Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Έλεγχος ρομποτικού οχήματος με τεχνικές μηχανικής μάθησης και αναγνώρισης φωνής

Διπλωματικη Εργασια

του

ΓΕΩΡΓΙΟΥ-ΤΑΞΙΑΡΧΗ ΓΙΑΝΝΙΟΥ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π

Συμμετοχή: Γεώργιος Σιόλας
Ε.Δ.Ι.Π

Εργαστηριο Τεχνητης Νοημοσυνης και Συστηματων Μαθησης
Αθήνα, Οκτώβριος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης

Έλεγχος ρομποτικού οχήματος με τεχνικές μηχανικής μάθησης και αναγνώρισης φωνής

Διπλωματικη Εργασια

του

ΓΕΩΡΓΙΟΥ-ΤΑΞΙΑΡΧΗ ΓΙΑΝΝΙΟΥ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π

Συμμετοχή: Γεώργιος Σιόλας
Ε.Δ.Ι.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26η Οκτωβρίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης	Γεώργιος Στάμου	Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2021

(Υπογραφή)

.....

Γεωργιος-Ταξιαρχης Γιαννιος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.
© 2021 – All rights reserved

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης

Copyright ©–All rights reserved Γεώργιος-Ταξιάρχης Γιαννιός, 2021.
Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Ευχαριστίες

Ολοκληρώνοντας την εκπόνηση της διπλωματικής μου εργασίας, κλείνει και ένας κύκλος φοίτησης στο Εθνικό Μετσόβιο Πολυτεχνείο. Μια πορεία, κατα την οποία πέρα απο τον προσωπικό μόχθο, συντέλεσαν καθοριστικά και κάποια πρόσωπα. Τα πρόσωπα αυτά, αισθάνομαι την ανάγκη να ευχαριστήσω.

Πρώτον θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ.Ανδρέα Σταφυλοπάτη, ο οποίος αρχικά με ενέπνευσε να ασχοληθώ με το συγκεκριμένο αντικείμενο, στην συνέχεια με εμπιστεύτηκε και τέλος μου έδωσε την ευκαιρία να εμβαθύνω περαιτέρω. Παράλληλα, ευχαριστώ τον Δρ. Γεώργιο Σιόλα για τις πολύτιμες συμβουλές που παρείχε και για την άψογη συνεργασία που θεμελίωσε, ήδη απο την αρχή. Τέλος, δε θα μπορούσα να παραλείψω και τα υπόλοιπα μέλη του εργαστηρίου, για το θεωρητικό αλλά και πρακτικό υπόβαθρο που μου διαμόρφωσαν, κατα την διάρκεια της προπτυχιακής μου εκπαίδευσης.

Ταυτόχρονα θα ήθελα να ευχαριστήσω μέσα απο την καρδία μου τους γονείς μου Νικόλαο και Μαρία, την αδερφή μου Ειρήνη, γενικότερα την οικογένειά και τους φίλους μου. Η βοήθεια τους ήταν διαρκής, ουσιαστική και πολύτιμη.

Περίληψη

Αντικείμενο της διπλωματικής εργασίας αποτελεί ο σχεδιασμός, η υλοποίηση καθώς και η αξιολόγηση ενός συστήματος το οποίο θα συμβάλει στην γόνιμη αλληλεπίδραση ενός ατόμου με ένα ρομποτικό όχημα. Η αλληλεπίδραση αυτή θα πραγματοποιείται με την χρήση συγκεκριμένων φωνητικών εντολών (πχ. 'stop', 'go', 'right'...), μέσω των οποίων θα ελέγχεται η κατεύθυνση του οχήματος.

Αρχικά συγκεντρώθηκαν ορισμένα ηχογραφημένα ακουστικά μηνύματα για κάθε μια απο τις εντολές. Τα δείγματα αυτά μετατράπηκαν σε κατάλληλη αναπαράσταση, ώστε να τροφοδοτηθούν σε ένα μοντέλο βαθιάς μηχανικής μάθησης. Ο πειραματισμός αφορούσε τόσο την εξαγωγή χαρακτηριστικών, όσο και τη σύνθεση της αρχιτεκτονικής του νευρωνικού δικτύου.

Ως προς την εξαγωγή χαρακτηριστικών, αρχικά πραγματοποιήθηκε ένας πειραματισμός στο πεδίο του χρόνου, με το κάθε ακουσικό σήμα να μελετάται σαν μια χρονοσειρά. Γρήγορα αναδείχτηκαν ορισμένες αδυναμίες, που έκαναν επιτακτική την ανάγκη μετάβασης στο πεδίο της συχνότητας. Στο πεδίο αυτό, μετά απο πειράματα αναδείχτηκε το μοντέλο αλλά και τα χαρακτηριστικά που επιτυγχάνουν το μεγαλύτερο ποσοστό ακρίβειας. Το τελικό μοντέλο που κατασκευάστηκε διατίθεται και μπορεί να δοκιμαστεί πάνω σε ακουστικά σήματα, και μάλιστα σε πραγματικό χρόνο.

Λέξεις Κλειδιά

Αναγνώριση φωνής, Πεδίο του χρόνου, Χρονοσειρά, Πεδίο της Συχνότητας, Φασματογράφημα, Μετασχηματισμός Fourier, MFCC, Βαθιά Μηχανική Μάθηση, Συνελικτικά Δίκτυα, Αναδρομικά Νευρωνικά Δίκτυα, Μεταφορά Μάθησης.

Abstract

The object of the present diploma thesis is the design, implementation and evaluation of a system that will contribute to the interaction between a person and a robotic vehicle. This interaction will be executed using specific voice commands (eg 'stop', 'go', 'right' ...), through which the direction of the vehicle will be controlled.

Initially, some audio recordings were assembled for each of the commands. These samples were then converted into appropriate representation to be fed into a deep machine learning model. The experimentation concerned both the extraction of features and the synthesis of the neural network architecture.

Regarding feature extraction, a time-field experiment was first conducted with each acoustic signal being examined as a time series. Some weaknesses quickly emerged, which made it imperative to switch to the frequency domain. In this field, after experimentation, the model has emerged, as well as features that achieve the highest accuracy. The final model that was created is available and can be tested on acoustic signals, and even in real time.

Keywords

Speech Recognition, Time series, Spectrogram, Fourier Transform, MFCC, Deep Learning, Recurrent Neural Networks, Convolutional Neural Networks, Transfer Learning

Περιεχόμενα

Ευχαριστίες **1**

Περίληψη **3**

Abstract **5**

Περιεχόμενα **9**

Κατάλογος Σχημάτων **13**

Κατάλογος Πινάκων **15**

1 Εισαγωγή **17**
1.1 Ιστορική Αναδρομή 17
1.2 Εφαρμογές Αναγνώρισης Ομιλίας 19
1.3 Πρόβλημα προς επίλυση 22
1.4 Διάρθρωση Εργασίας 22

2 Θεωρητικό Υπόβαθρο **23**
2.1 Το σήμα φωνής 23
2.2 Ανάλυση στο πεδίο του χρόνου 24
2.2.1 Διαδικασία Δειγματοληψίας 24
2.2.2 Χαρακτηριστικά στο πεδίο του χρόνου 24
2.2.3 Προεπεξεργασία στο πεδίο του Χρόνου 25
2.3 Ανάλυση στο πεδίο της συχνότητας 26
2.3.1 Μετάβαση στο πεδίο της συχνότητας 26
2.3.2 Ανάλυση με Spectrogram 29
2.3.3 Ανάλυση με Mel Spectrograms 31
2.3.4 Ανάλυση με MFCC 33
2.3.5 Ανάλυση με delta - MFCC 33

3 Γενικά περί Μηχανικής Μάθησης **35**
3.1 Γενικά περί Μηχανικής Μάθησης 35

3.2 Αλγόριθμοι Μηχανικής Μάθησης . . . 36
3.2.1 Επιβλεπόμενη Μάθηση . . . 36
3.2.2 Μη επιβλεπόμενη Μάθηση . . . 39
3.3 Ανάλυση Χρονοσειρών . . . 40
3.3.1 Ορισμός . . . 40
3.3.2 Αλγόριθμοι . . . 40
3.4 Νευρωνικά Δίκτυα . . . 42
3.4.1 Νευρώνας Ανθρώπινου Εγκεφάλου . . . 42
3.4.2 Τεχνητός Νευρώνας - Perceptron . . . 42
3.4.3 Πολυεπίπεδα νευρωνικά δίκτυα . . . 43
3.4.4 Συναρτήσεις ενεργοποίησης . . . 44
3.4.5 Έξοδος Νευρωνικού . . . 45
3.4.6 Συναρτήσεις Κόστους . . . 45
3.4.7 Εκπαίδευση νευρωνικού . . . 46
3.4.8 Αλγόριθμοι Βελτιστοποίησης . . . 48
3.4.9 Το πρόβλημα των vanishing gradients . . . 49
3.4.10 Το πρόβλημα της υπερποσαμογής - overfitting . . . 49
3.4.11 Τρόποι Αντιμετώπισης overfitting . . . 50
3.4.12 Αξιολόγηση Απόδοσης Νευρωνικού . . . 52
3.5 Συνελικτικά δίκτυα . . . 53
3.5.1 Ορισμός . . . 53
3.5.2 Επίπεδα Επεξεργασίας . . . 53
3.5.3 Υπερπαράμετοι Συνελικτικού Επιπέδου . . . 55
3.6 Αναδρομικά Δίκτυα . . . 56
3.7 Συνδυασμός Συνελικτικών - Αναδρομικών δικτύων . . . 57
3.8 Μεταφορά Μάθησης . . . 58
3.8.1 EfficientNet . . . 59
3.8.2 ResNet . . . 60
3.8.3 Xception . . . 61

4 Τεχνικές Λεπτομέρειες - Εργαλεία 63

5 Περιγραφή Αρχιτεκτονικής 65
5.1 Καθορισμός Προβλήματος- Στόχου . . . 65
5.2 Επισκόπηση papers - Συγγενείς Εργασίες . . . 65
5.3 Δομή Παρούσας Εργασίας . . . 66
5.4 Συλλογή δεδομένων . . . 66
5.5 Προεπεξεργασία Δεδομένων στο πεδίο του Χρόνου . . . 67
5.6 Εξαγωγή Χαρακτηριστικών . . . 69
5.6.1 Πεδίο του Χρόνου . . . 69
5.6.2 Πεδίο της Συχνότητας . . . 70

5.7 Επιλογή Αλγορίθμου Μάθησης - Αρχιτεκτονικής 72
5.7.1 Πεδίο του Χρόνου 72
5.7.2 Πεδίο της Συχνότητας 72
5.8 Εκπαίδευση 72
5.8.1 Πεδίο του Χρόνου 72
5.8.2 Πεδίο της Συχνότητας 77
5.9 Αξιολόγηση 84
5.9.1 Πεδίο του Χρόνου 84
5.9.2 Πεδίο της Συχνότητας 84

6 Σύνοψη-Μελλοντικές Επεκτάσεις 87
6.1 Σύνοψη 87
6.2 Μελλοντικές Επεκτάσεις 87

Βιβλιογραφία 89

Κατάλογος Σχημάτων

2.1 Απεικόνιση σήματος για την εντολή *three* στο πεδίο του χρόνου ($Fs = 16kHz$) 23
2.2 Εξαγωγή RMS τιμών (K = 80) για την εντολή *three* 25
2.3 Εξαγωγή ZCR τιμών (K = 80) για την εντολή *three* 25
2.4 Εφαρμογή pre-emphasis φίλτρου 26
2.5 Διαφορές απο την εφαρμογή pre-emphasis φίλτρου 26
2.6 Βήματα εξαγωγής χαρακτηριστικών για ένα ακουστικό σήμα 26
2.7 Αμφίπλευρη-Μονόπλευρη αναπαράσταση M/Σ Fourier 28
2.8 Επικάλυψη παραθύρων κατα την εφαρμογή FFT 29
2.9 Εξαγωγή spectrogram για την εντολή *three* 30
2.10 Εφαρμογή συνάρτησης λογαρίθμου στο φασματογράφημα της εντολής *three* . . 30
2.11 Μετατροπή κλίμακας Hertz στην κλίμακα mel 31
2.12 Κατασκευή τριγωνικών φίλτρων για την εξαγωγή mfcc χαρακτηριστικών . . . 32
2.13 Εξαγωγή MFCC χαρακτηριστικών για την εντολή *three* 34
2.14 Εξαγωγή delta MFCC χαρακτηριστικών για την εντολή *three* 34
2.15 Εξαγωγή delta-delta MFCC χαρακτηριστικών για την εντολή *three* 34

3.1 Περιοχές απόφασης με βάση τον αλγόριθμο *knn* 36
3.2 Κατασκευή Δέντρου Απόφασης με βάση τα στοιχεία του πίνακα 37
3.3 Κατασκευή Τυχαίου Δάσους χρησιμοποιώντας Δέντρα Απόφασης 38
3.4 Διαδοχικές (a) - (f) επαναλήψεις του αλγορίθμου k-means. 39
3.5 Υπολογισμός (a) Ευκλείδειας Απόστασης σε σχέση με (b) DTW μεταξύ των δειγμάτων δύο χρονοσειρών. 40
3.6 Δομή νευρικού κυττάρου του ανθρώπου 42
3.7 Δομή Τεχνητού Νευρώνα - Perceptron 43
3.8 Πολυεπίπεδο Νευρωνικό Δίκτυο, συνδυάζοντας τεχνητούς νευρώνες 44
3.9 Συναστήσεις Ενεργοποίησης: ReLU, tanh, sigmoid, αντίστοιχα 45
3.10 Συνάρτηση κόστους log loss για πρόβλημα δυαδικής ταξινόμησης 46
3.11 Υπολογισμός κλίσης για τη συνάρτηση $f(x, y) = x^2 + y^2$ στο [1,1] 47
3.12 Δομή Τεχνητού Νευρώνα - Perceptron 48
3.13 Ταξινόμησης δειγμάτων: Υποπροσαρμογή (αριστερά), Υπερπροσαρμογή (δεξιά) 50
3.14 Πίνακας Σύγχυσης για 3 κατηγορίες 52
3.15 Δομή Συνελικτικού Νευρωνικού Δικτύου (CNN) 53

3.16 Επίδραση διφορετικών φίλτρων σε εικόνες 54
3.17 Εφαρμογή συνέλιξης φίλτρου σε εικόνα εισόδου 54
3.18 Εφαρμογή Υποδειγματοληψίας σε Image Map 55
3.19 Δομή RNN στην κανονική μορφή (αριστερά) και κατόπιν ξεδίπλωσης (δεξιά) . 56
3.20 Συνδυασμός συνελικτικού (αριστερά) - αναδρομικού (δεξιά) δικτύου 58
3.21 Διαδικασία Μεταφοράς Γνώσης απο ένα προεκπαιδευμένο μοντέλο 59
3.22 Επέκταση Νευρωνικού δικτύου (a), ως προς το πλάτος (b), το μήκος (c), την ανάλυση (d) και βέλτιστα ως προς όλες (d) 59
3.23 Επίδραση επέκτασης δικτύου σε δύο εικόνες 60
3.24 Σύκριση training-test error του δικτύου ResNet αναφορικά με το βάθος ενός δίκτύου 60
3.25 Δομικό στοιχείο ResNet 61
3.26 Δομικό στοιχείο Xception 61

5.1 Διαχωρισμός και Οργάνωση Εργασίας σε 6 στάδια 66
5.2 Παρουσιάση τριών κυματομοφών στο Πεδίο του χρόνου για όλες τις εντολές . 68
5.3 Επίδραση εφαρμογής pre-emphasis στο πεδίο του χρόνου 69
5.4 Επίδραση εφαρμογής pre-emphasis στο πεδίο της συχνότητας 69
5.5 Εξαγωγή RMS τιμών στο πεδίο του χρόνου 70
5.6 Εξαγωγή ZCR τιμών στο πεδίο του χρόνου 70
5.7 Εξαγωγή log spectrograms για τα ακουστικά σήματα των εντολών 71
5.8 Εξαγωγή MFCCs για τα ακουστικά σήματα των εντολών 73
5.9 Εξαγωγή delta-MFCCs για τα ακουστικά σήματα των εντολών 74
5.10 Επίδραση του αριθμού των γειτόνων του αλγορίθμου KNN, στο test accuracy 75
5.11 Επίδραση του αριθμού των εκτιμητών του αλγορίθμου Random Forest, στο test accuracy 75
5.12 Επίδραση του αριθμού των εκτιμητών του αλγορίθμου RISE, στο test accuracy 76
5.13 Καμπύλη εκπαίδευσης απλού νευρωνικού δικτύου χρησιμοποιώντας τις RMS τιμές των ακουστικών σημάτων 76
5.14 Συνελικτικό Δίκτυο (CNN) εφαρμοσμένο σε χαρακτηριστικά MFCC 78
5.15 Επίδραση αύξησης βάθους του συνελικτικού δικτύου 78
5.16 Επίδραση του μεγέθους της εικόνας εισόδου 79
5.17 Πειραματισμός με συναρτήσεις ενεργοποίησης στα Συνελικτικά επίπεδα 79
5.18 Πειραματισμός με συναρτήσεις ενεργοποίησης στα υπόλοιπα επίπεδα 80
5.19 Πειραματισμός με βελτιστοποιητές 80
5.20 Μελέτη sgd για περισσότερες εποχές 80
5.21 Πειραματισμός με learning rate 81
5.22 Συνελικτικό Δίκτυο (CNN) ακολουθούμενο από αναδρομικό δίκτυο (RNN) σε MFCC 81
5.23 Πειραματισμός με πλήθος μονάδων RNN 82
5.24 Επαίδευση 20 % των επιπέδων του ResNet 83

5.25 Καμπύλες εκπαίδευσης για διάφορες αρχιτεκτονικές 84
5.26 Πίνακας Σύγχησης για την αριχτεκτονική CNN 85
5.27 Πίνακας Σύγχησης για την αριχτεκτονική CNN-RNN 86
5.28 Πίνακας Σύγχησης για την αριχτεκτονική ResNet-152 86

Κατάλογος Πινάκων

5.1 Πλήθος recordings για κάθε εντολή 67
5.2 Accuracy (%) για Αρχιτεκτονικές- Χαρακτηριστικά, χωρίς βλετιστοποιήση . . 77
5.3 Σύνοψη αρχιτεκτονικών (baseline) που πρόκειται να βελτιστοποιηθούν 77
5.4 Επιδραση του πλήθους των μονάδων RNN στο test accuracy 82
5.5 Διερεύνηση επεκτάσεων ορισμένων αρχιτεκτονικών 83
5.6 Accuracy (%) για Αρχιτεκτονικές- Χαρακτηριστικά, μετά την βλετιστοποιήση 85

Κεφάλαιο 1

Εισαγωγή

Η αναγνώριση φωνητικών εντολών απο ένα υπολογιστικό σύστημα και η επικοινωνία ανθρώπου - μηχανής μέσω της ομιλίας, συνιστούν ένα δυναμικό και συνάμα συνεχώς εξελισσόμενο πεδίο για περισσότερο από 70 χρόνια [1]. Η επινόηση του αλγορίθμου του γρήγορου μετασχηματισμού Fourier (Fast Fourier Transform, FFT) από τους Jim Cooley και John Turkey το 1965 [1], και η ακόλουθη ευρεία διάδοση του σε εφαρμογές γρήγορης συνέλιξης και φασματικής ανάλυσης, κατέρριψαν τα εμπόδια και τους περιορισμούς της αναλογικής τεχνολογίας. Σιγά σιγά, εδραιώθηκαν τα θεμέλια του πεδίου της ψηφιακής επεξεργασίας φωνής, χωρίς το οποίο δε θα ήταν εφικτή η αντίληψη ομιλίας απο τον υπολογιστή. Τα τελευταία χρόνια, οι εξελίξεις στην τεχνολογία ολοκληρωμένων κυκλωμάτων, στους αλγορίθμους DSP (Digital Signal Processing) και στην αρχιτεκτονική υπολογιστών ευθυγραμμίστηκαν και δημιούργησαν πρόσφορο έδαφος για καινοτομία και έρευνα γύρω απο την αναγνώριση ομιλίας.

Σταθμό στην έρευνα, αποτέλεσε και η χρήση τεχνικών βαθιάς μηχανικής μάθησης. Έχοντας στη διάθεση χαρακτηριστικά απο την ψηφιακή επεξεργασία ακουστικών σημάτων, σχεδιάστηκαν και υλοποιήθηκαν αλγόριθμοι οι οποίοι ήταν ικανοί να αναδείξουν τόσο αυτοσυσχετίσεις στα σύνολα των δεδομένων, όσο και διαφορές. Με τον τρόπο αυτό, κατέστη εφικτή η επιτυχής η διαδικασία της μάθησης απο ένα υπολογιστικό σύστημα και η προσομοίωση μιας απο τις βασικότερες λειτουργίες του ανθρώπινου συστήματος, αυτή της νοημοσύνης. Μοντελοποιώντας λοιπόν τον τρόπο που ο ανθρώπινος εγκέφαλος αναλύει και επεξεργάζεται πληροφορίες απο το περιβάλλον,όπως ακουστικά ερεθίσματα, καταλήξαμε σε επιτυχείς εφαρμογές αναγνώρισης ομιλίας. Βέβαια, οι ερευνητές για να καταλήξουν σε συστήματα που επιτελούν έγκυρες προβλέψεις και λαμβάνουν ορθές αποφάσεις, χρειάστηκε να μελετήσουν, να εξετάσουν και να αξιολογήσουν διαφορετικές αρχιτεκτονικές, μια διαδικασία που συνεχίζεται με αμείωτο ενδιαφέρον μέχρι και σήμερα.

