnology, «V+ Language Reference Guide,» 1997.

[21] A. Technology, «V+ Language User's Guide, Ver. 12.1,» 1997.

[22] B. Automation, «15 Robot End Effector Types and Selection Criteria,» 2022. [Ηλεκτρονικό]. Available: https://www.b2eautomation.com/insights/15-robot-end-effector-types-and-selection-criteria.

[23] A. N. S. S. M.M. Hasan, «An approach to voice conversion using feature statistical mapping,» *Elsevier,* p. 21, May 2005.

[24] D. Eringis και G. Tamulevičius, «Improving Speech Recognition Rate through Analysis Parameters,» *De Gruyter,* 2014.

[25] S. K. Kumar, B. Yazdanpanah και D. G. S. N. Raju, «Performance Comparison of Windowing Techniques for ECG Signal Enhancement,» *International Journal of Engineering Research,* p. 4, December 2014.

[26] M. Puckette, «Taxonomy of filters,» σε *Theory and Techniques of Electronic Music*, University of California, San Diego, World Scientific, 2003.

[27] G.-C. Vosniakos και P. Benardos, «Artificial Neural Networks in Manufacturing Systems,» National Technical University of Athens.

[28] P. Warden, «Speech Commands: A public dataset for single-word speech recognition - Copyright Google 2017,» [Ηλεκτρονικό]. Available: http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz. [Πρόσβαση 2017].

[29] V+ Language Users Giude Version 12.1, USA, 1997.

[30] R. M. V. V. L. Svitlana Maksymova, «Software for Voice Control Robot: Example of Implementation,» *Open Access Library Journal,* p. 12, 2017.

[31] L. Muda, M. Begam και I. Elamvazuthi, «Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques,» p. 6, March 2010.

[32] S. Khawatreh, B. Ayyoub, A. Abu-Ein και Z. Alqadi, «A Novel Methodology to Extract Voice Signal Features,» *International Journal of Computer Applications,* τόμ. 179, p. 4, 2018.

[33] H. Hofling, T. Berglund και A. Vaara, «Audio Compression,» Uppsala University, Uppsala, 2002.

[34] J. P. Egan και H. W. Hake, «On the masking pattern of a simple auditory stimulus,» *The Journal of the Acoustical Society of America,* pp. 622-630, 1950.

[35] J. V. Tobias, «Low-frequency masking patterns,» *The Journal of the Acoustical Society of America,* pp. 571-575, 1977.

[36] B. Y. D. G. S. N. R. K.Sravan Kumar, «Performance Comparison of Windowing Techniques for ECG Signal Enhancement,» *International Journal of Engineering Research,* p. 4, December 2014.

[37] W. L. Hosch, «Machine Learning definition,» Encyclopaedia Britannica, 2024. [Ηλεκτρονικό].

[38] A. Bryson και Y.-C. Ho, Applied optimal control, Hemisphere Pub. Corp., 1975.