1.1 Ιστορική Αναδρομή

Τα πρώτο σύστημα αναγνώρισης ομιλίας, που φέρει τον τίτλο *Audrey*, αναπτύχθηκε στα μέσα του 19ου αιώνα (1952) και συγκεκριμένα απο ερευνητές του Bell Labs. Βασικό στόχο της υλοποίησης αποτέλεσε η αναγνώριση ψηφίων (zero - nine), ώστε να εφαρμοστεί στην

τηλεφωνία. Το σύστημα αυτό, όταν δοκιμάστηκε σε συγκεκριμένους ομιλητές, σημείωσε ικανοποιητικά ποσοστά ακρίβειας (70 - 80 %) [16]. Ωστόσο δεν οδήγησε στην αντικατάσταση των κοινών πληκτρολογίων, που διέθεταν μέχρι τότε οι συσκευές τηλεφωνίας. Άλλωστε, δεν ήταν αυτός ο βασικός λόγος που κατασκευάστηκε. Σύμφωνα με μαρτυρίες, μεσω της αναγνώρισης φωνημάτων που πετύχαινε, θα μειωνόταν το εύρος ζώνης (bandwidth) καθώς, αντί για ολοκληρα τα ακουστικά σήματα, χρειάζοταν να μεταδοθούν μονο τα φωνήματα ως σύμβολα. Το γεγονός όμως ότι η τηλεφωνία έγινε ψηφιακή (γύρω στο 1970), φέρνοντας γρήγορες ταχύτητες μετάδοσης με χαμηλό κόστος, αποτέλεσε τροχοπέδη για τη καθιέρωση του συστήματος *Audrey*. Άλλοι λόγοι που συνέβαλαν σε αυτό, ήταν η δυσκολία του να ερμηνεύσει άγνωστους μέχρι τότε ήχους, η πολυπλοκότητα, και φυσικά η υψηλή κατανάλωση ενέργειας που σημείωνε. Άλλωστε στην εποχή αυτή τα υπολογιστικά συστήματα χαρακτηριζόταν από περιορισμένες δυνατότητες αποθήκευσης και μειωμένες ταχύτητες επεξεργασίας δεδομένων. Ωστόσο πρόκειται για ένα σημαντικό επίτευγμα, που σηματοδότησε την αφετηρία της έρευνας σε ένα καινοτόμο πεδίο, αυτό της αναγνώρισης ομιλίας. [16]

Το δεύτερο σύστημα αναγνώρισης ομιλίας αναπτύχθηκε απο την IBM στις αρχές της δεκαετίας του 1960, και έμεινε γνωστό στην ιστορία με το όνομα *Shoebox*. Οι λειτουργικότητες που προστέθηκαν σε σχέση με το σύστημα *Audrey*, ήταν η αναγνώριση φωνητικών εντολών που αφορούσαν πράξεις ('plus', 'minus', 'total', 'subtotal'), καθώς και η εκτέλεση τους σε πραγματικό χρόνο. Πρόκειται για μια κυκλωματική διάταξη, η οποία περιλάμβανε μεταξύ άλλων τρία αναλογικά φίλτρα (high - middle - low pass) και λυχνίες LED για την απεικόνιση τόσο των εντολών όσο και των αποτελεσμάτων. Ο χρήστης μπορούσε να προφέρει μια απο τις εντολές αυτές στο μικρόφωνο της συσκευής, καθεμία απο τις οποίες ενεργοποιούσε διαφορετική λυχνία LED. Το ακουστικό σήμα διαιρούνταν σε τρία παράθυρα, κάθε ένα απο τα οποία μπορούσε να ερμηνευτεί ως high, middle, low pitched ανάλογα με το φίλτρο που ενεργοποιούσε. Για παράδειγμα το ψηφίο 'five' μπορούσε να κωδικοποιηθεί σαν 'High-Middle-High' ενώ το ψηφίο 'zero' σαν 'High-Middle-Low'. Το δεύτερο λοιπόν σύστημα αντίληψης ανθρώπινης ομιλίας, όχι μόνο πέτυχε την αναγνώριση περισσότερων εντολών, αλλά μέσω της απάντησης σε μαθηματικές πράξεις έθεσε και τις βάσεις για τη γόνιμη επικοινωνία ανθρώπου - υπολογιστή. Βέβαια η επικοινωνία αυτή πραγματοποιούνταν μέσω της αναγνώρισης μεμονωμένων εντολών και όχι προτάσεων. [18]

Το 1971, πραγματοποιήθηκε η μετάβαση απο την αποκωδικοποίηση μεμονωμένων εντολών στην ερμηνεία ολοκληρωμένων προτάσεων. Πρόκειται για ένα σημαντικό άλμα, που ξεκίνησε απο το ερευνητικό γραφείο του Υπουργείου Άμυνας των ΗΠΑ. Στο πρόγραμμα που εκπόνησε το συγκεκριμένο Υπουργείο, συμμετείχαν εταιρείες όπως η IBM και ακαδημαϊκά ιδρύματα, μεταξύ των οποίων το Stanford και το πανεπιστήμιο Carnegie Mellon. Το σύστημα που αναπτύχθηκε έμεινε γνωστό ως *Harpy* και το λεξιλόγιο που μπορούσε να αναγνωρίσει περιλάμβανε πάνω απο 1000 λέξεις. Γρήγορα όμως οι ερευνητές ήρθαν αντιμέτωποι με προβλήματα, που ανακύπτουν απο την ερμηνεία προτάσεων, όπως η σύγχυση φράσεων ή η δυσκολία του συστήματος να διαχωρίσει τις επιμέρους λέξεις. Για παράδειγμα, όπως αναφέρει ο καθηγητής επιστήμης υπολογιστών του πανεπιστημίου Carnegie Mellon, Alexander Waibel, η φράση "euthanasia" θα μπορούσε να αποκωδικοποιηθεί και ως "youth in Asia". Παρόλες τις δυσκο-

λίες το σύστημα μπορούσε με αξιόλογα ποσοστά ακρίβειας να αναγνωρίσει πληθώρα λέξεων. Πλέον είχε ανοίξει ο δρόμος για τη χρήση γλωσσικών μοντέλων στην αποκωδικοποίηση των προτάσεων, με το σύστημα *Harpy* να αποτελεί τον πρόγονο των σύγχρονων συστημάτων. [16]

Περίπου δέκα χρόνια μετά, το 1980, υλοποιήθηκε από την IBM το σύστημα *Tangora*, το οποίο έθεσε ψηλά τον πήχη για την αναγνώριση ενός μεγαλύτερου συνόλου λέξεων. Πιο ειδικά καταρρίφθηκε για πρώτη φορά το φράγμα των 20.000 λέξεων. Αυτό κατέστη εφικτό χρησιμοποιώντας τα Κρυφά Μαρκοβιανά Μοντέλα (HMMs), τα οποία και κρίθηκαν κατάλληλα για επεξεργασία ακολουθιακής πληροφορίας, όπως ένα ακουστικό σήμα [16]. Ωστόσο όμως το πρόβλημα της αποθήκευσης και της διαχείρισης μεγάλων δεδομένων για εκπαίδευση αυτών των μοντέλων, συνέχειζε να παρέμενει στο προσκήνιο.

Στις αρχές του 21ου αιώνα, περίοδο που σημαδεύτηκε απο το cloud computing, κατέστη εφικτή η συγκέντρωση, αποθήκευση και επεξεργασία μεγαλύτερων συνόλων δεδομένων. Συγκεκριμένα η Google, αξιοποιώντας αυτή τη τεχνολογία, κυκλοφόρησε την αναζήτηση στον παγκόσμιο ιστό μέσω φωνητικών εντολών για συσκευές Iphone. Έπειτα, το 2011, η Apple παρουσίασε έναν εικονικό βοηθό με όνομα *SIRI* με αρκετά προηγμένες λειτουργικότητες στις συσκευές της, όπως ενημέρωση για τον καιρό, διαχείριση τηλεφωνικών κλήσεων, προσθήκη σημειώσεων, περάτωση οικονομικών συναλλαγών ακόμα και έλεγχο συμβατών οικιακών συσκευών - όλα με τη χρήση φωνητικών εντολών. Βέβαια ο εικονικός βοηθός μπορούσε να συνδεθεί μόνο με συσκευές Apple. Λίγα χρόνια μετά, η Amazon παρουσίασε την Alexa, έναν εικονικό βοηθό με αντίστοιχες λειτουργικότητες.

Εν κατακλείδι, η έρευνα και ο πειραματισμός γύρω απο συστήματα αναγνώρισης ομιλίας σημάδεψαν τον επιστημονικό χώρο, ηδη απο τα μέσα του 20ου αιώνα. Η έρευνα αυτή, συνεχίζεται με αδιάλειπτο ενδιαφέρον μέχρι και σήμερα. Η πλήρης κατανόηση ήχου ακόμα και σε θορυβώδη περιβάλλοντα, η ερμηνεία του σημασιολογικού περιβάλλοντος στο οποίο ανήκει η λέξη, η γρήγορη απόκριση των συστημάτων, η ταύτιση όρων που προφέρονται διαφορετικά (π.χ. αμερικάνικη - σκωτσέζικη προφορά) αλλά και ο σεβασμός στα δεδομένα προσωπικού χαρακτήρα είναι προκλησεις τις οποίες καλούνται να επιλύσουν τα state of the art μοντέλα. Νέες προκλήσεις που κρίνεται αναγκαίο να αντιμετωπιστούν άμεσα, αν αναλογιστεί κανείς ότι οι εφαρμογές της αναγνώρισης ομιλίας καλύπτουν ένα ευρύ φάσμα της ζωης του ανθρώπου. [19]

1.2 Εφαρμογές Αναγνώρισης Ομιλίας

Όπως έγινε αντιληπτό από την ιστορική αναδρομή, οι ερευνητές έστρεψαν απο νωρίς το ενδιαφέρον τους στην υλοποίηση συστημάτων, τα οποία αντιλαμβάνονται την ανθρώπινη ομιλία. Κινητήριο δύναμη αποτέλεσε το όραμα για εφαρμογές σε διάφορες πτυχές της ανθρώπινης ζωής. Πιο ειδικά, οι εφαρμογές αυτές αφορούν, μεταξύ άλλων, τους τομείς [20]:

Υγειονομική περίθαλψη: Νέα δεδομένα στο πεδίο της ιατρικής γνωμάτευσης φέρνει η αναγνώριση ομιλίας. Πιο ειδικά, μέσω της μετατροπής ομιλίας σε κείμενο καθίσταται εφικτή η αυτοματοποιημένη καταγραφή των συμπτωμάτων ή άλλων πληροφοριών που αφορούν τους

ασθενείς. Χαρακτηριστικό παράδειγμα αποτελεί η συστηματοποιημένη συλλογή πληροφοριών, και η αποθήκευση τους στο ηλεκτρονικό αρχείο υγείας (EHR) απλά με τη χρήση των φωνητικών εντολών [21]. Έτσι, άτομα με ειδικές ανάγκες ακόμα και ηλικιωμένοι που αδυνατούν να χειριστούν το πληκτρολόγιο ή το ποντίκι, μπορούν πλέον να απολαύσουν τα εργονομικά οφέλη αυτής της καινοτομίας. Ταυτόχρονα με τη συλλογή περισσότερων διαγνώσεων, επωφελείται σύσσωμη η ιατρική κοινότητα, αφού έχουν στη διάθεση τους περισσότερα δεδομένα για κάποια ασθένεια. Άλλα παραδείγματα στον τομέα αυτό, αποτελούν εικονικοί βοηθοί που υποβοηθουν το προσωπικό ενός νοσοκομείου. Ερωτήσεις όπως "πόσες κλίνες είναι διαθέσιμες" ή "πόσοι ασθενείς είναι σε ένα δωμάτιο", μπορούν να απαντηθούν άμεσα και να εξοικονομήσουν πολύτιμο χρόνο απο το ιατρικό προσωπικό.

Τραπεζικές Συναλλαγές: Η τεχνολογία της αυτόματης αναγνώρισης ομιλίας δεν μπορούσε να μην υιοθετηθεί απο τράπεζες. Έτσι, σε διάφορες υπηρεσίες ή εφαρμογές ενσωματώνονται προηγμένες λειτουργικότητες, που καθιστούν πιο εύχρηστη την διεπαφή με αυτές. Αντιπροσωπευτικά δείγματα αποτελούν τράπεζες (N26, Bank of Canada, Garanti Bank) και εταιρείες (Venmo, PayPal) [20], οι οποίες παρέχουν στους πελάτες την δυνατότητα να ενημερώνονται για το ιστορικό πληρωμών, ή να προβαίνουν σε μεταφορές χρηματικών ποσών απλά με την ομιλία τους.

Ηλεκτρονικό Εμπόριο: Στις μέρες, μας έχει αλλάξει ο τρόπος με τον οποίο οι καταναλωτές προσεγγίζουν τα προϊόντα τους. Διάφορες πλατφόρμες προσφέρουν την δυνατότητα στους αγοραστές να πλοηγηθούν μέσω φωνητικών εντολών σε ηλεκτρονικά καταστήματα, να φιλτράρουν τις επιλογές τους ως προς τα επιθυμητα κριτηρια και εν τέλει να καταλήξουν στην αγορά των προϊόντων τους, με απλό και αποτελεσματικό τρόπο. Απο την πλευρά των επιχειρήσεων τώρα, μπορούν να συλλέξουν χρήσιμες πληροφορίες για τους χρήστες, δημιουργώντας φόρμες που επιδέχονται φωνητικές απαντήσεις.

Οικιακή Χρήση: Ήδη απο τις αρχές του 21ου αιώνα, αρκετά σπίτια φιλοξενούν εφαρμογές της τεχνητής νοημοσύνης. Ηλεκτρικές συσκευές, φωτιστικά, πόρτες - παράθυρα, κάμερες, σώματα θέρμανσης αποτελούν ένα μικρό δείγμα πραγμάτων, που μπορούν να ελεγχθούν απλά επικοινωνώντας στην ανθρώπινη γλώσσα. Κάτι τέτοιο αφενός προκαλεί άνεση στους παρόχους, αφετέρου προσφέρει και άλλα οφέλη όπως εξοικονόμηση ενέργειας ακόμα και ασφάλεια. Αρκεί κανείς να αναλογιστεί, ότι ένα τέτοιο σύστημα μπορεί να αναγνωρίσει και τον ομιλητή, οπότε να αποτρέπεται ο έλεγχος του απο τρίτους.

Διαδίκτυο των πραγμάτων: Το IoT, πέρα απο την οικιακή χρήση, συναντάται σε πληθώρα εφαρμογών. Στις μέρες μας βρίσκεται σε πλήρη ευθυγράμμιση με την τεχνητή νοημοσύνη και πιο ειδικά με τεχνικές επεξεργασίας και κατανόησης της φυσικής γλώσσας. Έξυπνες πόλεις και δίκτυα (smart grids), καλλιέργειες, βιομηχανικές διατάξεις, ακόμα και μέσα μεταφοράς είναι μερικά παραδείγματα όπου ο άνθρωπος μπορεί να επικοινωνήσει στην ίδια του τη γλώσσα με τις ενσωματωμένες συσκευές.

Μεταφορές - Συγκοινωνίες: Η επανάσταση στον τομέα της μηχανικής μάθησης, με εφαρμογές στην αναγνώριση ομιλίας, έγινε αισθητή ακόμα και στο πεδίο των μεταφορών - συγκοινωνιών. Σήμερα λοιπόν οι άνθρωποι μπορούν να πληροφορηθούν για δρομολόγια, να ενημερωθούν για τις συντομότερες διαδρομές, απλά θέτοντας μια ερώτηση σε έναν εικονι-

κό βοηθό. Προφανώς η διαδικασία αυτή προσφέρει άνεση στους χρήστες και εξοικονόμηση χρόνου, ωστόσο τα πραγματικά οφέλη αφορούν κυρίως άτομα με ειδικές ανάγκες, όπως προβλήματα όρασης.

Αυτόνομη οδήγηση: Η "έξυπνη" οδήγηση είναι μια διαδικασία, που τα τελευταία χρόνια έχει μονοπωλήσει το ενδιαφέρον των ερευνητών - μηχανικών. Βασικός τους στόχος, είναι ο οδηγός να βρίσκεται απόλυτα προσηλωμένος στον έλεγχο του αμαξιού, εκμηδενίζοντας έτσι τον κίνδυνο πρόκλησης κάποιου τροχαίου ατυχήματος. Μπορεί λοιπόν να ενημερωθεί για τα μηνύματα του, να χειριστεί το ραδιόφωνο, να πληροφορηθεί για οδηγίες πλοήγησης και πολλά άλλα, χρησιμοποιώντας μόνο και μόνο μερικές φωνητικές εντολές.

Αεροδιαστημική: Η NASA συμπεριέλαβε στο διαστημικό σκάφος Mars Polar Lander μια τεχνολογία αναγνώρισης ομιλίας. Πρόκειται για ένα μικρόφωνο το οποίο δημιουργήθηκε απο την εταιρεία Sensory, Inc. [22]

Εκπαίδευση: Το εκπαιδευτικό σύστημα δε θα μπορούσε να μείνει ανεπηρέαστο απο τις ραγδαίες τεχνολογικές εξελίξεις. Εικονικοί βοηθοί μπορούν, συμπληρωματικά με το έργο του δασκάλου, να συμβάλλουν στην εκμάθηση κάποιας ξένης γλώσσας ακόμα και να αξιολογήσουν την προφορά του εκάστοτε μαθητή με μεθόδους NLP. Επίσης, αξιοποιώντας τεχνικές μετατροπής ομιλίας σε κειμένου, μπορούν να συμμετάσχουν πλήρως στην εκπαιδευτική διαδικασία και άτομα που δυσκολεύονται στην γραφή (λόγω δυσλεξίας ή λόγω κάποιας κάκωσης στο χέρι).

Τέχνη: Είναι αδιαμφισβήτητο ότι η τεχνητή νοημοσύνη μπορεί να προάγει ακόμα και την τέχνη. Για παράδειγμα μπορούν, άτομα με προβλήματα όρασης ακόμα και κακώσεις στα άκρα, να επιδοθούν στην συγγραφή κάποιου ποιήματος ή λογοτεχνικού κειμένου, χρησιμοποιώντας απλά speech to text τεχνικές. Επίσης καθίσταται εφικτή η πλοήγηση σε χώρους πολιτισμού όπως μουσεία, με εικονικούς βοηθούς που υποβοηθούν στο έργο του ξεναγού. Τα παραπάνω αποτελούν μόνο ένα μικρό δείγμα απο εφαρμογές AI και speech recognition στην τέχνη.

Στρατιωτικοί Λόγοι: Ουσιαστικές προσπάθειες έχουν αφιερωθεί την τελευταία δεκαετία στη δοκιμή και την αξιολόγηση της αναγνώρισης ομιλίας σε μαχητικά αεροσκάφη. Ιδιαίτερη σημασία έχουν το πρόγραμμα αναγνώρισης ομιλίας των ΗΠΑ για τα αεροσκάφη Advanced Fighter Technology Integration (AFTI) / F-16 (F-16 VISTA) [23]. Σε αυτά τα προγράμματα, τα αναγνωριστικά ομιλίας λειτουργούν επιτυχώς σε μαχητικά αεροσκάφη, με εφαρμογές όπως: ρύθμιση ραδιοσυχνοτήτων, διοίκηση συστήματος αυτόματου πιλότου, ρύθμιση συντεταγμένων διεύθυνσης ακόμα και παραμέτρων απελευθέρωσης όπλων.

Δημοσιογραφία - καταγραφή γεγονότων: Ένα απο τα προβλήματα που αντιμετωπίζουν οι δημοσιογράφοι, καλείται να επιλύσει σήμερα η χρήση τεχνικών speech to text. Πιο ειδικά, η πλήρης κάλυψη της ροής των γεγονότων αλλά και οι ρυθμοί με τους οποίους αλλάζουν τα δεδομένα και έρχονται στο προσκήνιο νέες εξελίξεις, πολλές φορές επιβάλλουν τη χρήση τέτοιων συστημάτων. Συστήματα τα οποία, μετατρέποντας τις ομιλίες απευθείας σε κείμενο, θα εξοικονομήσουν πολυτιμο χρόνο απο τους δημοσιογράφους, συμβάλλοντας έτσι στην καταγραφή μεγαλύτερου όγκου δεδομένων και στη διάθεση περισσότερης πληροφορίας στους πολίτες.

Εγκληματολογία - Ασφάλεια: Η αναγνώριση προσώπου απο την ομιλία, μπορεί να συμ-

βάλλει θετικά προς τη κατεύθυνση της αναζήτησης ενός δράστη. Ο Παγκόσμιος Ιστός και ειδικά πλατφόρμες όπως Facebook, YouTube μπορούν να αποτελέσουν πηγή δεδομένων με τα οποία θα συγκριθεί το υλικό των εγκληματολόγων. Μια τέτοια έρευνα ωστόσο, καθώς διαχειρίζεται δεδομένα προσωπικού χαρακτήρα, εγείρει νομικά θέματα.

Υποστήριξη ΑΜΕΑ: Κοινό παράγοντα σε πολλούς τομείς που αναφέρθηκαν παραπάνω, αποτέλεσε η υποβοήθηση ατόμων με ειδικές ανάγκες. Πρόκειται για καινοτομίες που απλοποιούν τη χρήση διαφόρων υπηρεσιών και καθιστούν εφικτή τη πρόσβαση σε διάφορες υπηρεσίες απο ΑΜΕΑ.

1.3 Πρόβλημα προς επίλυση

Από όλες τις παραπάνω εφαρμογές, η παρούσα διπλωματική εστιάζει στην τελευταία εφαρμογή των συστημάτων αναγνώρισης ομιλίας, και ειδικά στην υποστήριξη ΑΜΕΑ. Στόχος είναι ο σχεδιασμός, η υλοποίηση καθώς και η αξιολόγηση ενός συστήματος το οποίο θα συμβάλει στην γόνιμη αλληλεπίδραση του ατόμου με ένα ρομποτικό όχημα. Η αλληλεπίδραση αυτή θα πραγματοποιείται με την χρήση συγκεκριμένων φωνητικών εντολών μέσω των οποίων θα ελέγχεται η κατεύθυνση του οχήματος. Πιο αναλυτικά, τα επιμέρους στάδια που ακολουθήθηκαν για την εκπόνηση της Εργασίας, αναλύονται συνοπτικά στην επόμενη Παράγραφο.

1.4 Διάρθρωση Εργασίας

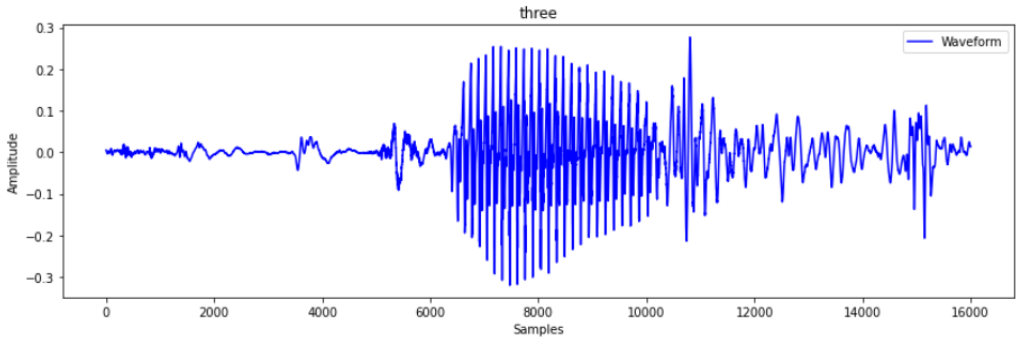
Η παρούσα Διπλωματική Εργασία διαρθρώνεται σε 7 κεφάλαια. Στο Κεφάλαιο 1, πραγματοποιείται μια εισαγωγή στο θέμα, παρουσιάζονται κάποιοι σταθμοί στην έρευνα που συντελέστηκε τον περασμένο αιώνα, δηλώνεται το προς επίλυση πρόβλημα, και τέλος καταγράφεται η δομή της Εργασίας. Στο Κεφάλαιο 2, καλύπτεται το θεωρητικό υπόβαθρο που απαιτείται ώστε να είναι δυνατή η κατανόηση εννοιών γύρω απο την σύνθεση - παραγωγή ανθρώπινης ομιλίας καθώς και της επεξεργασίας - μοντελοποίησης της απο ένα υπολογιστικό σύστημα. Στο Κεφάλαιο 3, επεξηγούνται βασικές θεωρητικές έννοιες αναφορικά με τους αλγορίθμους, ή τις αρχιτεκτονικές που θα διερευνηθούν για την επίλυση του προβλήματος. Το Κεφάλαιο 4, παρέχει μια επισκόπηση στα προγραμματιστικά εργαλεία που χρησιμοποιούνται για την υλοποίηση του συστήματος. Το Κεφάλαιο 5, αναλύει τα εργαλεία καθώς και τη διαδικασία που ακολουθήθηκε σε πρακτικό επίπεδο για την σχεδίαση υλοποίηση, και την αξιολόγηση του τελικού συστήματος. Το Κεφάλαιο 6, συνοψίζει και ολοκληρώνει την παρούσα εργασία με ορισμένες κατευθύνσεις και μελλοντικές επεκτάσεις για την εξέλιξη του συγκεκριμένου συστήματος.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Το σήμα φωνής

Κύριο σκοπό της ομιλίας αποτελεί η επικοινωνία, δηλαδή η ανταλλαγή μηνυμάτων μεταξύ ενός ομιλητή και ενός ακροατή. Τόσο στην ανθρώπινη ομιλία όσο και στην επικοινωνία ηλεκτρονικών συστημάτων, η πληροφορία (σήμα φωνής) που μεταδίδεται κωδικοποιείται σε μια μορφή συνεχώς μεταβαλλόμενης (αναλογικής) κυματομορφής. Η κυματομορφή αυτή, στη θεμελιώδη του αναπαράσταση, περιγράφει πώς μεταβάλλεται το πλάτος (Amplitude) αναφορικά με το χρόνο (time).



Σχήμα 2.1: Απεικόνιση σήματος για την εντολή *three* στο πεδίο του χρόνου ($Fs = 16kHz$)

Παρατηρεί κανείς ότι στον οριζόντιο άξονα είναι τα δείγματα (samples) σε χρονική σειρά ($1sec$), και στον κατακόρυφο το πλάτος. Στον κατακόρυφο άξονα παρατηρούνται και αρνητικές τιμές, γεγονός που έρχεται σε αντιπαράθεση με τον όρο "πλάτος", στη γενική βέβαια περίπτωση. Λαμβάνοντας υπόψιν το γεγονός ότι ο ήχος δεν είναι τίποτα άλλο από ένα κύμα "πίεσης", όπου υπάρχουν περιοχές με υψηλότερη και χαμηλότερη πίεση, τότε μπορεί να γίνει κατανοητή η ερμηνεία του αρνητικού πλάτους. Πιο ειδικά, αν η απουσία ήχου δεν δημιουργεί πίεση (μηδενική), τότε η πρόκληση υψηλότερης πίεσης ερμηνεύεται με θετικό πρόσημο και χαμηλότερης με αρνητικό. Απο την αναπαράσταση αυτη (πλάτος ακουστικού σήματος συναρτήσει του χρόνου), είναι δυνατόν να εξαχθούν χαρακτηριστικά που θα συμβάλλουν στην αναγνώριση ομιλίας.

2.2 Ανάλυση στο πεδίο του χρόνου

2.2.1 Διαδικασία Δειγματοληψίας

Προτού εξαχθούν τα χαρακτηριστικα στο πεδίο του χρόνου, προηγείται μια διαδικασία που καλείται υποδειγματοληψία. Πιο ειδικά, ο αριθμός των δειγμάτων σε ένα αναλογικό σήμα τείνει στο άπειρο. Το γεγονός αυτό καθιστά αδύνατη την αποθήκευση αλλά και τη διαχείριση του απο ένα υπολογιστικό σύστημα. Για το λόγο αυτό επιλέγεται ένας πεπερασμένος αριθμός συνεχόμενων δειγμάτων, με το σήμα να μετατρέπεται απο την αναλογική στη ψηφιακή του μορφή. Το σήμα αυτό στην συνέχεια αναφέρεται ως $s(k)$.

2.2.2 Χαρακτηριστικά στο πεδίο του χρόνου

Έστω $s(k)$ το σήμα που αναπαριστά το πλάτος του k-οστού δείγματος, Kτο μέγεθος του παραθύρου (ή το πλήθος των δειγμάτων που το αποτελούν), και t ο αριθμός του παραθύρου. Μεταξύ άλλων, μπορεί κανείς πλέον να ξεχωρίσει τα ακόλουθα χαρακτηριστικά:

Amplitude Envelope (AE): Στόχος της περιβάλλουσας πλάτους είναι να εξαχθεί το μέγιστο πλάτος μέσα στα πλαίσια (frames), στα οποία διαχωρίζεται το σήμα. Για κάθε πλαίσιο t μεγέθους K, η περιβάλλουσα υπολογίζεται από τον τύπο:

$$AE_t = \max_{k=t\cdot K}^{(t+1)\cdot K-1} s(k) \tag{2.1}$$

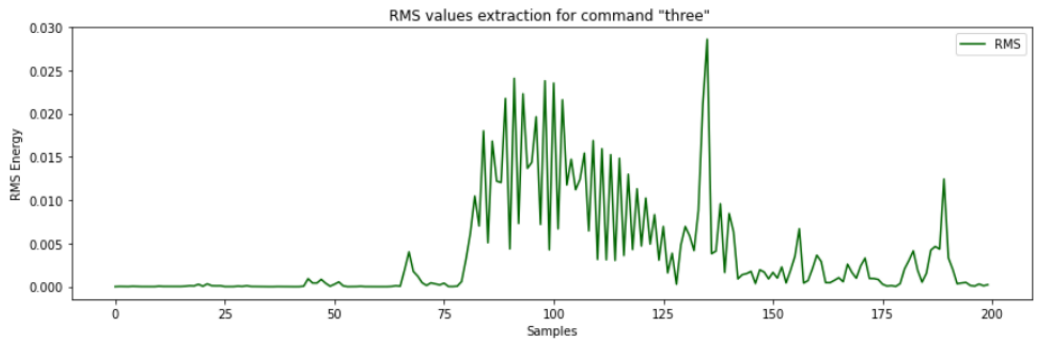
Η διαδικασία αυτή μπορεί να χρησιμοποιηθεί σε εφαρμογές, όπως η μελέτη του επιπέδου έντασης ενός ακουστικού σήματος, η ανίχνευση κάποιας σημαντικής πληροφορίας στον ήχο (φωνήματα, νότες, κλπ), η κατηγοριοποίηση μουσικών ειδών κλπ. Ωστόσο η μέθοδος αυτή παρουσιάζει ευαισθησία ως προς τις ακραίες τιμές, μπορεί δηλαδή να επιλεχθεί κάποια τιμή που δεν είναι αντιπροσωπευτική του frame.

Root Mean Square Energy (RMS):Αντίστοιχα με την περιβάλλουσα πλάτους, το σήμα διαμερίζεται σε παράθυρα μεγέθους K. Τα επιμέρους πλάτη μέσα σε αυτό το παράθυρο τετραγωνίζονται και στην συνέχεια αφού αθροιστούν διαιρούνται με το μέγεθος του παραθύρου. Λαμβάνοντας τη τετραγωνική ρίζα αυτής της ποσότητας, καταλήγουμε στην ενέργεια RMS. Τα παραπάνω εκφράζονται μαθηματικά ως εξής:

$$RMS_t = \sqrt{\frac{1}{K} \cdot \sum_{k=t\cdot K}^{(t+1)\cdot K-1} s(k)^2} \tag{2.2}$$

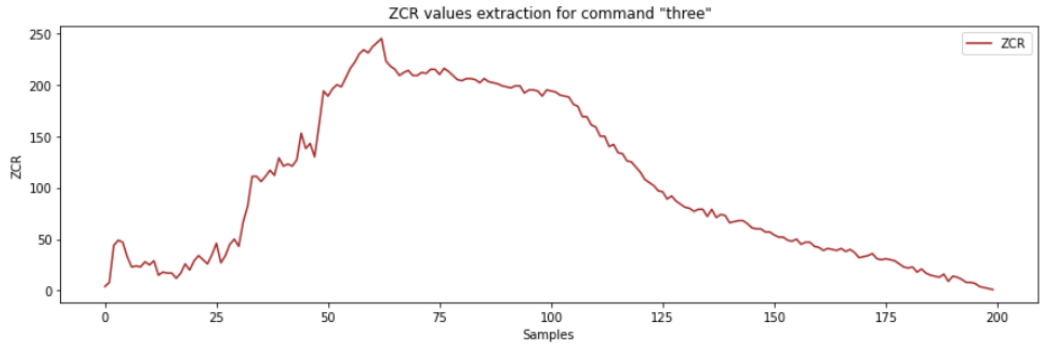
Η μέθοδος αυτή, λαμβάνει υπόψιν το σύνολο των δειγμάτων ενός frame και όχι απλά τη μέγιστη τιμή. Οι εφαρμογές είναι αντίστοιχες με όσες αναφέρθηκαν για την περιβάλλουσα πλάτους.

Zero Crossing Rate (ZCR): Μια διαφορετική προσέγγιση, στην εξαγωγή χαρακτηριστικών (στο πεδίο του χρόνου), αποτελεί ο ρυθμός εναλλαγής προσήμου. Όπως και προηγουμένως, το σήμα εξετάζεται σε παράθυρα. Σε κάθε παράθυρο λοιπόν, αθροίζεται η απόλυτη



Σχήμα 2.2: Εξαγωγή RMS τιμών (K = 80) για την εντολή *three*

διαφορά του δείγματος με το επόμενο του. Στην συνέχεια αφού διαιρείται δια του 2, λαμβάνεται η τιμή ZCR για το εν λόγω παράθυρο. Η διαδικασία συνεχίζεται και για τα υπόλοιπα παράθυρα. Μαθηματικά τα παραπάνω εκφράζονται απο τη σχέση 2.3:



Σχήμα 2.3: Εξαγωγή ZCR τιμών (K = 80) για την εντολή *three*

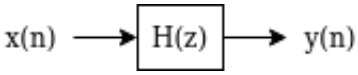
$$ZCR_t = \frac{1}{2} \cdot \sum_{k=t\cdot K}^{(t+1)\cdot K-1} |\mathrm{sgn}(s(k)) - \mathrm{sgn}(s(k-1))| \tag{2.3}$$

Τα παραπάνω συνοψίζουν κάποια βασικά χαρακτηριστικά απο την ανάλυση στο πεδίο του χρόνου. Ένα πεδίο που δεν περιλαμβάνει ωστόσο καμία πληροφορία για το ρυθμο (rate) του σήματος. Με κατάλληλους μετασχηματισμούς λοιπόν, το σήμα μπορεί να αναπαρασταθεί σε άλλους χώρους - πεδία, όπως ο χώρος των συχνοτήτων.

2.2.3 Προεπεξεργασία στο πεδίο του Χρόνου

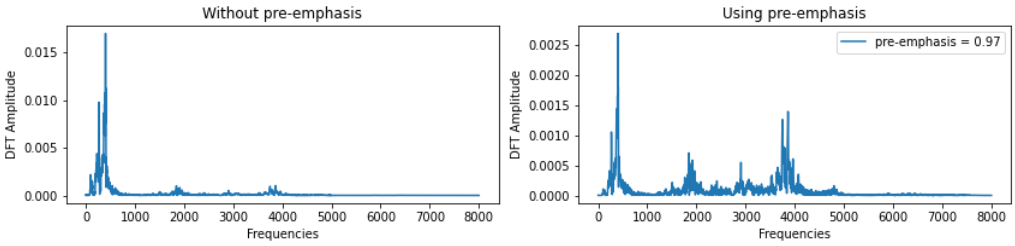
Προτού το σήμα $x(n)$ αναλυθεί στο πεδίο των συχνοτήτων, εφαρμόζεται ένα υψιπερατό pre-emphasis φίλτρο $H(z)$, και έτσι προκύπτει το σήμα $y(n)$ (βλ. Σχήμα 2.4). Ο μετασχηματισμός Z του φίλτρου είναι: $H(z) = 1 - az^{-1}$, οπότε το σήμα που προκύπτει ορίζεται απο τη σχέση: $y(t) = x(t) - ax(t-1)$.

Τυπικές τιμές της παραμέτρου α ειναι 0.95 - 0.97. Η διαδικασία αυτή πραγματοποιείται κυρίως για 2 λόγους: Πρώτος λόγος είναι η ενίσχυση υψηλών συχνοτήτων. Χωρίς pre-emphasis, οι υψηλές συχνότητες έχουν μικρότερα πλάτη από ότι οι χαμηλές (βλ. Σχήμα 2.5).



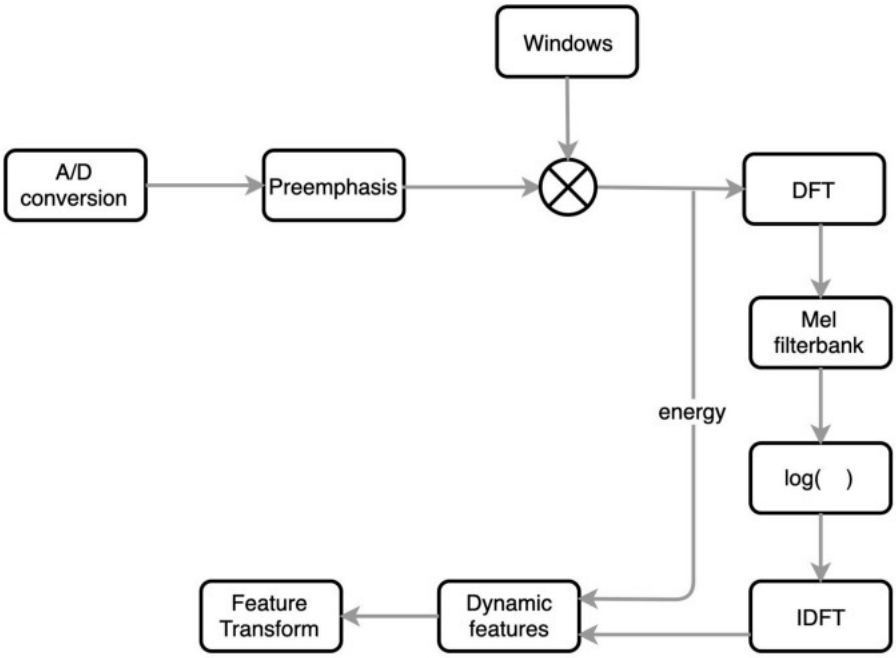
Σχήμα 2.4: Εφαρμογή pre-emphasis φίλτρου

Δεύτερος λόγος είναι ότι μπορεί να βελτιωθεί ο σηματοθορυβικός λόγος SNR.



Σχήμα 2.5: Διαφορές απο την εφαρμογή pre-emphasis φίλτρου

2.3 Ανάλυση στο πεδίο της συχνότητας



Σχήμα 2.6: Βήματα εξαγωγής χαρακτηριστικών για ένα ακουστικό σήμα

2.3.1 Μετάβαση στο πεδίο της συχνότητας

Χρησιμοποιώντας τον Μετασχηματισμο Fourier, μια θεμελιώδη έννοια των Σημάτων και Συστημάτων, το σήμα πλέον μπορεί να αναλυθεί στο πεδίο των συχνοτήτων. Πιο ειδικά αναπαρίσταται σαν μια συνάρτηση του πλάτους (amplitude) σε σχέση με τη συχνότητα (frequency).

Η διαδικασία που ακολουθείται για τη μετάβαση στο πεδίο της συχνότητας και την εξαγωγή χαρακτηριστικών περιγράφεται ακολουθως μέσα απο 6 Βήματα:

Βήμα 0 Εφαρμογή pre-emphasis, όπως περιγράφηκε παραπάνω.

Βήμα 1 (Παραθυροποίηση Ακουστικών Σημάτων): Ο αλγόριθμος FFT, όπως θα δούμε και παρακάτω, θα εφαρμοστεί σε ένα σήμα με πεπερασμένο μήκος. Εν γένει, το σήμα στο οποίο εφαρμόζεται ο συγκεκριμένος αλγόριθμος, αποτελεί περίοδο μιας ευρύτερης ακολουθίας (σήματος), της οποίας το αρχικό και το τελικό σημείο πρέπει να ταυτίζονται. Στη πράξη όμως δεν είναι όλα τα σήματα περιοδικά. Για το λόγο αυτό εφαρμόζουμε παραθυροποίηση. Μέσω αυτής της διαδικασίας, το σήμα πολλαπλασιάζεται με ένα παράθυρο, το πλάτος του οποίου μειώνεται σταδιακα προς τα άκρα της ακολουθίας. Κάτι τέτοιο, έχει ως αποτέλεσμα η αρχική τιμή του σήματος και η τελική να ταυτίζονται. Ένας άλλος λόγος, για τον οποίο εφαρμόζουμε παραθυροποίηση, ειναι ο περιορισμός του φαινομένου *spectral leakage*. Πρόκειται για ένα φαινόμενο κατά το οποίο ενέργεια απο μια συχνότητα "διαρρέει" σε γειτονικές, γεγονός που καθιστά και πιο δύσκολη την ανακατασκευή του σήματος.

Βήμα 2 (Μετασχηματισμός DFT): Μετά την παραθυροποίηση, εφαρμόζεται ο Μετασχηματισμός DFT. Ο Διακριτός Μετασχηματισμός Fourier μετασχηματίζει μια ακολουθία N δειγμάτων $x_n = x_0, x_1, ..., x_{N-1}$, σε μία ακολουθία N δειγμάτων $X_k = X_0, X_1, ..., X_{N-1}$, σύμφωνα με τη σχέση:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N} \tag{2.4}$$

Τα παραπάνω μπορούν να εκφραστούν και σε πινακική μορφή ως εξής:

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ ... \\ X_{N-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & ... & 1 \\ 1 & \omega_n & \omega_n^2 & ... & \omega_n^{N-1} \\ 1 & \omega_n^2 & \omega_n^4 & ... & \omega_n^{2(N-1)} \\ ... & ... & ... & ... & ... \\ 1 & \omega_n^{N-1} & \omega_n^{2(N-1)} & ... & \omega_n^{(N-1)^2} \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ ... \\ x_{N-1} \end{bmatrix} \tag{2.5}$$

, όπου $w_n = e^{-i2\pi/N}$

Η πολυπλοκότητα του συγκεκριμένου αλγορίθμου είναι $O(N^2)$ αφού για κάθε ένα απο τα N δείγματα απαιτούνται N πολλαπλασιασμοί. Ωστόσο η επινόηση του αλγορίθμου FFT, μειώνει σημαντικά την πολυπλοκότητα υπολογισμου του DFT.

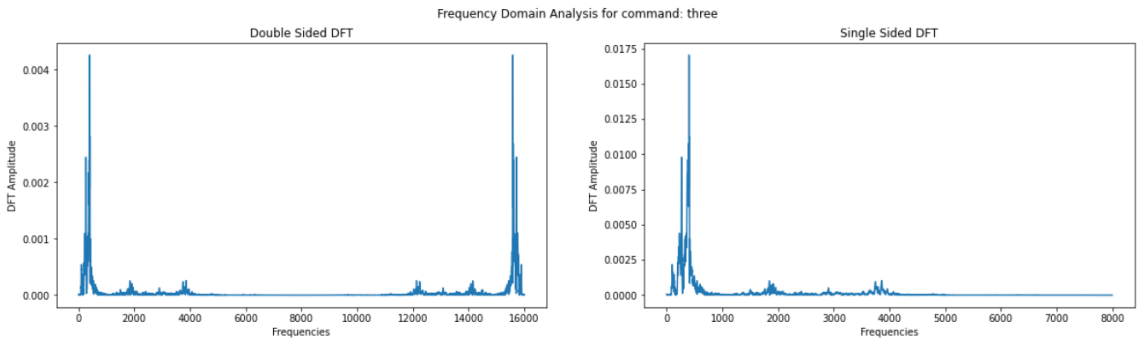
Βήμα 2.1 (Αλγόριθμος FFT για υπολογισμό DFT): Η κατασκευή των περισσοτέρων αλγορίθμων, βασίζεται στην αναζήτηση συμμετρίας. Στη προκειμένη περίπτωση αποδεικνύεται ότι:

$$
\begin{aligned}
X_{N+k} &= \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi(N+k)n/N} \\
&= \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N} \cdot e^{-i2\pi n} \\
&= \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N}
\end{aligned}
$$

Συμπερασματικά

$$X_{N+k} = X_k \tag{2.6}$$

Αρχικά πραγματοποείται zero padding ώστε το πλήθος των δειγμάτων (του αρχικού σήματος) να είναι δύναμη του 2. Με τον τρόπο αυτό το παράθυρο έχει μέγεθος $fftsize = 512, 1024, 2048...$. Στην συνέχεια, είναι δυνατή η αναδιάταξη των εξισώσεων (2.5), ώστε να υπολογίζονται ξεχωριστά οι άρτιοι και οι περιττοί όροι. Ορίζοντας λοιπόν δύο υποπροβλήματα, το κόστος μειώνεται σε $N \times N/2$. Η διαδικασία "διαίρει και βασίλευε" οδηγεί τελικά σε πολυπλοκότητα $O(NlogN)$. Με τον αλγόριθμο FFT επιτεύχθηκε λοιπόν, ο μετασχηματισμός DFT της ακολουθίας σε μια άλλη ακολουθία μιγαδικών αριθμών. Η γραφική παράσταση θα προκύψει προφανώς λαμβάνοντας το μέτρο αυτών των μιγαδικών ποσοτήτων. Όπως φαίνεται και παρακάτω (Σχήμα 2.7, αριστερά), ο DFT είναι συμμετρικός:



Σχήμα 2.7: Αμφίπλευρη-Μονόπλευρη αναπαράσταση Μ/Σ Fourier

Η συμμετρία έγκειται ως προς το μισό του ρυθμού δειγματοληψίας (συχνότητα Nyquist). Σύμφωνα με το θεώρημα δειγματοληψίας του Nyquist-Shannon, ένα σήμα μπορεί να ανακατασκευαστεί πλήρως αν περιλαμβάνει συχνότητες οι οποίες είναι μικρότερες απο το μισό της συχνότητας δειγματοληψίας. Με άλλα λόγια, η μέγιστη συχνότητα που επιστρέφει ο DFT, είναι μισή απο το ρυθμό δειγματοληψίας. Για το λόγο αυτό δεν χρειάζεται να λάβουμε υπόψιν το συμμετρικό κομμάτι (Σχήμα 2.7, δεξιά). Η ενέργεια του σήματος σε μια συχνότητα, εκφράζεται ισόποσα απο δύο συμμετρικές. Έτσι, στον μονόπλευρο DFT, όλες οι τιμές πλάτους πρέπει να διπλασιαστούν.

Επίσης κάθε τιμή διαιρείται με το μέγεθος του παραθύρου. Όπως φαίνεται και απο τις εξισώσεις (2.5), κάθε τιμή X έχει προκύψει απο το άθροισμα $fftsize$ συντελεστών. Διαιρώντας με το μέγεθος του παραθύρου επιτυγχάνουμε μια κανονικοποίηση, με την τιμή που προκύπτει να είναι πιο αντιπροσωπευτική για το παράθυρο Τέλος να σημειωθεί ότι λόγω DFT οι συχνότητες διακριτοποιούνται σε bins τα οποία συνολικά είναι $fftsize$ / 2.

2.3.2 Ανάλυση με Spectrogram

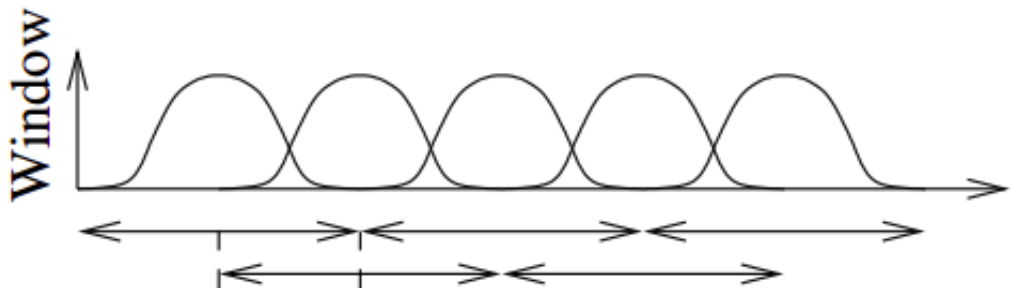
Προηγουμένως, έγινε αντιληπτό ότι η ανάλυση στο πεδίο του χρόνου αγνοεί το συχνοτικό περιεχόμενο, ενώ στο πεδίο της συχνότητας χάνεται κάθε έννοια χρονικής ακολουθίας. Τη λύση στο παραπάνω έρχεται να δώσει η χρήση φασματογραφήματος (spectrogram) που συνδυάζει τα δύο παραπάνω πεδία.

Πρόκειται για μια δισδιάστατη απεικόνιση μεταξύ χρόνου - συχνότητας. Πιο ειδικά, στον οριζόντιο άξονα είναι ο χρόνος, στον κατακόρυφο η συχνότητα, ενώ η κλιμακα των χρωμάτων αναφέρεται στο πλάτος των συχνοτήτων. Για παράδειγμα ένα τυχαίο σημείο πάνω στο φασματογράφημα υποδηλώνει μια συγκεκριμένη χρονική στιγμή (οριζόντιος άξονας), τι πλάτος έχουν (χρωματική κλίμακα) οι επιμέρους συχνότητες (κατακόρυφος άξονας). Για να γίνει κατανοητή αυτή η απεικόνιση, επεξηγείται μια βασική έννοια που θεμελιώνει την ύπαρξη του spectrogram.

Βήμα 3 (Short time Fourier Transform STFT): Ένα σήμα μεταβάλλεται συνεχώς και ακαθόριστα σε σχέση με το χρόνο. Θεωρώντας όμως ότι το σήμα, σε ένα σχετικά μικρό (framesize) παράθυρο, είναι στατιστικά σταθερό, λαμβάνεται το συχνοτικό περιεχόμενο μόνο για αυτό το χωρίο με τη χρήση DFT. Προφανώς για τους λόγους που εξηγήθηκαν και στην προηγούμενη παράγραφο κρίνεται αναγκαία η παραθυροποίηση. Στην συνέχεια το παράθυρο κυλίεται προς τα δεξιά, με τη διαδικασία να σταματά όταν ληφθεί υπόψιν όλο το σήμα. Στην παραπάνω διαδικασία να σημειωθούν τα ακόλουθα:

Αν το μέγεθος του παραθύρου (στο εξής framesize) είναι μικρότερο απο το μέγεθος fftsize πραγματοποιείται *zero padding* εως ότου εξισορροπηθούν οι τιμές.. Με την παράμετρο αυτή fftsize, καθορίζεται ο αριθμός των bins στα οποία θα διακριτοποιηθούν οι συχνότητες.

Τα παράθυρα είναι εν γένει επικαλυπτόμενα, καθώς εφαρμόζοντας παραθυροποίηση στα επιμέρους τμήματα, χάνεται πληροφορία που εμπεριέχεται στα άκρα των παραθύρων αυτών.



Σχήμα 2.8: Επικάλυψη παραθύρων κατα την εφαρμογή FFT

Ενδιαφέρον παρουσιάζει η σύγκριση DFT-STFT:

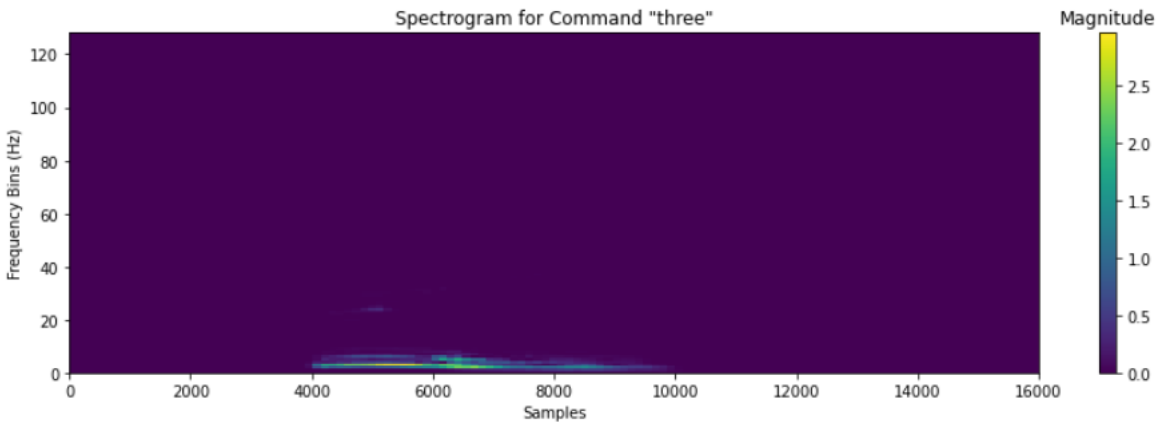
DFT:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N} \tag{2.7}$$

STFT

$$S[m,k] = \sum_{n=0}^{N-1} x(n+mH) \cdot w(n) \cdot e^{-i2\pi kn/N} \tag{2.8}$$

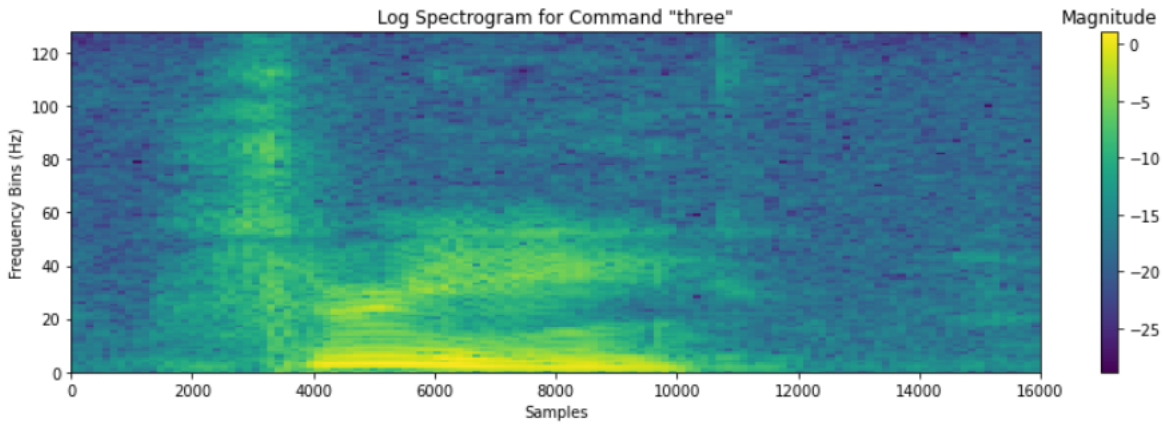
όπου m ο αριθμός του frame, k η συχνότητα και H το hop size. Πρόκειται λοιπόν για έναν 2D πίνακα μεγέθους frames x frequency bins που περιλαμβάνει μιγαδικούς αριθμούς (στη μορφή πλάτος-φάση). Τα παραπάνω μπορούν να υπολογιστούν από τις σχέσεις: frequency bins = framesize/2 +1 και frames = (samples - framesize)/hopsize +1.



Σχήμα 2.9: Εξαγωγή spectrogram για την εντολή *three*

Με τον αλγόριθμο STFT ο ήχος μπορεί να οπτικοποιηθεί χρησιμοποιώντας τη σχέση :

$$Y = 1/N \cdot |S[m,k]|^2, N = fftsize \tag{2.9}$$



Σχήμα 2.10: Εφαρμογή συνάρτησης λογαρίθμου στο φασματογράφημα της εντολής *three*

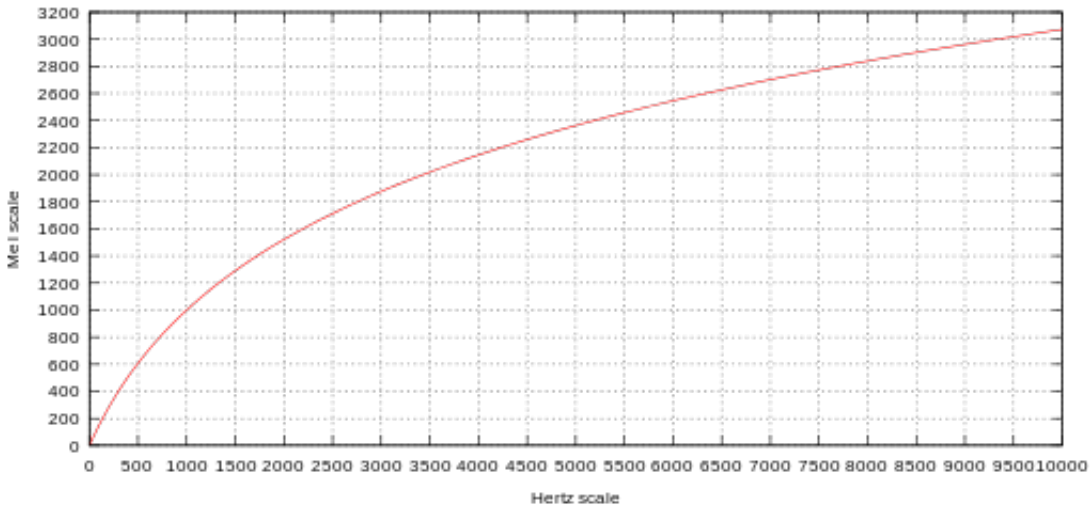
Παρατηρεί κανείς ότι η παραπάνω γραφική απεικόνιση (Σχήμα 2.9), περιέχει πληροφορία μόνο σε ένα συγκεκριμένο τμήμα. Οι περισσότερες τιμές είναι κοντά στο μηδέν, και για το λόγο αυτό το μεγαλύτερο μέρος καλύπτεται από μωβ χρώμα. Για το λόγο αυτό χρειάζεται μια συνάρτηση που θα αναδείξει τις τιμές αυτές.

Βήμα 4 (Μετατροπή πλάτους σε λογαριθμική κλίμακα): Με τη συγκεκριμένη μετατροπή, το πλατος πλεον εκφράζεται λογαριθμικά στην κλιμακα των χρωμάτων. Όπως σημειώθηκε και παραπάνω κάτι τέτοιο ήταν αναγκαίο για να ανακτηθεί πληροφορία και από άλλα τμήματα του spectrogram. Το παραπάνω spectrogram (Σχήμα 2.10), αν και ικανό να αναδείξει ομοιότητες - διαφορές μεταξύ των σημάτων, ωστόσο δεν λαμβάνει υπόψιν τον τρόπο με τον οποίο ο άνθρωπος αντιλαμβάνεται τα ακουστικά ερεθίσματα.

2.3.3 Ανάλυση με Mel Spectrograms

Εν γένει, ο τρόπος με τον οποίο ο άνθρωπος αντιλαμβάνεται τα ακουστικά ερεθίσματα είναι μια πολύπλοκη διαδικασία. Πειράματα έχουν αποδείξει ότι ο άνθρωπος δεν αναγνωρίζει με την ίδια ευκολία όλες τις συχνότητες. Για παράδειγμα στο ακουστό φάσμα του ανθρώπου ($20Hz$ - $20kHz$) είναι εύκολο να εντοπιστούν διαφορές μεταξύ σημάτων $800Hz$, $900Hz$, $1000Hz$. Κάτι τέτοιο δεν συμβαίνει με μεγαλύτερες συχνότητες όπου είναι αδύνατον το ανθρώπινο αυτί να ξεχωρίσει ήχους που ανήκουν για παράδειγμα στις συχνότητες $7000Hz, 7100Hz, 7200Hz$.

Εξαιτίας αυτού του γεγονότος, οι Stevens, Volkmann, και Newman, μετά απο ψυχοακουστικά πειράματα, δημιούργησαν μια κλίμακα η οποία πήρε την ονομασία mel-scale. Η γραφική της απεικόνιση φαίνεται παρακάτω:



Σχήμα 2.11: Μετατροπή κλίμακας Hertz στην κλίμακα mel

Σύμφωνα με την παραπάνω γραφική, ο άνθρωπος αντιλαμβάνεται λογαριθμικά τις επιμέρους συχνότητες, και όχι γραμμικά. Όπως γίνεται αντιληπτό, για μικρότερες συχνότητες η διαφορά των τιμών στον κατακόρυφο άξονα είναι μεγαλύτερη. Απο την άλλη, μεγαλύτερες συχνότητες λαμβάνουν σχεδόν ίδιες τιμές καθώς η καμπύλη τείνει να γίνεται ευθεία. Πλέον λοιπόν, μέσω της συνάρτησης

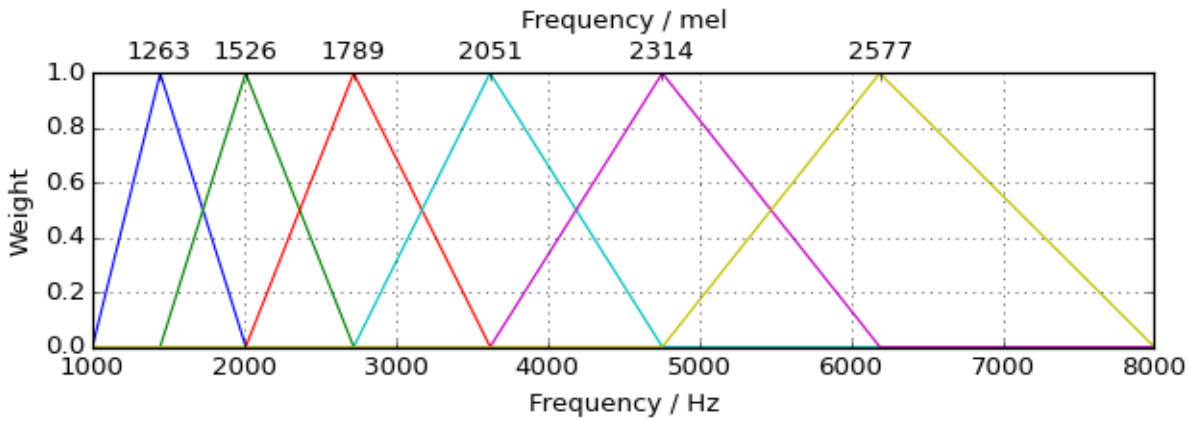
$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{10}\right) \tag{2.10}$$

$$f = 700(10^{m/2595} - 1) \tag{2.11}$$

πραγματοποιείται μια αντιστοίχιση τιμών ενός συνόλου (συχνότητες σε Hz) σε ένα άλλο σύνολο τιμών (συχνότητες σε mel scale). Στο σύνολο αυτό, αυτό που επιτυγχάνεται είναι: η διαφορά μεταξύ ακουστικών σημάτων 500 - 510 mels είναι τόσο αισθητή όσο και σε 1000 - 1010 mels.

Βήμα 5 (Μετατροπή συχνοτήτων σε κλίμακα mel). Στα παραπάνω 3 βήματα για εξαγωγή spectrograms, προστίθεται ακόμα ένα βήμα για την εξαγωγή mel spectrograms, που δεν είναι άλλο απο τη μετατροπή των συχνοτήτων στην κλιμακα mel. Τα βήματα περιγράφονται αναλυτικά ως ακολούθως:

1. Επιλογή αριθμού mel bands: Συνήθως λαμβάνει τιμές μεταξύ του διαστήματος $40-120$. Πρόκειται για μια υπερπαράμετρο, η οποία διαφέρει ανάλογα με το πρόβλημα που εξετάζεται. Η βέλτιστη τιμή μπορεί να βρεθεί με πειραματισμό και η σημασία της επεξηγείται παρακάτω:



Σχήμα 2.12: Κατασκευή τριγωνικών φίλτρων για την εξαγωγή mfcc χαρακτηριστικών

2. Κατασκευή mel filterbanks ως εξής:

   (α′) Μετατροπή επιθυμητής ελάχιστης (συνήθως 1000 Hz) και μέγιστης (συνήθως $samplerate/2$) συχνότητας στη κλίμακα *mel* συμφωνα με τη σχέση (2.10).

   (β′) Χωρισμος του διαστήματος που ορίζουν τα δύο παραπάνω σημεία σε $melbands + 1$ χωρία.

   (γ′) Κατασκευή τριγώνων, με βάση των κανόνα ότι ένα τρίγωνο ξεκινά απο την μέση του προηγούμενου. Κάθε τρίγωνο έχει $frequency\ bins$ σημεία (όπως έχει οριστεί απο STFT)

   (δ′) Μετατροπή των σημειών αυτών πάλι στην κλίμακα Hz σύμφωνα με τη σχέση 2.11

Στο παραπάνω παράδειγμα έχουμε 6 *mel bands*. Όπως φαίνεται, τα κέντρα των τριγωνικών φίλτρων έχουν την ίδια απόσταση (σε κλίμακα *mel*). Επίσης τα άκρα ενος τριγωνικού φίλτρου (αριστερό - δεξί), είναι αντίστοιχα η κορυφή του προηγούμενου και του επόμενου τριγώνου. Για το λόγο αυτό όσο κινούμαστε προς τα δεξιά τόσο "απλώνει" το τριγωνικό φίλτρο. Εν κατακλείδι, το μέγεθος του πίνακα που περιγράφει τα τριγωνικά φίλτρα είναι $M = (bands, frequency\ bins)$.

3. Εφαρμογή mel filterbanks σε spectrogram: Προηγουμένως είχε υπολογιστεί ότι το μέγεθος ενός spectrogram είναι: $Y = (frames, frequency\ bins)$. Η εφαρμογή mel spectrogram, πλεον ανάγεται στον πολλαπλασιασμό δύο πινακων: Mel Spectrogram = $log(MY^T)$, και έτσι προκύπτει ένας πίνακας μεγέθους $(frames, bands)$.

2.3.4 Ανάλυση με MFCC

Στην αναγνώριση ομιλίας, και ειδικά στην εξαγωγή χαρακτηριστικών, χρησιμοποιούνται μεταξύ άλλων και οι Mel Frequency Cepstral Coefficients ή συντομογραφικά MFCCs. Εμπνευστές της ιδέας αυτής, υπήρξαν οι Davis and Mermelstein το 1980. Η εξήγηση του όρου mel frequency έχει δοθεί στην προηγούμενη παράγραφο. Ο όρος cepstral προέρχεται απο τον cepstrum (αναγραμματισμός σπεστρυμ), και δίνεται απο τη σχέση:

$$C(x(t)) = F^{-1}[\log(F(x(t)))] \tag{2.12}$$

Βήμα 6 (Υπολογισμός DCT). Τόσο οι επικαλύψεις μεταξύ των παραθύρων, όσο και οι επικαλυψεις στα επιμέρους filterbanks (τριγωνικά φίλτρα), δημιουργούν συσχετίσεις μεταξύ των τελικών features. Κάτι τέτοιο δημιουργεί προβλήματα στην διαδικασία της μάθησης. Τη λύση σε αυτό έρχεται να δώσει ο μετασχηματισμός DCT, ο οποίος εν τέλει δίνει και τους mfcc συντελεστές. Η σχέση του DCT για 1-D δεδομένα (frames) είναι:

$$F(u) = \left(\frac{2}{N}\right)^{\frac{1}{2}} \sum_{i=0}^{N-1} \Lambda(i) \cos(\frac{\pi u}{2N}(2i+1)) f(i) \tag{2.13}$$

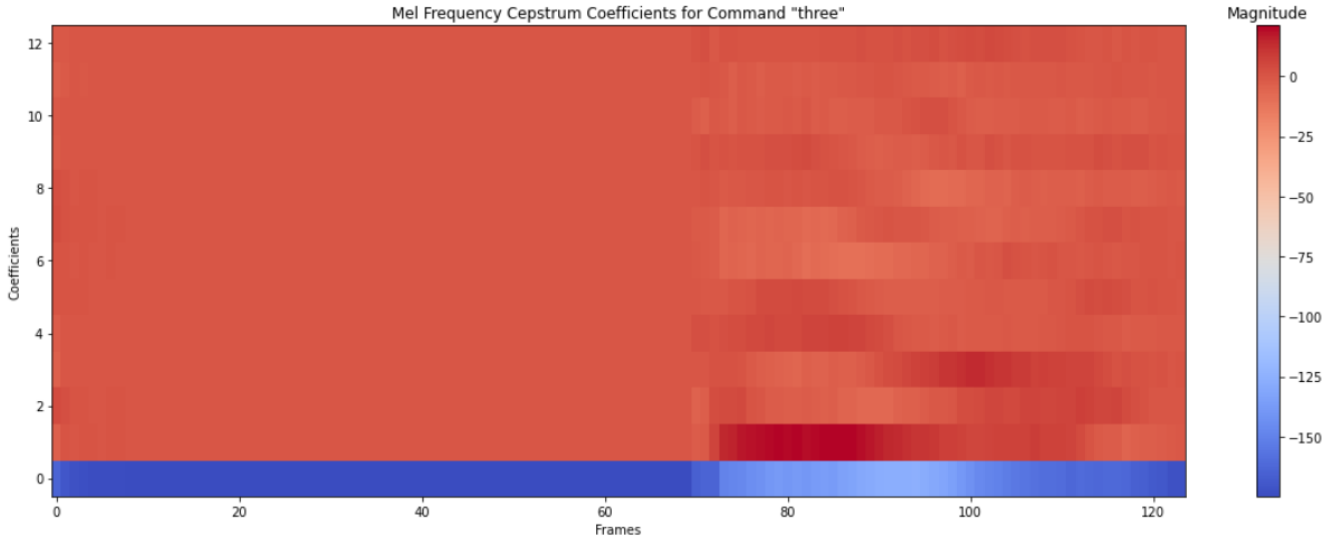
Συνήθως μας ενδιαφέρουν οι πρώτοι 12 -13 συντελεστές mfcc για κάθε frame οι οποίοι περιέχουν πληροφορία για φωνήματα, περιβάλλουσα φάσματος, ενώ οι μεγαλύτεροι συντελεστές δεν περιέχουν ουσιαστική πληροφορία.

2.3.5 Ανάλυση με delta - MFCC

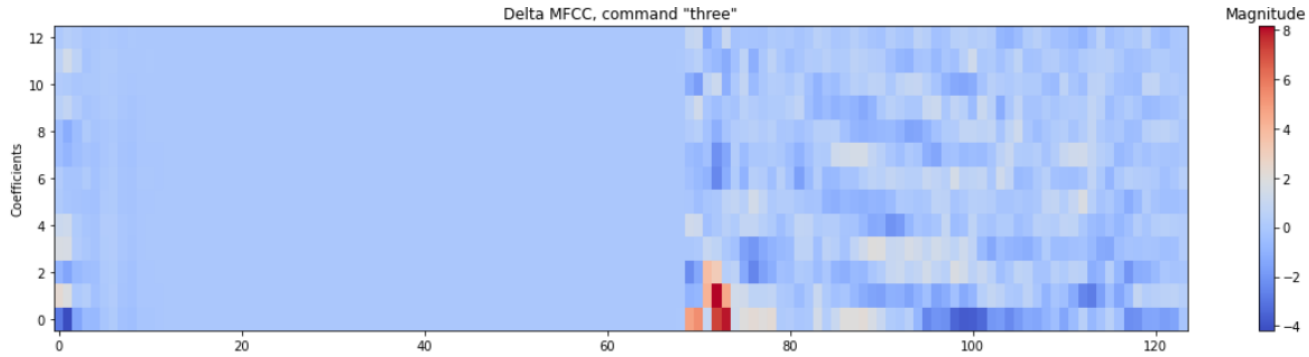
Ενδιαφέρον παρουσιάζουν και οι συντελεστές deltas-deltas που αναφέρονται στην πρώτη και δεύτερη παράγωγο των mfcc's αντίστοιχα.

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2.14}$$

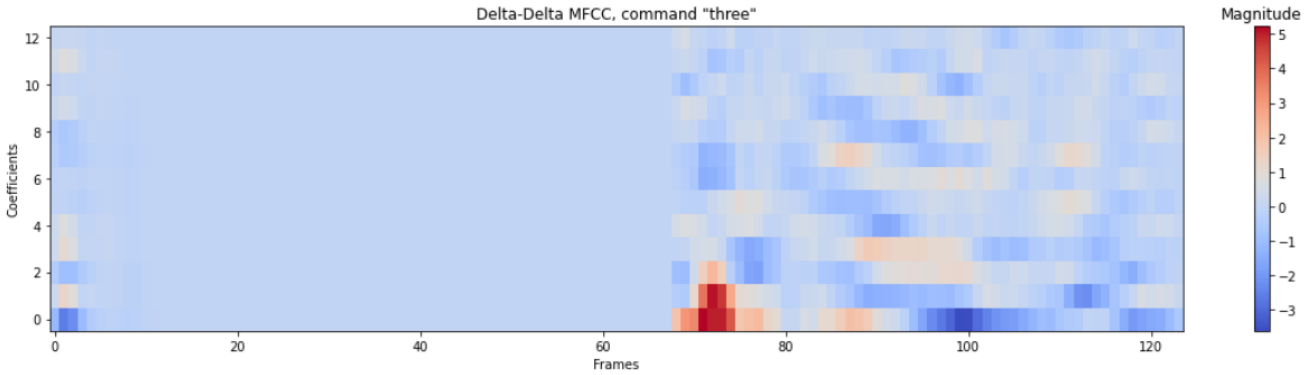
όπου N ο αριθμός προηγούμενων - επόμων features που θα ληφθούν υπόψιν, t ο αριθμός του frame n ο αριθμός του n- επόμενου ή προηγούμενου frame και c τα features.



Σχήμα 2.13: Εξαγωγή MFCC χαρακτηριστικών για την εντολή *three*



Σχήμα 2.14: Εξαγωγή delta MFCC χαρακτηριστικών για την εντολή *three*



Σχήμα 2.15: Εξαγωγή delta-delta MFCC χαρακτηριστικών για την εντολή *three*

Κεφάλαιο 3

Γενικά περί Μηχανικής Μάθησης

3.1 Γενικά περί Μηχανικής Μάθησης

Η Μηχανική Μάθηση, ένας υποκλάδος της Τεχνητής Νοημοσύνης και γενικότερα της επιστήμης των υπολογιστών, αναφέρεται στον σχεδιασμό και στην ανάπτυξη αλγορίθμων που δέχονται ως είσοδο εμπειρικά δεδομένα, αναζητούν σε αυτά πρότυπα χρησιμοποιώντας στατιστικές μεθόδους, και επιτελούν προβλέψεις για τους εμπλεκόμενους μηχανισμούς που δημιούργησαν τα δεδομένα αυτά. Εναλλακτικά μπορεί να οριστεί ως το επιστημονικό πεδίο που προσδίδει στους υπολογιστές την ικανότητα να μάθουν, χωρίς να έχουν άμεσα προγραμματιστεί για το σκοπό αυτό (Samuel 1959). Ένας πλέον τυπικός ορισμός σχετικός με την Εκμάθηση Μηχανών αναφέρει ότι 'προγράμματα υπολογιστών μπορούν να μάθουν την εμπειρία E, όσον αφορά κάποιες διεργασίες T και μετρικές απόδοσης P, αν οι αποδόσεις τους στα έργα T όπως μετρήθηκαν με P, βελτιώνονται με την εμπειρία E (E. Mitchell 1997). Εν γένει, ο τομέας της Μηχανικής Μάθησης αναπτύσσει τρεις τρόπους μάθησης ανάλογους με τους τρόπους με τους οποίους μαθαίνει ο άνθρωπος: επιβλεπόμενη μάθηση, μη επιβλεπόμενη μάθηση και ενισχυτική μάθηση.

Επιβλεπόμενη Μάθηση (Supervised Learning): είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε επιθυμητές εξόδους με απώτερο σκοπό τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα ταξινόμησης (classification), πρόγνωσης (prediction) διερμηνείας (interpretation) και παλινδρόμησης (regression)

Μη επιβλεπόμενη Μάθηση (Unsupervised Learning): είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπο μορφή παρατηρήσεων, χωρίς να γνωρίζει τις επιθυμητές εξοδους. Χρησιμοποιείται σε προβλήματα ανάλυσης συσχετισμών (association analysis) και ομαδοποίησης (clustering).

Ενισχυτική Μάθηση (Reinforcement Learning): είναι η διαδικασία όπου ο αλγόριθμος, μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.

3.2 Αλγόριθμοι Μηχανικής Μάθησης

3.2.1 Επιβλεπόμενη Μάθηση

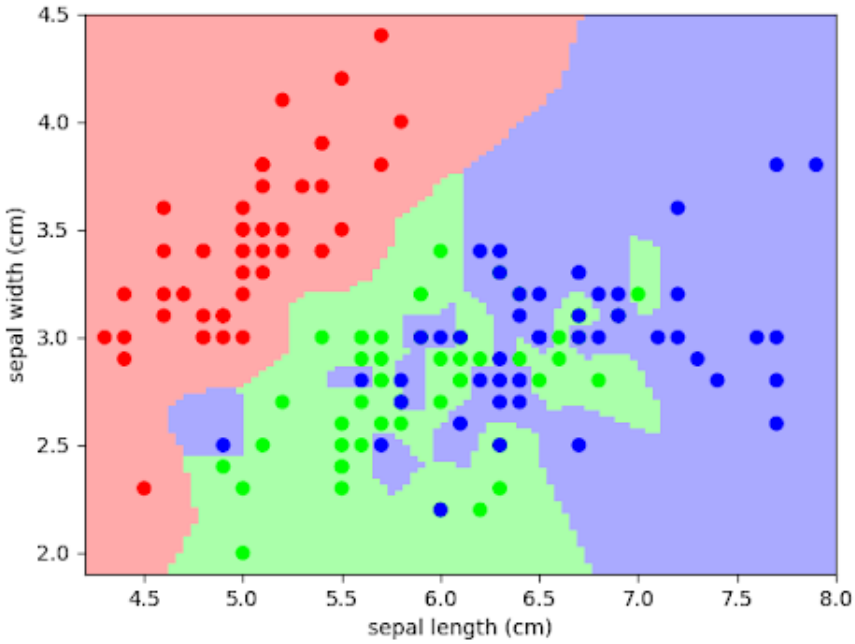
kNN

Πρόκειται για έναν αλγόριθμο επιβλεπόμενης μηχανικής μάθησης, ο οποίος υποθέτει ότι παρόμοια πρότυπα βρίσκονται στην ίδια "γειτονιά". Το πλήθος των γειτόνων καθορίζεται μέσω μιας υπερπαραμέτρου k. Η αρχή λειτουργίας του είναι απλή: Έστω ότι τα χαρακτηριστικά ανήκουν στον n-διάστατο χώρο, δηλαδή είναι διανύσματα της μορφής $x = [x_1, x_2, ..., x_n]$. Για κάθε νέο δείγμα x' (που δεν ανήκει στο σύνολο εκπαίδευσης), υπολογίζονται όλες οι αποστάσεις με τα δείγματα εκπαίδευσης, με βάση κάποια συνάρτηση απόστασης, συνήθως Ευκλείδεια:

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + ... + (x_n - x'_n)^2} \tag{3.1}$$

Στην συνέχεια, επιλέγονται οι k μικρότερες αποστάσεις για να προσδιοριστούν οι k γείτονες του νέου δείγματος. Με βάση λοιπόν την *πλειοψηφία* των ετικετών των γειτόνων, προσδιορίζεται και η ετικέτα (κλάση) του νέου δείγματος.

Στο παρακάτω παράδειγμα, επιθυμούμε να ταξινομήσουμε άνθη σε τρεις κατηγορίες με βάση κάποια χαρακτηριστικά που διαθέτουν. Τα χαρακτηριστικά είναι στον 2D χώρο, δηλαδή είναι της μορφής $x = [x_1, x_2]$, όπου το x_1 είναι το χαρακτηριστικό sepal width (πλάτος φύλλου ανθού) και x_2 το χαρακτηριστικό sepal length (μήκος φύλλου ανθού).



Σχήμα 3.1: Περιοχές απόφασης με βάση τον αλγόριθμο knn

Οι κουκίδες αναφέρονται στα δείγματα του συνόλου εκπαίδευσης, ενώ το χρώμα τους

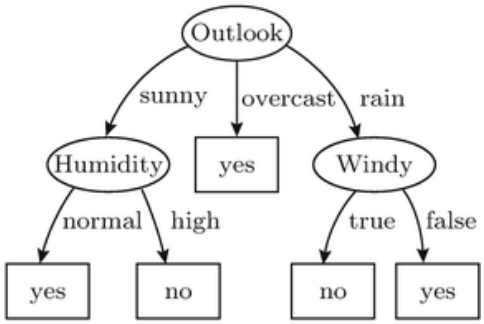
υποδηλώνει την κλάση στην οποία ανήκουν. Με αντίστοιχο χρώμα έχουν επισημειωθεί και οι περιοχές απόφασης. Η περιοχή απόφασης για μια κατηγορία A ορίζεται η περιοχή του n-διάστατου χωρου χαρακτηριστικών, στην οποία, αν βρεθεί ένα νέο δείγμα θα ταξινομηθει στην κατηγορία A, Για παράδειγμα, ένα νέο δείγμα με χαρακτηριστικά sepal length $= 4.5\ cm$ και sepal width $= 4\ cm$ δηλαδή το $x = [4.5, 4]$ θα ταξινομηθει στην κατηγορία με κόκκινο χρώμα, καθώς ο ένας ($k = 1$) κοντινότερος του γείτονας ανήκει σε αυτή την κατηγορία (έχει σημειωθεί με κόκκινο χρώμα).

Δέντρα απόφασης

Ένα δέντρο απόφασης, μπορεί να προβλέψει την κλάση που ανήκει ένα δείγμα, διενεργώντας μια σειρά ερωτήσεων στα χαρακτηριστικά του. Πρόκειται για ένα διάγραμμα ροής (flowchart) το οποίο έχει ανάποδη δενδρική δομή. Πιο ειδικά, σε κάθε εσωτερικό κόμβο πραγματοποιείται ένας έλεγχος (ερώτηση) για κάποιο χαρακτηριστικό. Κάθε κλαδί αυτού του κόμβου αναπαριστά και διαφορετική απάντηση (έξοδο) απο τον έλεγχο που πραγματοποιήθηκε. Τα φύλλα αυτού του δέντρου, αντιπροσωπεύουν τις ετικέτες των κλάσεων (labels).

Στο παρακάτω παράδειγμα, έχουμε ένα πρόβλημα δυαδικής ταξινόμησης (yes / now) στο ερώτημα: *Play Golf*. Για την κατασκευή του δέντρου απόφασης, χρησιμοποιήθηκαν 14 δείγματα εκπαίδευσης, κάθε ένα απο τα οποία διέθετε 4 χαρακτηριστικά (Outlook, Temp, Humidity, Windy).

Outlook	*Temp*	*Humidity*	*Windy*	*Golf?*
rainy	hot	high	false	no
rainy	hot	high	true	no
overcast	hot	high	false	yes
sunny	mild	high	false	yes
sunny	cool	normal	false	yes
sunny	cool	normal	true	no
overcast	cool	normal	true	yes
rainy	mild	high	false	no
rainy	cool	normal	false	yes
sunny	mild	normal	false	yes
rainy	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
sunny	mild	high	true	no



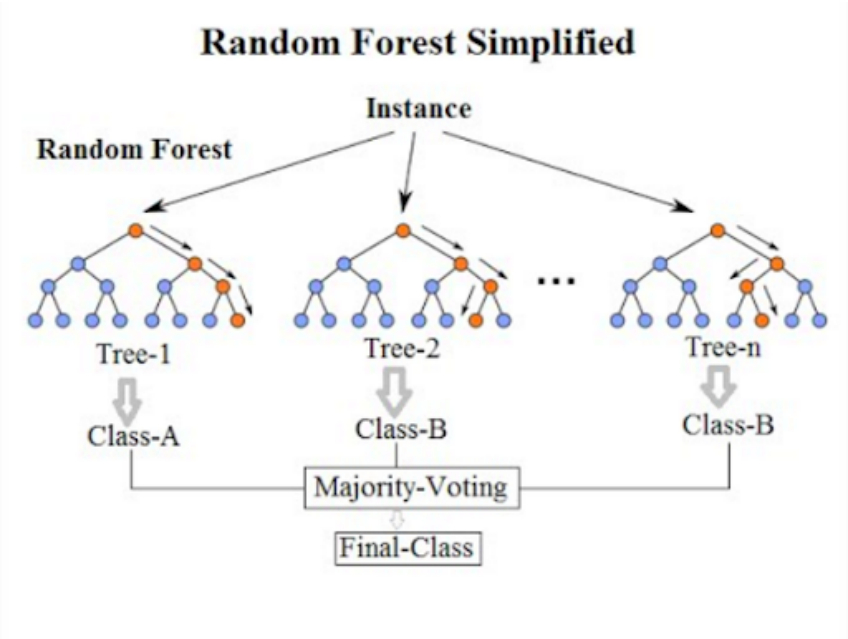
Σχήμα 3.2: Κατασκευή Δέντρου Απόφασης με βάση τα στοιχεία του πίνακα

Τα δέντρα απόφασης συνιστούν ένα ιδιαίτερα δημοφιλές εργαλείο, λόγω της ταχύτητας τους και της ευκολίας που παρέχουν τόσο σε επίπεδο κατανόησης, όσο και υλοποίησης. Ωστόσο, όπως αναφέρεται "Τα δέντρα απόφασης δεν μπορούν να γίνουν το καλύτερο εργαλειο ταξινόμησης, λόγω ανακρίβειας" [The elements of statistical learning [2]]. Με άλλα λόγια, σε κατάλληλα σύνολα δεδομένων (όπου δεν υπάρχουν δείγματα με πανομοιότυπα χαρακτηριστικά και διαφορετικές ετικέτες), είναι πάντα πιθανόν να κατασκευαστούν δενδρικές δομές με μηδενικό σφάλμα ταξινόμησης, στο σύνολο εκπαίδευσης. Ωστόσο κάτι τέτοιο εγείρει τον κίνδυνο overfit, δηλαδή την αδυναμία του δέντρου να ανταποκριθεί σε δειγματα που δεν έχουν

χρησιμοποιηθεί κατά την εκπαίδευση και να γενικευσει κατάλληλα. Η ιδέα ωστόσο ότι, για ένα σύνολο δεδομένων μπορούν να κατασκευαστούν διαφορετικά δέντρα απόφασης (δίνοντας το κάθε ένα διαφορετική βαρύτητα σε ορισμένα χαρακτηριστικά), οδήγησε στην μετάβαση απο τα δέντρα απόφασης στα δάση από δέντρα (Random Forest).

Τυχαία Δάση - Random Forest

Μία συνηθισμένη τεχνική στην Επιβλεπόμενη Μάθηση είναι η εκπαίδευση πολλαπλών ταξινομητών, ανεξάρτητα μεταξύ τους, συνδυάζοντας τα αποτελέσματά τους στο τέλος για την πρόβλεψη της τελικής τιμής (Ensemble Models). Μια μέθοδος, για τη δημιουργία ensemble μοντέλων είναι η Bootstrap AGGregating (Συντομογραφικά BAGG). Πιο αναλυτικα, όταν δίνεται ένα σύνολο δεδομένων, εξάγονται bootstrapped συνολα. Ένα τέτοιο σύνολο δημιουργείται επιλέγοντας τυχαία δείγματα απο το αρχικό dataset, χωρίς να αποκλείεται να επιλεγεί το ίδιο δείγμα, περισσότερες απο μια φορές. Στην συνέχεια κατασκευάζεται ένα δέντρο απόφασης το οποίο όμως περιλαμβάνει ένα υποσύνολο των αρχικών χαρακτηριστικών στους κόμβους του. Η διαδικασία αυτή (δημιουργία bootstrapped συνόλου - εκπαίδευση δέντρου με ορισμένα χαρακτηριστικά), επαναλαμβάνεται ως ότου συμπληρωθεί ο επιθυμητός αριθμός δέντρων. Κάθε ένα απο τα διαφορετικά δέντρα, δίνει μια πρόβλεψη για την κλάση στην οποία ανήκει το δείγμα, η οποία μπορει να είναι διαφορετική για κάθε δέντρο απόφασης. Οι προβλέψεις αυτές συλλέγονται (aggregated), και η τελική απόφαση προκύπτει με βάση την πλειοψηφία των προβλέψεων, των επιμέρους δέντρων απόφασης. Η διαδικασία αυτή περιγράφεται σχηματικά ακολούθως:



Σχήμα 3.3: Κατασκευή Τυχαίου Δάσους χρησιμοποιώντας Δέντρα Απόφασης

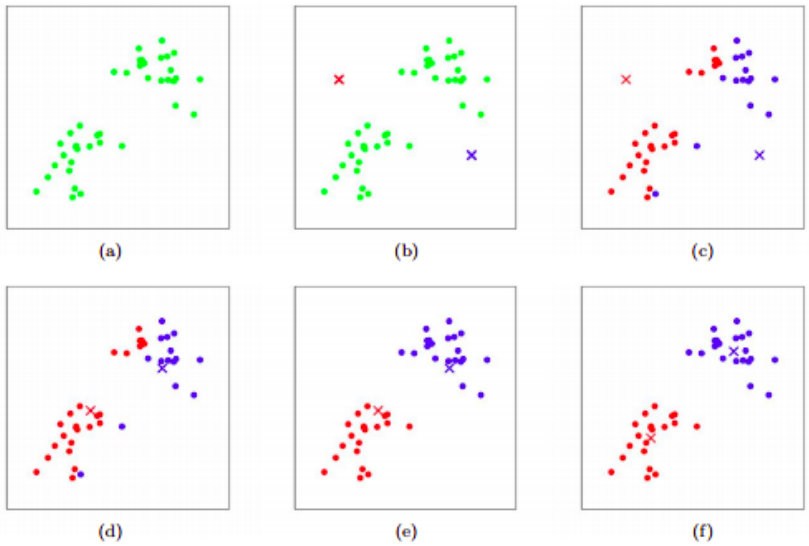
3.2.2 Μη επιβλεπόμενη Μάθηση

k-means

Ένας δημοφιλής αλγόριθμος μη επιβλεπόμενης μάθησης είναι ο k-means. Τυπικά, οι αλγόριθμοι μη επιβλεπόμενης μάθησης αναλύουν τα δείγματα σε συστάδες, χωρίς να είναι γνωστές οι ετικέτες τους. Μια συστάδα μπορεί να θεωρηθεί σαν ένα σύνολο δειγμάτων που ομαδοποιούνται, λόγω των κοινών χαρακτηριστικών τους. Βασική παράμετρο του συγκεκριμένου αλγορίθμου αποτελεί ο αριθμός των clusters k. Κάθε cluster εκπροσωπείται απο ένα κεντροειδές, ένα σημείο δηλαδή που αποτελεί το κέντρο μια ομάδας. Η λειτουργία του συγκεκριμένου αλγορίθμου είναι απλή:

Στο χώρο των δειγμάτων τοποθετούνται τυχαία k κεντροειδή, καθε ένα απο τα οποία αντιπροσωπεύει και ένα cluster. Στην συνέχεια για κάθε δείγμα υπολογίζεται η απόστασή του (συνήθως Ευκλείδεια) απο κάθε κεντροειδές. Έτσι με βάση την μικρότερη απόσταση, το δείγμα εκχωρείται και στο cluster που εκπροσωπείται απο το κεντροειδές. Πλεον λοιπόν έχουν δημιουργηθεί k clusters. Έπειτα υπολογίζονται ξανά τα κέντρα αυτών των clusters και η παραπάνω διαδικασία συνεχίζεται με τα νεα κεντροειδή.

Ο αλγόριθμος τερματίζει όταν τα κεντροειδή δεν αλλάζουν σε κάθε βήμα του αλγορίθμου, ή επιτεύχθηκε ο επιθυμητός αριθμός επαναλήψεων του αλγορίθμου. Είναι σημαντικο να τονιστούν τα εξής: Εξάρχής είναι άγνωστος ο αριθμός των συστάδων στις οποοίες θα διαχωριστούν τα δέιγματα, και δέυτερον μια «κακή» αρχικοποίηση των κεντροειδών, μπορεί να οδηγήσει σε κακής ποιότητας συστάδες στην πορεία.



Σχήμα 3.4: Διαδοχικές (a) - (f) επαναλήψεις του αλγορίθμου k-means.

3.3 Ανάλυση Χρονοσειρών

3.3.1 Ορισμός

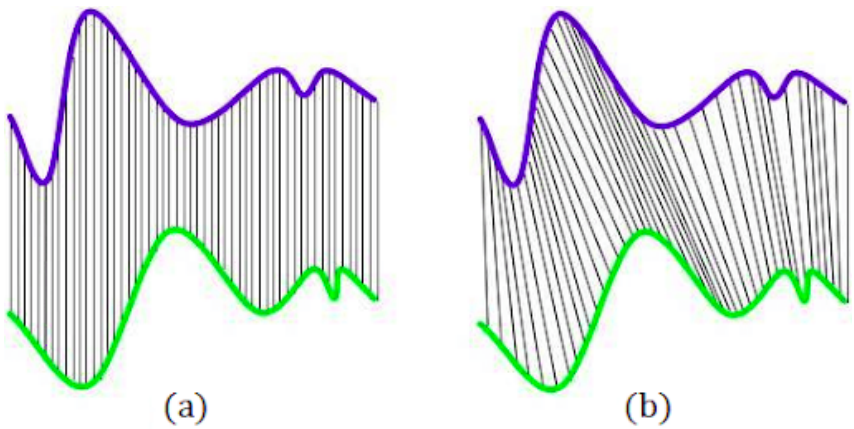
Μια χρονοσειρά περιγράφει την μεταβολή μιας ποσότητας, με τη πάροδο του χρόνου. Ένα ακουστικό σήμα, δεν είναι τίποτα άλλο παρά δείγματα, διατεταγμένα σε χρονική σειρά, που αναφέρονται στη μεταβολή του πλάτους. Μια πρώτη προσέγγιση θα ήταν η διαχείριση του κάθε σημείου σαν χαρακτηριστικό και η απευθείας εφαρμογή κάποιου αλγορίθμου μηχανικής μάθησης στα χαρακτηριστικά αυτά. Στη περίπτωση αυτή όμως, αγνοείται κάθε έννοια χρονικής ακολουθίας με αποτέλεσμα, αν τα δείγματα αναδιαταχτούν τυχαία, οι προβλέψεις να μην αλλάζουν. Για το λόγο αυτό χρησιμοποιούνται αλγόριθμοι που λαμβάνουν υπόψιν την χρονική ακολουθία.

3.3.2 Αλγόριθμοι

Οι επιμέρους αλγόριθμοι που εξετάζονται, μπορούν να χωριστούν σε τρεις κατηγορίες. Αλγόριθμοι βασισμένοι σε (a) Απόσταση, (b) Διαστήματα, (c) Συχνότητα.

Αλγόριθμοι Βασισμένοι στην Απόσταση

Ένας αλγόριθμος που μπορεί να χρησιμοποιηθεί στην περίπτωση αυτή είναι ο kNN, ο οποίος έχει εξηγηθεί στην Υποπαράγραφο 3.2.1. Στην περίπτωση των χρονοσειρών, βέβαια αντί της Ευκλείδειας Απόστασης μπορεί να χρησιμοποιηθεί η μετρική DTW (Dynamic Time Wrapping).



Σχήμα 3.5: Υπολογισμός (a) Ευκλείδειας Απόστασης σε σχέση με (b) DTW μεταξύ των δειγμάτων δύο χρονοσειρών.

Στην (a) περίπτωση, μέσω της Ευκλείδειας Απόστασης οι αποστάσεις που λαμβάνονται (σε γραμμικό χρόνο) μεταξύ των δειγμάτων θα οδηγήσουν στην λανθασμένη πρόβλεψη ότι οι δύο χρονοσειρές δεν σχετίζονται. Αντιθέτως στην (b) περίπτωση, η μετρική DTW θα εκτιμήσει ορθώς (σε τετραγωνικό χρόνο) ότι η μια χρονοσειρά είναι μετατοπισμένη έκδοση της άλλης .

Αλγόριθμοι Βασισμένοι σε Διαστήματα

Ένας αλγόριθμος που συσχετίζει χρονοσειρές, λαμβάνοντας τα επιμέρους διαστήματα είναι ο Random Forest που εξηγήθηκε στην προηγούμενη Υποπαράγραφο 3.2.1. Πιο ειδικά κάθε χρονοσειρά διαμερίζεται σε τυχαία διαστήματα, τα οποία έχουν εν γένει διαφορετικό μήκος. Στην συνεχεια, για κάθε ένα απο αυτά τα διαστήματα εξάγονται χαρακτηριστικά όπως μέση τιμή ή διασπορά. Τέλος εκπαιδεύεται ένα δέντρο απόφασης στα χαρακτηριστικά αυτά. Η διαδικασία αυτή (διαμερισμος χρονοσειράς, εξαγωγή χαρακτηριστικών, εκπαίδευση δέντρου απόφασης) επαναλαμβάνεται εως ότου κατασκευαστεί ο επιθυμητός αριθμός δέντρων. Αφού κατασκευαστούν τα δέντρα, κάθε νέα χρονοσειρά ταξινομείται με βάση την πλειοψηφική ψήφο των δέντρων που απαρτίζουν το δάσος.

Αλγόριθμοι Βασισμένοι στη Συχνότητα

Μια παραλλαγή του Time Series Forest, που εξηγήθηκε παραπάνω, είναι ο αλγόριθμος ομαδοποίησης Random Interval Spectral Ensemble (συντ. RISE). Πιο ειδικά, για κάθε δέντρο δεν χρησιμοποιούνται χαρακτηριστικά που αναφέρονται στο πεδίο του χρόνου (μέση τιμή - διασπορά), αλλά στο πεδίο της συχνότητας. Τα χαρακτηριστικά αυτά μπορεί να είναι:

Συντελεστές Αυτοσυσχέτισης: Οι συντελεστές αυτοσυσχέτισης r_k, χρησιμοποιούνται για να αναδείξουν τη σχέση που έχουν ορισμένες τιμές της χρονοσειράς, που απέχουν ένα ορισμένο διάστημα k. Πιο ειδικά, μέσω του συντελεστή r_1, μελετούμε την απόσταση όλων των δειγμάτων y_t, y_{t-1}, μέσω του r_2 τη σχέση y_t και y_{t-2}, κτλ χρησιμποιοώντας και τη μέση τιμή $\overline{y}$ των δειγμάτων. Γενικά υπολογίζονται βάσει του τύπου:

$$r_k = \sum_{t=k+1}^{T} \frac{(y_t - \overline{y})(y_{t-k} - \overline{y})}{\sum_{t=1}^{T}(y_t - \overline{y})^2} \tag{3.2}$$

Συντελεστές φάσματος ισχύος. Πρόκειται για συντελεστές που προκύπτουν από την εφαρμογή μετασχηματισμού Fourier, όπως εξηγήθηκε στην Υποπαράγραφο 2.3.1.

Τα παραπάνω χαρακτηριστικά μπορούν να συγχωνευτούν προκειμένου να εκπαιδευτούν δέντρα απόφασης. Πιο αναλυτικά ο αλγόριθμος RISE περιλαμβάνει τα ακόλουθα βήματα:

1. Επιλογή ενός τυχαίου διαστήματος απο μια χρονοσειρά. Συνήθως το μήκος του διαστήματος είναι δύναμη του 2.

2. Για το ίδιο διάστημα, εφαρμόζεται η εξαγωγή χαρακτηριστικών που αναφέρθηκε προηγουμένως (συγχώνευση συντελεστών αυτοσυσχέτισης και συντελεστών φάσματος ισχύος) σε όλες τις χρονοσειρές

3. Δημιουργείται ένα νέο training set, με βάση τα χαρακτηριστικά που εξήχθησαν απο τα επιμέρους διαστήματα όλων των χρονοσειρών

4. Εκπαιδεύεται ένα δέντρο απόφασης

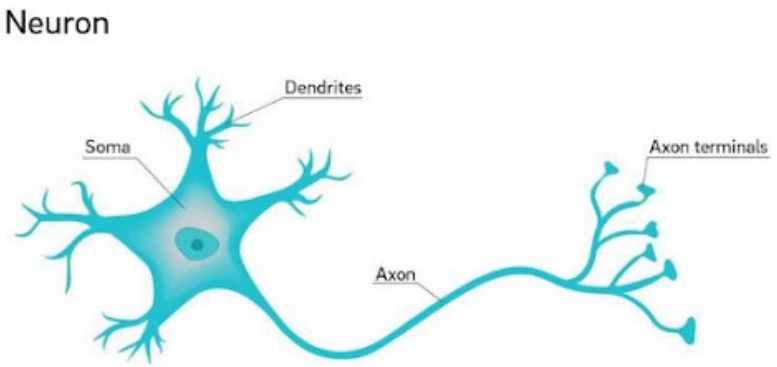
5. Επαναλαμβάνονται τα βήματα 1-4, εως ότου εκπαιδευτεί ο επιθυμητός αριθμός δέντρων απόφασης.

3.4 Νευρωνικά Δίκτυα

Το κέντρο του ανθρώπινου συστήματος είναι ο ανθρώπινος εγκέφαλος, ο οποίος αναπαριστάνεται απο το νευρωνικό (νευρικό) δίκτυο. Το δίκτυο αυτό προσλαμβάνει συνεχώς πληροφορίες, που αφορούν αλλαγές τοσο στο εξωτερικό περιβάλλον, όσο και στον ίδιο μας τον οργανισμό. Οι πληροφορίες συλλέγονται μέσω των υποδοχέων, οι οποίοι διακρίνονται σε κατηγορίες όπως μηχανοϋποδοχείς, χημειοϋποδοχείς, θερμοϋποδοχείς κλπ. Με λίγα λόγια, οι υποδοχείς μετατρέπουν τα ερεθίσματα σε ηλεκτρικά σήματα (ώσεις), τα οποία διαβιβάζονται στο νευρικό δίκτυο.

3.4.1 Νευρώνας Ανθρώπινου Εγκεφάλου

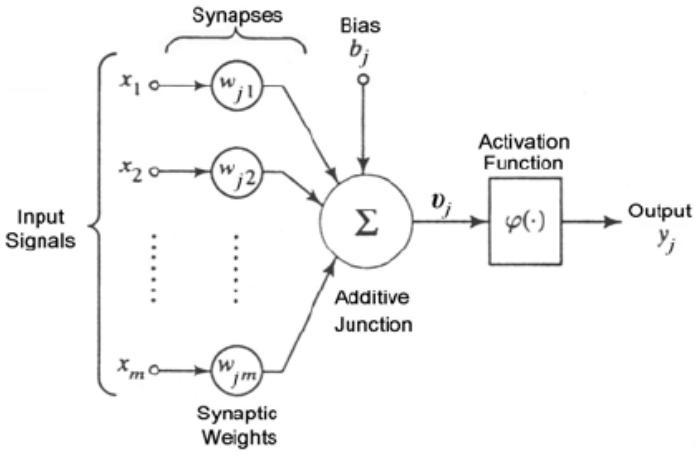
Ο αγώνας για την κατανόηση του ανθρώπινου εγκεφάλου έχει γίνει ευκολότερος χάρη στο πρωτοποριακό έργο του Ramon y Cajal (1911), ο οποίος εισήγαγε την ιδέα των νευρώνων ως δομικά συστατικά του εγκεφάλου. Ένας νευρώνας δεν είναι τίποτα άλλο πέρα απο ένα κύτταρο (νευρικό κύτταρο). Το κύτταρο αυτό μέσω των δενδριτών συλλέγει τις νευρικές ώσεις (σήματα), που προέρχονται απο τους υποδοχείς. Οι ώσεις συγκεντρώνονται στο σώμα του κυττάρου, όπου και βρίσκεται ο πυρήνας. Στην συνέχεια μεταδίδονται μέσω μιας λεπτής ίνας, που ονομάζεται νευράξονας ή απλά αξονας, στις συναπτικές απολήξεις (axon terminals). Τα σήματα αυτά μπορούν να μεταδοθούν και σε άλλους νευρώνες που επικοινωνούν μεταξύ τους μέσω συνάψεων



Σχήμα 3.6: Δομή νευρικού κυττάρου του ανθρώπου

3.4.2 Τεχνητός Νευρώνας - Perceptron

Με πηγή έμπνευσης τη νευροφυσιολογία του ανθρώπινου εγκεφάλου αναπτύχθηκαν τεχνητοί νευρώνες, οι οποίοι στην συνέχεια μοντελοποιήθηκαν χρησιμοποιώντας θεωρητικά και υπολογιστικά εργαλεία. Στο παρακάτω σχήμα, παρουσιάζεται το μοντέλο ενός τεχνητού νευρώνα, στο οποίο κανείς μπορεί να διακρίνει τρια βασικά στοιχεία:



Σχήμα 3.7: Δομή Τεχνητού Νευρώνα - Perceptron

1. Ένα σύνολο συνάψεων (ή διασυνδέσεων), κάθε μια απο τις οποίες χαρακτηρίζεται απο το δικό της βάρος

2. Έναν αθροιστή για την άθροιση των σημάτων εισόδου, σταθμισμένων απο τα αντίστοιχα συναπτικά βάρη του νευρώνα

3. Μια συνάρτηση ενεργοποίησης φ() για τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα σε κάποια πεπερασμένη τιμή.

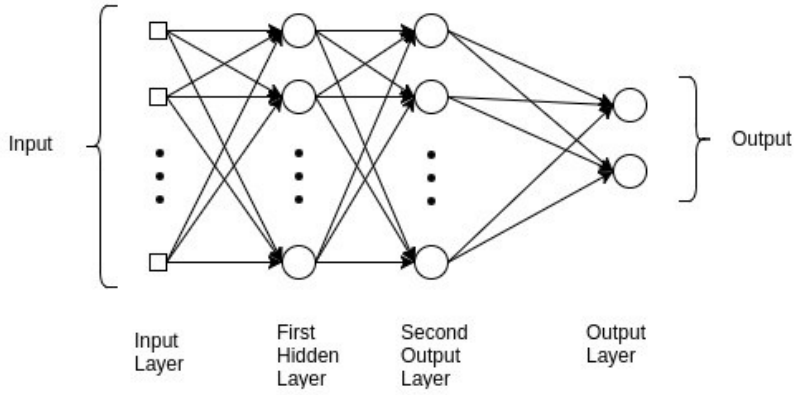
4. Με μαθηματικούς όρους, ένας νευρώνας μπορεί να περιγραφεί από το ζεύγος των ακόλουθων εξισώσεων:

$$u_j = \sum_{i=1}^{m} w_{ji} x_i$$
$$y_j = \phi(b_j + u_j)$$

3.4.3 Πολυεπίπεδα νευρωνικά δίκτυα

Για την επίλυση πιο σύνθετων προβλημάτων, οι επιμέρους νευρώνες (perceptrons) συνδυάζονται και δημιουργούν πολυεπίπεδες αρχιτεκτονικές νευρώνων. Οι δομές αυτές καλούνται Multilayer Perceptrons (βλ. Σχήμα 3.8), και τα επίπεδα στα οποία οργανώνονται οι νευρώνες χωρίζονται στις εξής κατηγορίες: (a) Επίπεδο εισόδου (Input layer), (b) Κρυφά επίπεδα (Hidden layer), (c) Επίπεδο εξόδου (Output layer)

Να σημειωθεί ότι τα πολυεπίπεδα δίκτυα, μπορούν να είναι πλήρως συνδεδεμένα (fully connected), δηλαδή κάθε νευρώνας ενός επιπέδου να συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου ή σε αντίθετη περίπτωση μερικώς συνδεδεμένα (partially connected). Επίσης, τα δίκτυα μπορούν να χωριστούν σε δίκτυα πρόσθιας τροφοδότησης (feedforward), στην περίπτωση που η πληροφορία μεταβιβάζεται σειριακά απο το επίπεδο εισόδου στο επίπεδο εξόδου, ή σε αναδρομικά δίκτυα (recurrent) που περιλαμβάνουν βρόχους ανάδρασης.



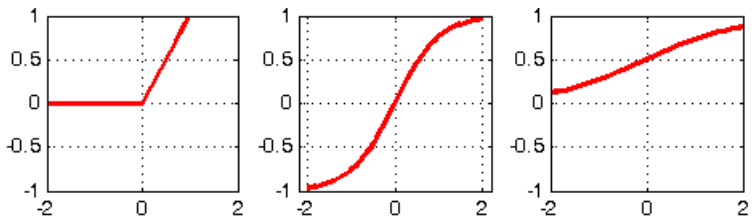
Σχήμα 3.8: Πολυεπίπεδο Νευρωνικό Δίκτυο, συνδυάζοντας τεχνητούς νευρώνες

3.4.4 Συναρτήσεις ενεργοποίησης

Ένα απο τα βασικά στοιχεία ενός Perceptron, αποτελεί η Συνάρτηση Ενεργοποίησης (Activation Function), όπως φαίνεται και στο σχήμα 3.9. Μέσω αυτών των συναρτήσεων καθορίζεται η τιμή της εξόδου που θα διαβιβαστεί στον επόμενο νευρώνα με βάση την είσοδο. Χρησιμοποιώντας μια γραμμική συνάρτηση $\phi() = x$, δεν εφαρμόζεται κανένας μετασχηματισμός στην είσοδο. Στην περίπτωση αυτή, το νευρωνικό δίκτυο είναι πιο εύκολο να εκπαιδευτεί (βλ παρακάτω 3.4.7), ωστόσο είναι δύσκολη η εκμάθηση περίπλοκων σχέσεων και απεικονίσεων μεταξύ εισόδου - εξόδου. Για το λόγο αυτό, στους κόμβους των hidden layers χρησιμοποιούνται μη γραμμικές συναρτήσεις. Πιο ειδικά το άθροισμα των σημάτων εισόδου σε ένα κόμβο, σταθμισμένων απο τα αντίστοιχα βάρη, αντιστοιχίζεται και ερμηνεύεται πλεον σε ένα επιθυμητό διάστημα. Το διάστημα αυτό ορίζεται απο την εκάστοτε συνάρτηση ενεργοποίησης. Ορισμένες απο τις συναρτήσεις που χρησιμοποιούνται στην πράξη είναι:

- ReLU: Η συνάρτηση ReLU μηδενίζει τις αρνητικές εισόδους και εφαρμόζει ένα γραμμικό μετασχηματισμό στις θετικές. Το πεδίο τιμών της είναι το διάστημα $[0, \infty)$. Πρόκειται για μια ευρύτατα διαδεδομένη συνάρτηση ενεργοποίησης που δεν αντιμετωπίζει το πρόβλημα της εξαφάνισης κλίσης (gradient vanishing). Το πρόβλημα αυτό εξηγείται στην παράγραφο Εκπαίδευση Νευρωνικού (βλ. 3.4.9)

- Tanh-sigmoid: Η tanh είναι μια μη γραμμική συνάρτηση που αντιστοιχίζει τις τιμές εισόδου στο σύνολο $(-1, 1)$, ενώ η σιγμοειδής συνάρτηση, που ονομάζεται και logistic, μετατρέπει την είσοδο στο διάστημα $(0, 1)$. Και οι δύο συναρτήσεις, οι οποίες έχουν παρόμοιο σχήμα (βλ. 3.9), οδηγούν σε κορεσμό. Αυτό πρακτικά σημαίνει ότι, οι τιμές που είναι πολύ μεγάλες αντιπροσωπεύονται απλά απο την μονάδα, ενώ οι πολύ μικρές τιμές αντιπροσωπεύονται από το -1 για tanh (αντ. 0 για sigmoid). Για το λόγο αυτό, οι συναρτήσεις αυτές είναι ευαίσθητες ως προς αλλαγές τιμών γύρω απο τα κέντρα τους (0 για tanh, 0.5 για sigmoid). Αυτές οι συναρτήσεις χρησιμοποιούνται συχνα στους κόμβους των hidden layers.



Σχήμα 3.9: Συναστήσεις Ενεργοποίησης: ReLU, tanh, sigmoid, αντίστοιχα

Κάποιες συναρτήσεις ενεργοποίησης, χρησιμοποιούνται συχνότερα στο επίπεδο εξόδου (Output Layer) των νευρωνικών δικτύων. Οι συναρτήσεις αυτές, δεν δέχονται σαν είσοδο μια τιμή x, αλλά ένα διάνυσμα τιμών που προέκυψαν απο το προηγούμενο επίπεδο. Συχνά χρησιμοποιείται η softmax

softmax: Με βάση το διάνυσμα εισόδου $x = [x_1, x_2, x_3, ..., x_J]$ υπολογίζεται για κάθε $0 <= i <= J$ η τιμή

$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}} \tag{3.3}$$

Το αποτέλεσμα είναι ένα διάνυσμα J στοιχείων.

3.4.5 Έξοδος Νευρωνικού

Στην περίπτωση λοιπόν, που το πρόβλημα προς επίλυσης είναι δυαδικό, χρησιμοποιείται ένας κόμβους εξόδου με συνάρτηση ενεργοποίησης σιγμοειδή (αντ. tanh). Έτσι αν η τιμή που προκύπτει είναι μεγαλύτερη απο ένα κατώφλι (συνήθως 0.5 για tanh και αντίστοιχα 0 για σιγμοειδή), τότε ανήκει στην κατηγορία A, ενώ σε αντίθετη περίπτωση στην κατηγορία B.

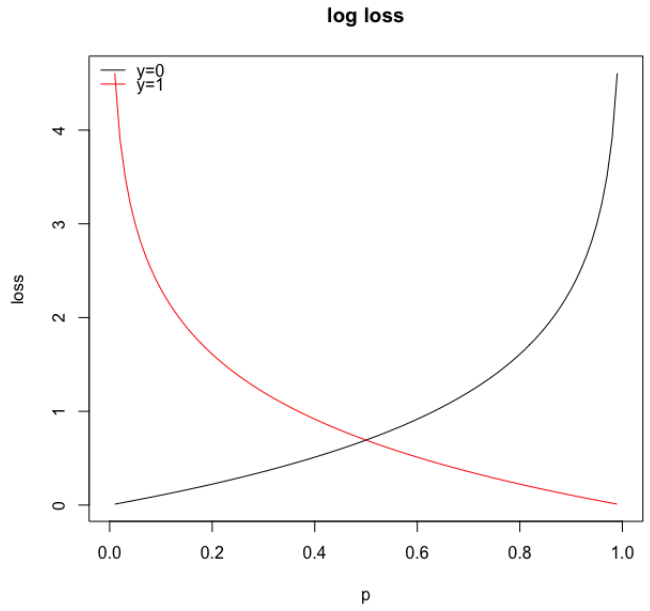
Στην περίπτωση ωστόσο,που το πρόβλημα προς επίλυση είναι ταξινόμηση πολλών κλάσεων, χρησιμοποιούνται τόσοι κόμβοι εξόδου όσες και οι κατηγορίες. Χρησιμοποιώντας συνάρτηση ενεργοποίησης softmax, δίνεται για κάθε κατηγορία (κόμβο) ένας αριθμός που δηλώνει την πιθανότητα να ανήκει το δείγμα στην κλάση αυτή. Βέβαια μπορεί να χρησιμοποιηθεί σε κάθε κόμβο tanh ή σιγμοειδή και το δείγμα να αποδοθει στον κόμβο-κατηγορία με την μεγαλύτερη τιμή εξόδου.

3.4.6 Συναρτήσεις Κόστους

Οι συναρτήσεις κόστους, αποτελούν μια μέθοδο αποτίμησης της απόδοσης ενός νευρωνικού δικτύου, κατά την διάρκεια της εκπαίδευσης. Πιο ειδικά για κάθε δείγμα εκπαίδευσης, συγκρίνεται η έξοδος του δικτύου (prediction) με την αναμενόμενη τιμή (label). Συχνά χρησιμοποείται το κόστος διασταυρούμενης εντροπίας (Cross Entropy Loss), το οποίο αρχικά εξηγείται για ένα πρόβλημα δυαδικής ταξινόμησης.

Cross Entropy Loss:

$$L = -(y \log(p) + (1 - y) \log(1 - p)) \tag{3.4}$$



Σχήμα 3.10: Συνάρτηση κόστους log loss για πρόβλημα δυαδικής ταξινόμησης

Έστω λοιπόν ότι ένα δείγμα που ανήκει στην κατηγορία A, αποδίδεται απο το νευρωνικό ότι ανήκει στην κατηγόρία A, με πιθανότητα p. Στην περίπτωση αυτή, η πρόβλεψη είναι σωστή ($y = 1$) και ο τύπος μετασχηματίζεται σε -log(p) που απεικονίζεται παραπάνω με κόκκινο χρώμα. Επιθυμητό είναι να λάβουμε $L = 0$, δηλαδή οι σωστές προβλέψεις να δίνονται με μεγάλη πιθανότητα. Με τον τρόπο αυτό, 'τιμωρούνται' προβλέψεις που ήταν σωστές αλλά δόθηκαν με μικρή πιθανότητα.

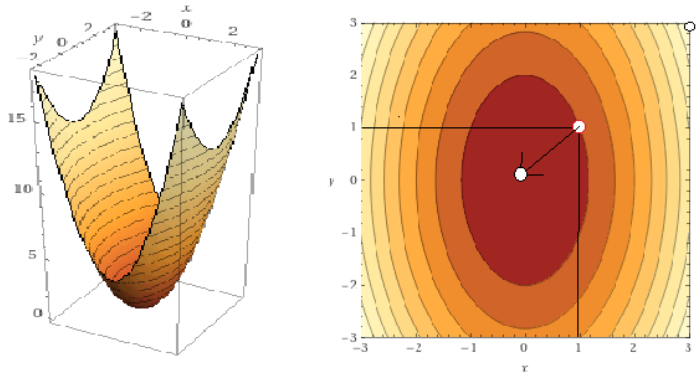
Έστω λοιπόν ότι ένα δείγμα που ανήκει στην κατηγορία A, αποδίδεται απο το νευρωνικό ότι ανήκει στην κατηγορία B, με πιθανότητα p. Στην περίπτωση αυτή, η πρόβλεψη είναι λάθος ($y = 0$) ο τύπος μετασχηματίζεται σε -log(1-p) που απεικονίζεται παραπάνω με μαύρο χρώμα. Είναι επιθυμητό είναι να λάβουμε $L = 0$, δηλαδή οι λάθος προβλέψεις να δίνονται με μικρή πιθανότητα. Με τον τρόπο αυτό, 'τιμωρούνται' προβλέψεις που ήταν λάθος, αλλά δόθηκαν με μεγάλη πιθανότητα (σιγουριά).

Τα παραπάνω μπορούν να γενικευτούν σε περιπτώσεις όπου διαθέτουμε περισσοτερες απο δυο κατηγορίες (έστω M) για ταξινόμηση

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{3.5}$$

3.4.7 Εκπαίδευση νευρωνικού

Η εκπαίδευση ενός νευρωνικού, είναι μια μέθοδος για την εύρεση των κατάλληλων τιμων των βαρών w, ελαχιστοποιώντας μια συνάρτηση κόστους. Σε πρώτη φάση αρχικοποιούνται τυχαία τα βάρη του μοντέλου. Στην συνέχεια, με βαση την συνάρτηση κόστους τροποποιούνται κατάλληλα, ώστε να ελαχιστοποιηθεί η συνάρτηση αυτή.



Σχήμα 3.11: Υπολογισμός κλίσης για τη συνάρτηση $f(x, y) = x^2 + y^2$ στο [1,1]

Γίνεται λοιπόν αντιληπτό ότι η τιμή της συνάρτηση κόστους για μια δεδομένη παρατήρηση, εξαρτάται απο τις τιμές των βαρών w. Η γνώσης της παραγώγου της συνάρτησης κόστους (μιας συνάρτησης πολλών μεταβλητών), μπορεί να δώσει χρήσιμη πληροφορία για την κατεύθυνση, όπου ελαχιστοποιείται η συνάρτηση. Πιο ειδικά, σε συναρτήσεις πολλών μεταβλητών, η παράγωγος της συνάρτησης (grad) δίνει την κατεύθυνση κατα την οποία η συνάρτηση αυξάνεται πιο απότομα. Για παράδειγμα, στο σημείο $[1, 1]$ το gradient ισούται με $[3, 3]$ που σημαινει οτι προς αυτή την κατεύθυνση η συνάρτηση αυξάνεται απότομα, οπότε κινούμαστε ανάποδα, όπως ορίζει το -grad. Η διαδικασία αυτή επαναλαμβάνεται.

Απο τη στιγμή που υπολογίζεται το gradient, χρησιμοποιείται για να ανανεωθούν οι τιμές των βαρών σύμφωνα με τη σχέση (3.9): Η διαδικασία αυτή (ανανέωση βαρών) μπορεί να γίνει, είτε μεμονωμένα για κάθε δείγμα είτε σε ομάδες (batches) δειγμάτων είτε για όλο το σύνολο εκπαίδευσης. Έτσι διακρίνονται τρεις κατηγορίες εκπαίδευσης:

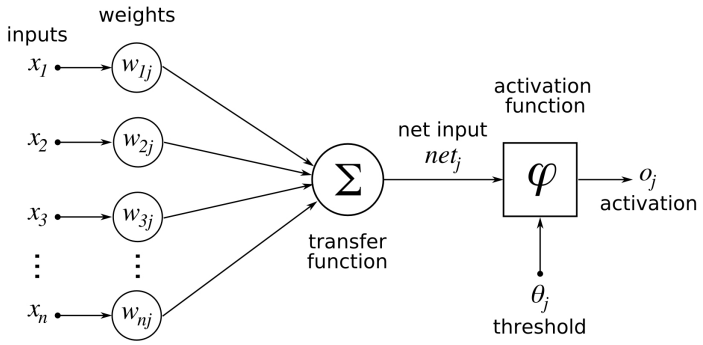
- Stochastic Gradient Descent: Στην περίπτωση αυτή, για κάθε νέο δείγμα εκπαίδευσης, υπολογίζεται το γραδιεντ της συνάρτησης και πραγματοποιείται ανανέωση των βαρών -grad

- Mini Batch Gradient Descent: Στην περίπτωση αυτή, τα δείγματα χωρίζονται σε ομάδες δεδομένων και όταν μια ομάδα τροφοδοτηθεί στο νευρωνικό, τότε λαμβάνεται ο μέσος όρος των γραδιεντς και ανανεώνονται τα βάρη.

- Gradient Descent: Στην περίπτωση αυτή, λαμβάνεται ο μέσος όρος των gradients για κάθε κατηγορία ξεχωριστά, αφού έχουν τροφοδοτηθεί όλα τα δείγματα εκπαίδευσης στο νευρωνικό (μια εποχή).

Ωστόσο αυτό που δεν αναφέρθηκε είναι ο τρόπος με τον οποίο υπολογίζεται το gradient της συνάρτησης κόστους για ένα βάρος. Ο αλγόριθμος που χρησιμοποιείται στην πράξη είναι ο backpropagation, που σχεδιαστηκε για τον υπολογισμό παραγώγων σε δομές γράφων, ξεκινώντας απο την έξοδο προς την είσοδο. Η διαδικασία αυτή γίνεται εφικτή χρησιμοποιώντας τον κανόνα της αλυσίδας (Chain Rule).

Αλγόριθμος Backpropagation



Σχήμα 3.12: Δομή Τεχνητού Νευρώνα - Perceptron

Η τιμή του βάρους w_{ij} ανανεώνεται σύμφωνα με τη σχέση $w_{ij} = w_{ij} - \partial L/\partial w_{ij}$. Η μερική παράγωγος $\partial L/\partial w_{ij}$ υπολογίζεται με βάση τον κανόνα της αλυσίδας:

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial o_j}\frac{\partial o_j}{\partial w_{ij}} = \frac{\partial L}{\partial o_j}\frac{\partial o_j}{\partial net_j}\frac{\partial net_j}{\partial w_{ij}} \tag{3.6}$$

Στον τελευταίο παράγοντα το άθροισμα net_j όταν παραγωγιστεί μερικώς ως προς w_{ij} θα δώσει:

$$\frac{\partial net_j}{\partial w_{ij}} = \frac{\partial}{w_{ij}}\left(\sum_{k=1}^{n} w_{kj}x_k\right) = \frac{\partial}{w_{ij}} w_{ij}x_i = x_i \tag{3.7}$$

Η έξοδος o_i προκύπτει απο την ενεργοποίηση του αθροίσματος net_j με βάση κάποια συνάρτηση φ. ($o_i = \phi(net_j)$:

$$\frac{\partial o_i}{\partial net_j} = \frac{\partial \phi(net_j)}{\partial net_j} \tag{3.8}$$

Αντικαθιστώντας τις εξισώσεις (3.6), (3.7) στην εξίσωση (3.8), αποτιμάται η τιμή της μερικής παραγώγου $\partial L/\partial w_{ij}$.

3.4.8 Αλγόριθμοι Βελτιστοποίησης

Με τον όρο βελτιστοποίηση αναφερόμαστε στην εύρεση των τιμών μιας συνάρτησης που την μεγιστοποιούν ή αντίστοιχα την ελαχιστοποιούν. Η εκπαίδευση ενός νευρωνικού είναι στην ουσία ένα πρόβλημα ελαχιστοποίησης της συνάρτησης κόστους. Το gradient που υπολογίστηκε με τον αλγόριθμο backpropagation, μπορεί να χρησιμοποιηθεί απο τους ακόλουθους αλγορίθμους βελτιστοποίησης, κάποιοι απο τους οποίους αναγράφονται ακολούθως:

- SGD

$$w_i(t) = w_i(t-1) - \eta \cdot \partial L/\partial w_i \tag{3.9}$$

Στην προκειμένη περίπτωση, η νέα τιμή του βάρους διαφέρει απο την προηγούμενη κατά ένα παράγοντα $\eta \cdot \partial L/\partial w_i$. Με η συμβολίζουμε το ρυθμό εκμάθησης (learning rate).

- Adam: Το όνομα του, προέρχεται απο τη φράση "adaptive moment estimation". Πρόκειται για μια επέκταση/τροποποίηση του αλγορίθμου SGD. Στον αλγόριθμο SGD, για όλες τις ανανεώσεις βαρών, χρησιμοποιείται το ίδιο learning rate, και μάλιστα δεν τροποποιείται κατά την διάρκεια της εκπαίδευσης. Αντιθέτως με τη χρήση του αλγορίθμου Adam, κάθε βάρος ανανεώνεται με διαφορετικό learning rate και μάλιστα αυτό μπορεί να αλλάξει κατά τη διάρκεια της εκπαίδευσης, δηλαδή για κάθε νέο δείγμα εκπαίδευσης.

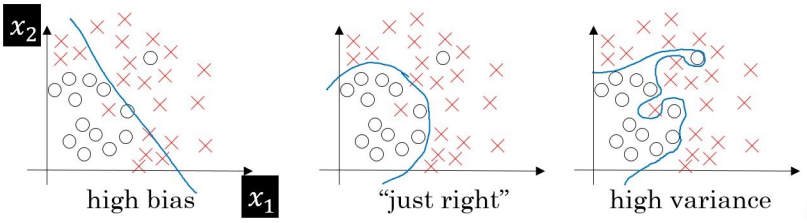
- Adamax: Πρόκειται για μια επέκταση του αλγορίθμου Adam. Η ανανέωση των βαρών πραγματοποιείται με αντίστοιχο τρόπο με τον αλγόριθμο Adam. Όταν λοιπόν η τιμή του gradient είναι πολύ μικρή, (κοντά στο μηδεν), αγνοείται και στη θέση της λαμβάνεται μια άλλη ποσότητα. Στην περίπτωση αυτή τα βάρη επηρεάζονται λιγότερες φορές απο gradients. Κάτι τέτοιο είναι αναγκαίο ορισμένες φορές που τα gradients αναπαριστούν θόρυβο, και όχι κάποια σημαντική πληροφορία για την κατεύθυνση του ελαχίστου κάποιας συνάρτησης κόστους.

3.4.9 Το πρόβλημα των vanishing gradients

Κατα την διάρκεια του backpropagation, η τιμή του κάθε βάρους προκύπτει αφαιρώντας από την τρέχουσα τιμή την μερική παράγωγο της συνάρτησης κόστους αναφορικά με το συγκεκριμένο βάρος .Η τιμή τελευταία τιμή υπολογίζεται με τον κανόνα της αλυσίδας. Εξαιτίας των διαδοχικών πολλαπλασιασμών , είναι πιθανό η τιμή που θα προκύψει να είναι αρκετά κοντά στο μηδέν. Αυτο πρακτικά σημαίνει, ότι η νέα τιμή βάρους θα είναι ίση με την προηγούμενη. Γενικεύοντας αυτό το πρόβλημα και για τα υπόλοιπα βάρη του δικτύου, το νευρωνικό δίκτυο πλέον δεν εκπαιδεύεται.

3.4.10 Το πρόβλημα της υπερποσαμογής - overfitting

Στόχος των αλγορίθμων βελτιστοποίησης, μεταξύ άλλων, είναι να επιλύσουν το ζήτημα της υποπορσαρμογής (underfitting). Για παράδειγμα σε ένα πρόβλημα δυαδικής ταξινόμησης το πρόβλημα (underfitting) συναντάται όταν η διαχωριστική καμπύλη δεν διαμερίζει πλήρως τα πρότυπα του συνόλου εκπαίδευσης (βλ. Σχήμα 3.13) Η κατάσταση αυτή χαρακτηρίζεται απο υψηλό bias, ενώ στην αντίθετη περίπτωση όπου τα πρότυπα εκπαίδευσης διαμερίζονται πλήρως, είμαστε σε μια κατάσταση low bias. Επιλέγοντας ωστόσο τη δεύτερη περίπτωση (με low bias), θα παρατηρήσουμε ότι στο σύνολο ελέγχου (test set), τα πρότυπα δεν διαχωρίζονται επιτυχώς. Η διαφορά (variation) αυτή που παρατηρείται, απο την επιλογή μιας καμπύλης, στο διαχωρισμό προτύπων που ανήκουν στο σύνολο εκπαίδευσης και προτύπων που ανήκουν στο σύνολο ελέγχου, καλείται variance. Υψηλο variance έχουμε στην εικόνα δεξιά (βλ. Σχήμα 3.13), όπου η καμπύλη, μπορεί να προσαρμόζεται άψογα στο σύνολο εκπαίδευσης, ωστόσο είναι πιθανόν να μην πετύχει στο σύνολο ελέγχου.



Σχήμα 3.13: Ταξινόμησης δειγμάτων: Υποπροσαρμογή (αριστερά), Υπερπροσαρμογή (δεξιά)

3.4.11 Τρόποι Αντιμετώπισης overfitting

Προκειμένου να αντιμετωπιστεί το πρόβλημα της Υπερπροσαρμογής (overfitting), μπορούν να χρησιμοποιηθούν οι ακόλουθες τεχνικές:

- Επαύξηση του συνόλου δεδομένων (Data Augmentation): Όταν υπάρχει ένας μικρός αριθμός δειγμάτων εκπαίδευσης, το μοντέλο μερικές φορές μαθαίνει από θορύβους ή ανεπιθύμητες λεπτομέρειες από αυτά τα δείγματα, σε βαθμό που επηρεάζεται αρνητικά την απόδοση του μοντέλου σε νέα παραδείγματα. Αυτό το φαινόμενο είναι γνωστό ως overfitting. Ένας τρόπος για να διορθώσουμε αυτό το πρόβλημα είναι να αυξήσουμε το σύνολο δεδομένων εκπαίδευσης, χρησιμοποιώντας τυχαίους μετασχηματισμούς (περιστροφές, μετατοπίσεις κ.τ.λ.) των αρχικών εικόνων.

- Πρόωρη διακοπή εκπαίδευσης (Early Stopping): Κατά την διάρκεια της εκπαίδευσης, χρησιμοποιείται ένα σύνολο δεδομένων (validation set) για την εκτιμηση της συμπεριφοράς του μοντέλου σε άγνωστα δεδομένα. Υπάρχει περίπτωση το μοντέλο να εκπαιδεύεται επιτυχώς στο σύνολο εκπαίδευση (η συνάρτηση κόστους να μειώνεται ανα εποχή), αλλά να μη συμβαίνει το ίδιο για το σύνολο επικύρωσης. Στη περίπτωση αυτή, το μοντέλο αδυνατεί να γενικεύσει και καλό είναι να διακόπτεται η διαδικασία της εκπαίδευσης, ώστε να μην παρατηρηθεί υπερπροσαρμογή στα δεδομένα εκπαίδευσης. Η διαδικασία διακοπής της εκπαίδευσης καλείται Early Stopping.

- Περιορισμός ενεργοποίησης (Dropout): Ορισμένα "μονοπάτια" σε ένα δίκτυο δεν αντιπροσωπεύουν πάντα κάποια κλάση, αλλά πιθανώς ανεπιθύμητο θόρυβο. Στις περιπτώσεις αυτές καλο είναι να παρεμβάλλεται ένα επίπεδο dropout, ώστε να ακυρωθούν κάποιες απο τις εισόδους του.

- Κανονικοποίηση L1 L2 : Μια μέθοδος για την αντιμετώπιση του overfitting είναι η κανονικοποίηση L1 ή αντίστοιχα η κανονικοποίηση L2. Στην πρώτη περίπτωση χρησιμοποιείται η νορμα L1 (3.10), ενώ στην δεύτερη περίπτωση η νόρμα L2 (3.11). Για βάρη ορίζονται ως εξής:

$$\|w\|_1 = |w_1| + |w_1| + ... + |w_n| \tag{3.10}$$

$$\|w\|_2 = (|w_1|^2 + |w_1|^2 + ... + |w_n|^2)^{1/2} \tag{3.11}$$

Η κανονικοποίηση αφορά ένα συγκεκριμένο επίπεδο, και πιο ειδικά τα βάρη του. Πιο ειδικά προστίθενται στην συνάρτηση κόστους L ($L = Error(y, \hat{y})$) οι παραπάνω ποσότητες σταθμισμένες με ένα παράγοντα λ.

$$L_1 = L + \lambda \sum_{i=1}^{N} |w_i| \tag{3.12}$$

$$L_2 = L + \lambda \sum_{i=1}^{N} w_i^2 \tag{3.13}$$

Στο απλοποιημένο παράδειγμα, όπου η συνάρτηση απωλειών L είναι διαφορά μεταξύ πραγματικής τιμής και πρόβλέψης και έχουμε ένα νευρώνα, τα βάρη ανανεώνονται ως εξής: L1:

$$\begin{aligned} w_{new} &= w - \eta \cdot \partial L_1 / \partial w \\ &= w - \eta \cdot [2x(wx + b - \hat{y}) + \lambda \frac{d\,|w|}{dw}] \\ &= \begin{Bmatrix} w - \eta[2x(wx + b - \hat{y}) + \lambda], w > 0 \\ w - \eta[2x(wx + b - \hat{y}) - \lambda], w < 0 \end{Bmatrix} \end{aligned}$$

L2:

$$\begin{aligned} w_{new} &= w - \eta \cdot \partial L_2 / \partial w \\ &= w - \eta \cdot [2x(wx + b - \hat{y}) + 2\lambda \dot{w}] \end{aligned}$$

Θέτοντας $H = 2x(wx + b - \hat{y})$, και για $\eta = 1$ οι παραπάνω εξισώσεις λαμβάνουν την μορφή:

$$w_{new} = w - H$$

Αντίστοιχα λοιπόν για L1, L2 ισχύουν:

$$\begin{aligned} w_{new} &= \begin{Bmatrix} (w - H) - \lambda, w > 0 \\ (w - H) + \lambda, w < 0 \end{Bmatrix} \\ w_{new} &= (w - H) - 2\lambda w \end{aligned}$$

Χωρίς κανονικοποίηση, αν τροφοδοτηθεί ξανά ένα δείγμα εκπαίδευσης η εξοδος του y θα είναι ακόμα πιο κοντά στην έξοδο $\hat{y}$. Ωστόσο εμείς δεν επιθυμούμε τόσο μεγάλη ακρίβεια στα βάρη κατα την εκπαίδευση , γιατί έτσι το μοντέλο δε θα μπορεί να γενικεύσει σε άγνωστα δεδομένα. Η πρόσθεση των όρων ($\pm\lambda$ για L1, $-2\lambda w$ για L2) είναι που καθιστά την επιλογή της νέας παραμέτρου λιγότερο κατάλληλη , κάτι που σύμφωνα με τα παραπάνω (βλ. 3.4.10), είναι επιθυμητό.

Ο δύο μέθοδοι κανονικοποίησης έχουν μια βασική διαφορά. Στην περίπτωση που χρησιμοποιείται κανονικοποίηση L1 να σημειωθεί επίσης ότι τα βάρη οδηγούνται στο μηδεν. Αν το βάρος είναι θετικό, αφαιρείται τιμή, αν είναι αρνητικό προστίθεται. Οπότε, με τον τρόπο αυτό μηδενίζονται βάρη που δεν συνεισφέρουν στην τελική πρόβλεψη και εν τέλει το μοντέλο τείνει να μην χαρακτηρίζεται απο μεγάλη πολυπλοκότητα. Στην περίπτωση που χρησιμοποιείται κανονικοποίηση L2 τα βάρη οδηγούνται στο μηδεν, αλλά με μικροτερο βήμα. Ο όρος που αφαιρείται δεν είναι σταθερός ($\pm\lambda$) αλλά της μορφής ($-2\lambda w$). Όσο πιο μικρό το w τόσο πιο μικρός ο όρος που αφαιρείται.

3.4.12 Αξιολόγηση Απόδοσης Νευρωνικού

Προκειμένου να ποσοτικοποιηθεί το πόσο επιτυχής ήταν η εκπαίδευση ενός νευρωνικού μοντέλου, χρησιμοποιούνται κάποιες μετρικές αξιολόγησης στο σύνολο ελέγχου (test set).

Μετρικές Απόδοσης

Ορθότητα - Accuracy: Πρόκειται για μια απο τις βασικότερες μετρικές αξιολόγησης ενός μοντέλου. Ορίζεται ως το κλάσμα των ορθών προβλέψεων προς τον αριθμό των συνολικών εκτιμήσεων που πραγματοποιήθηκαν

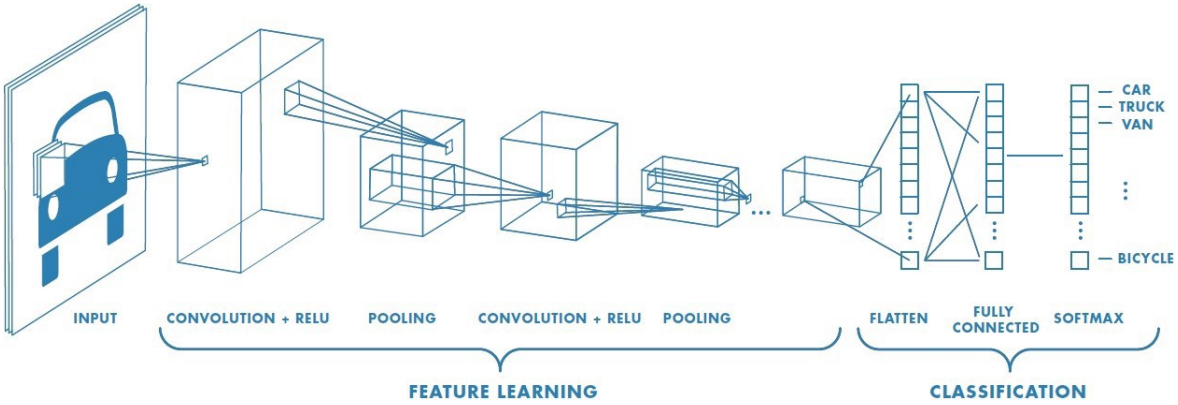
Η μετρική αυτή,ωστόσο καλό είναι να μην χρησιμοποιείται όταν το dataset δεν είναι ισορροπημένο (balanced). Έστω ότι η κατηγορία A ενός προβλήματος δυαδικής ταξινόμησης έχει 9 900 δείγματα, και η κατηγορία B 100. Αν όλα τα δείγματα της κατηγορίας B ταξινομηθούν λανθασμένα στην κατηγορία A, η ακρίβεια του μοντέλου θα είναι πολύ υψηλή και πιο ειδικα: $9900/(9900 + 100) = 99\%$. Για τον λόγο αυτό χρησιμοποιούνται στην πράξη και άλλες μετρικές απόδοσης.

Πίνακας Σύγχυσης: Στην συνέχεια, θα περιγραφεί ο Πίνακας Σύγχυσης. Πρόκειται για μια αναπαράσταση, συγκεκριμένα έναν τετραγωνικό πίνακα μεγέθους NxN, που δίνει χρήσιμη πληροφορία για τα λάθη που συμβαίνουν μεταξύ των N κλάσεων. Πιο ειδικά κάθε τιμή αυτού του πίνακα αναφέρεται στον αριθμό των προβλέψεων που αποδοθηκαν σε μια κλάση (Predicted Class), δεδομένου ότι τα δείγματα άνηκαν σε μια άλλη κλάση (True Class). Στόχος είναι ο πίνακας σύγχυσης να έχει μηδενικά στα στοιχεία εκτός της κύριας διαγωνίου. Στο παρακάτω παράδειγμα διαθέτουμε 3 κλάσεις ($3x3$) πίνακας. Η τιμή 9 του πίνακα (Σχήμα 3.14) ενημερώνει ότι 9 δείγματα αποδόθηκαν στην κλάση Apple ενώ πραγματικά ήταν Mango.

Predicted Class \ True Class	Apple	Orange	Mango
Apple	7	8	9
Orange	1	2	3
Mango	3	2	1

Σχήμα 3.14: Πίνακας Σύγχυσης για 3 κατηγορίες

3.5 Συνελικτικά δίκτυα



Σχήμα 3.15: Δομή Συνελικτικού Νευρωνικού Δικτύου (CNN)

3.5.1 Ορισμός

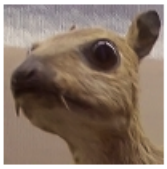
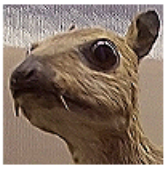
Τα συνελικτικά δίκτυα (Convolutional Neural Networks) αποτελούν μια κατηγορία δικτύων βαθιάς μηχανικής μάθησης, κατάλληλα για να αναλύουν οπτικό περιεχόμενο, δηλαδή εικόνες. Η ανάλυση αυτή, όπως υποδηλώνει και το όνομα, πραγματοποιείται κυρίως με τη χρήση μιας μαθηματικής πράξης, της συνέλιξης. Η αρχιτεκτονική ενός συνελικτικού δικτύου μπορεί να αναλυθεί στα ακόλουθα επίπεδα επεξεργασίας:

3.5.2 Επίπεδα Επεξεργασίας

Input Layer: Η είσοδος ενός συνελικτικού δικτύου μπορεί να είναι μια ή περισσότερες εικόνες. Η διάσταση της εισόδου είναι: (Πλήθος Εικόνων) x (Μήκος Εικόνας) x (Πλάτος Εικόνας) x (Αριθμός Καναλιών). Κάθε εικόνα είναι ένας πίνακας, και κάθε τιμή του πίνακα αντιπροσωπεύει την τιμή ενός εικονοστοιχείου (pixel). Για ασπρόμαυρες (grayscale) εικόνες, όπου ο αριθμός των καναλιών είναι 1, οι τιμές των pixel κυμαίνονται απο 0 (μαύρο) έως 255 (λευκό), ή με κανονικοποίηση στο διάστημα [0,1]. Για έγχρωμες εικόνες, χρησιμοποιείται συχνά ο χρωματικός χώρος RGB, όπου ο αριθμός των καναλιών είναι 3. Πιο ειδικά, κάθε pixel αναπαρίσταται απο τρεις τιμές, καθε μια απο τις οποίες αναφέρεται στο επίπεδο κόκκινου (Red), πρασίνου (Green) και μπλε (Blue).

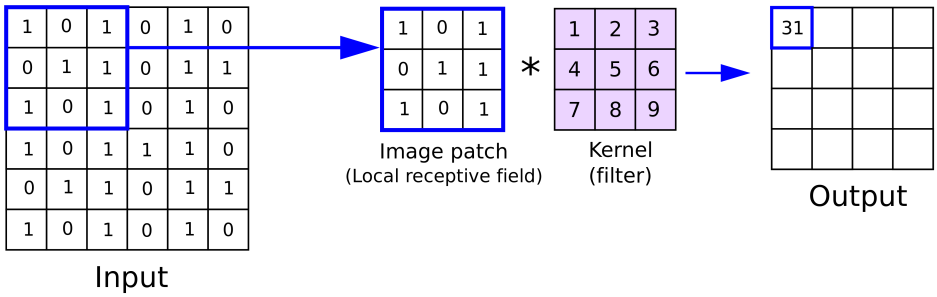
Image Maps: Η εξαγωγή των χαρακτηριστικών μιας εικόνας, πραγματοποιείται στα hidden layers. Η διαδικασία αυτή πραγματοποιείται σταδιακά. Στα πρώτα επίπεδα ενός συνελικτικού δικτύου, εξάγονται χαρακτηριστικά χαμηλού επιπέδου (low level), όπως ακμές. Στα επόμενα επίπεδα, τα χαρακτηριστικά αυτά συντίθεται δημιουργώντας πιο σύνθετες αναπαραστάσεις (μέρη αντικειμένων). Τέλος, όσο πιο βαθύ γίνεται το δίκτυο τα μέρη αυτά ομαδοποιούνται και δημιουργούν τα τελικά αντικείμενα. Προτού αναλυθεί η διαδικασία της εξαγωγής χαρακτηριστικών θεμελιώνονται κάποιες βασικές έννοιες

Πυρήνας (Φίλτρο): Ένας πυρήνας/φίλτρο είναι ένας πίνακας διαστάσεων: (Μήκος Φίλτρου) x (Πλάτος Φίλτρου) x (Πλήθος Καναλιών). Το πλήθος των καναλιών του φίλτρου θα πρέπει να είναι ίδιο με το πλήθος των καναλιών της εικόνας, στο οποίο εφαρμόζεται. Οι τιμές αυτου του φίλτρου ονομάζονται βάρη (weights). Κάθε πυρήνας συνοδεύεται και απο μια τιμή bias. Παρακάτω φαίνεται η επίδραση διαφορετικών πυρήνων σε μια εικόνα μετά την εφαρμογή συνέλιξης (εξηγείται ακολούθως). Να σημειωθεί ότι οι τιμές των βαρών αλλά και του bias προσδιορίζονται κατά την διάρκεια της εκπαίδευσης.

Original	Gaussian Blur	Sharpen	Edge Detection
$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$
			

Σχήμα 3.16: Επίδραση διφορετικών φίλτρων σε εικόνες

Συνέλιξη Φίλτρου με Εικόνα: Ένας πυρήνας πολλαπλασιάζεται με ένα τμήμα της εικόνας (το οποίο πρέπει να έχει τις ίδιες διαστάσεις). Πιο ειδικά πραγματοποιείται ξεχωριστά στοιχείο προς στοιχείο (element-wise) ο πολλαπλασιασμός πινάκων σε κάθε διάσταση και το αποτέλεσμα του πολλαπλασιασμού είναι ένας αριθμός (για κάθε διάσταση). Οι αριθμοί αυτόι αθροίζονται και με το bias και έτσι δίνουν μια τελική τιμή, όπως στο παρακάτω παράδειγμα.



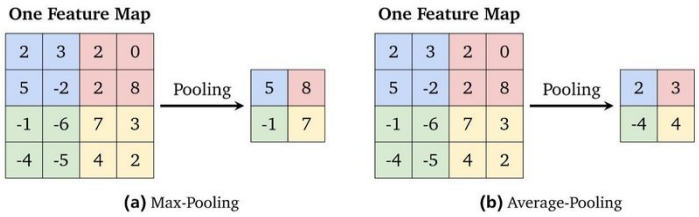
Σχήμα 3.17: Εφαρμογή συνέλιξης φίλτρου σε εικόνα εισόδου

Όσο ολισθαίνει ο πυρήνας πάνω στην εικόνα, παράγονται και άλλοι αριθμοι απο τον πολλαπλασιασμό πινάκων, και έτσι η πράξη της συνέλιξης δημιουργεί ένα Image map. Να σημειωθεί ότι είναι δυνατόν στο ίδιο επίπεδο να χρησιμοποιηθούν και άλλοι πυρήνες με ίδιες διαστάσεις, και έτσι να δημιουργηθουν και άλλα Image Maps (όσα και οι πυρήνες).

Συναρτήσεις Ενεργοποίησης (Activation Functions): Οι τιμές των image

maps μπορούν να 'περάσουν' από μια συνάρτηση ενεργοποίησης (βλ. 3.4.4) και έτσι να απεικονιστούν σε ένα διαφορετικό σύνολο τιμών. Για παράδειγμα στο σύνολο [0, ∞) αν πρόκειται για ReLU, στο (-1,1) για tanh ή στο (0,1) για sigmoid.

Επίπεδα Υποδειγματοληψίας (Pooling Layers): Μεταξύ των Convolutional Layers, είναι δυνατόν να παρεμβάλλονται κάποια επίπεδα που μειώνουν τις διαστάσεις των activation maps. Τα επίπεδα αυτά καλούνται pooling layers και στην πράξη χρησιμοποιούνται πιο συχνά δύο είδη υποδειγματοληψίας: Max pooling / Average pooling. Στις περιπτώσεις αυτές καθορίζεται το μέγεθος του παραθύρου στο οποίο θα εφαρμοστεί pooling.



Σχήμα 3.18: Εφαρμογή Υποδειγματοληψίας σε Image Map

Fully Connected Layers: Η αρχιτεκτονική ενός συνελικτικού δικτύου ολοκληρώνεται με ένα τελικό επίπεδο NN. Οι τιμές των τελικών activation maps, όπως προέκυψαν μετά απο διαδοχικά συνελικτικά επίπεδα, αποτελούν την είσοδο πλήρως συνδεδεμένων δικτύων. Για να πραγματοποιηθεί αυτό, μετατρέπονται σε μονοδιάστατα διανύσματα. Στην περίπτωση περισσοτέρων activation maps οι τιμές συγχωνεύονται κάθετα . Πλέον έχει διαμορφωθεί το διάνυσμα εισόδου που θα αποτελέσει την είσοδο ενός fully connected layer.

3.5.3 Υπερπαράμετοι Συνελικτικού Επιπέδου

Έστω ένα συνελικτικό επίπεδο, η είσοδος του οποίου έχει διαστάσεις $W_1 x H_1 x D_1$.

- Μέγεθος Φίλτρου: Οι διαστάσεις του φίλτρου (FxF) επιλέγονται να είναι μικρότερες απο τις διαστάσεις της εικόνας, στην οποία εφαρμόζεται.

- Βάθος Εξόδου: Το βάθος εξόδου (K) αναφέρεται στο πλήθος των φίλτρων που θα χρησιμοποιηθούν, με κάθε φιλτρο να εξαγει διαφορετικά χαρακτηριστικά.

- Βήμα: Η παράμετρος stride (S) αυτή ρυθμίζει το βημα ολίσθησης του φίλτρου πάνω στην εικόνα. Πολλές φορές το βήμα ολίσθησης πρέπει να επιλεγεί κατάλληλα. Για εικόνες NxN, και φίλτρο FxF, θα πρέπει η διαφορά $N - F$ να είναι πολλαπλάσιο του βήματος (stride). Σε διαφορετική περίπτωση, το φίλτρο δεν μπορεί να εφαρμοστεί πλήρως στα ακραία τμήματα της εικόνας.

- Συμπλήρωση Ακραίων τιμών (padding): Για να εφαρμοστεί πλήρως ένα φίλτρο επιθυμητών διαστάσεων με οποιοδήποτε stide σε μια εικόνα, πραγματοποιείται συμπλήρωση ακραίων τιμών. Οι τιμές αυτές είναι συνήθως μηδενικά και καθορίζουν ένα 'περίγραμμα' πάχους P μηδενικών.

Με βάση αυτές τις παραμέτρους η έξοδος έχει διαστάσεις: $W_2 x H_2 x D_2$, όπου:

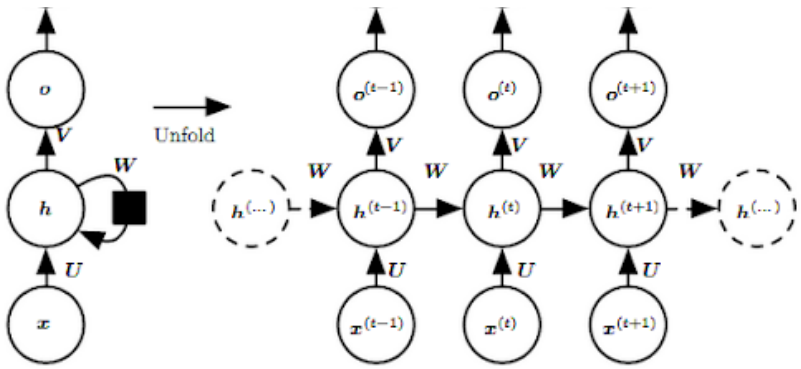
$$W_2 = (W_1 - F + 2P)/2 + 1$$
$$H_2 = (H_1 - F + 2P)/2 + 1$$
$$D_2 = K$$

3.6 Αναδρομικά Δίκτυα

Η αποκωδικοποίηση κάποιων δεδομένων είτε απο τον ανθρώπινο εγκέφαλο είτε απο κάποιο υπολογιστικό σύστημα, προϋποθέτει σειριακή επεξεργασία. Τα δεδομένα αυτά ονομάζονται ακολουθιακά. Για παράδειγμα, όταν κάποιος μελετά ένα κείμενο, δεν αποκωδικοποιεί το νόημα μιας λέξης αυτόνομα, αλλά με βάση το νόημα των προηγουμένων όρων - προτάσεων μπορεί εύστοχα να καταλήξει στην σωστή ερμηνεία της. Αντίστοιχα, η ερμηνεία ενός βίντεο ή ακουστικού σήματος βασίζεται στην ακολουθιακή ανάλυση των επιμέρους καρέ η φωνηματων αντίστοιχα. Ωστόσο τα feedforward δίκτυα που μελετήθηκαν προηγουμένως, δεν έχουν την δυνατότητα ούτε βραχυπρόθεσμης μνήμης, αλλά ούτε σειριακής ανάλυσης.

Την αδυναμία αυτή, έρχονται να καλύψουν τα αναδρομικά δίκτυα. Στα συγκεκριμένα δίκτυα, οι συνδέσεις μεταξύ των κόμβων σχηματίζουν έναν κατευθυνόμενο γράφο κατά μήκος της ακολουθίας. Με τον τρόπο αυτό μπορεί κάθε κόμβος να αναλύει ένα συγκεκριμένο μέρος της ακολουθίας και το συμπέρασμα που εξάγει, να μεταβιβάζεται στον επόμενο. Οι δομές αυτές διατηρούν ένα μηχανισμό εσωτερικής μνήμης.

Τα αναδρομικά νευρωνικά δίκτυα περιλαμβάνουν επαναληπτικούς (αναδρομικούς) βρόγχους, που επιτρέπουν την διατήρηση αλλά και τη διάδοση προηγούμενης γνώσης. Η δομή ενός RNN απεικονίζεται ακολούθως. Έστω μια ακολουθία τιμών $[x_1, x_2, ..., x_n]$. Στο σχήμα 3.19 αριστερά, ο συγκεκριμένος κόμβος με όνομα A, δέχεται κάποια είσοδο x_t, και η έξοδος (κάθε χρονική στιγμή) είναι h_t. Ο βρόχος ανάδρασης χρησιμεύει για την μεταβίβαση πληροφορίας απο μία κατάσταση την χρονική στιγμή $t-1$, στην κατάσταση την χρονική στιγμή t. Η δομή αυτή μπορεί να γίνει περισσότερο κατανοητή αν ξεδιπλωθεί, όπως φαίνεται στο σχήμα 3.19 δεξιά.



Σχήμα 3.19: Δομή RNN στην κανονική μορφή (αριστερά) και κατόπιν ξεδίπλωσης (δεξιά)

Εξειδικεύοντας στο εσωτερικό του κόμβου A συνδυάζονται η τρέχουσα είσοδος με την προηγούμενη κατάσταση και δίνουν την έξοδο. Αρχικά ορίζουμε μια βοηθητική μεταβλητή $a(t)$ που δεν φαίνεται στο σχήμα. Οι μεταβλητές των εξισώσεων επεξηγούνται στην συνεχεια:

$$
\begin{aligned}
a^{(t)} &= b + W_h^{(t-1)} + Ux^{(t)} \\
h^{(t)} &= f_1(a^{(t)}) \\
o^{(t)} &= c + Vh^{(t)}
\end{aligned}
$$

- Είσοδος $x(t)$: Αν και τα δεδομένα αναλύονται σειριακά, η είσοδος ενός RNN θα πρέπει να έχει σταθερές διαστάσεις. Για το λόγο αυτό πραγματοποιείται κάποιου είδους κωδικοποίηση. Έτσι για παράδειγμα αν ένα δείγμα ξ του συνόλου δεδομένου διαθέτει 10 τιμες $[v_0, v_1, \ldots, v_9]$ και είναι επιθυμητή η ακολουθιακή επεξεργασία, τότε οι είσοδοι την $t = 0$ θα είναι $[v_0, 0, \ldots, 0]$, και αντίστοιχα $[0, v_1, \ldots, 0]$ την χρονική στιγμή $t = 1$.

- Κρυφή Κατάσταση $h(t)$: Η κρυφή κατάσταση λειτουργεί σαν την 'μνήμη' του δικτύου. Ο υπολογισμός της βασίζεται τόσο στη τρέχουσα είσοδο και στην προηγούμενη κατάσταση. Στο άθροισμα αυτών των τιμών, σταθμισμένων κατάλληλα με τα βάρη, και του bias b επενεργεί μια συνάρτηση ενεργοποίησης f1 (πχ $tanh$)

- Έξοδος $o(t)$: Η κρυφή κατάσταση h σταθμίζεται απο τον πίνακα V, και αφού προστεθεί το bias c δίνει την της έξοδο $o(t)$.

- Βαρη (U, V, W): Η είσοδος προς την κρυφή κατάσταση παραμετροποίεται απο έναν πίνακα U, οι σχέσεις μεταξύ των κρυφών καταστάσεων απο έναν πίνακα W και η σύνδεση κρυφής κατάστασης και εξόδου από ένα πίνακα V. Να σημειωθεί ότι και σε αυτή την περίπτωση έχουμε διαμοιρασμό παραμέτρων δηλαδή οι τιμές των πινάκων δεν αλλάζουν κατά το forward pass κάθε χρονική στιγμή t.

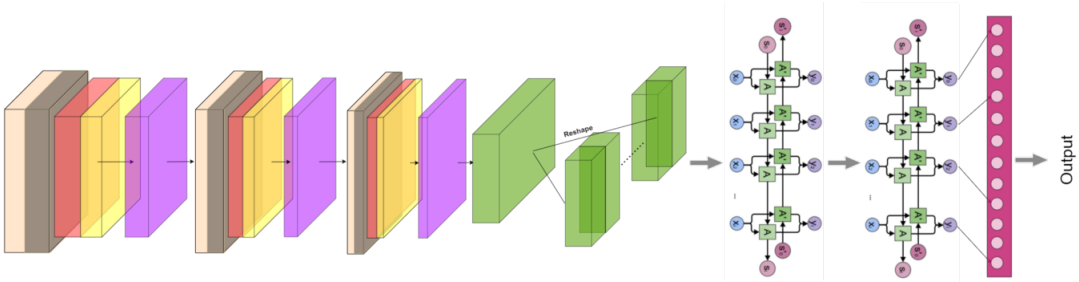
Εκπαίδευση Αναδρομικών Δικτύων:
Η διαδικασία εκπαίδευσης ενός RNN, περιλαμβάνει των προσδιορισμό των τιμών των πινάκων U,V,W αλλά και των bias b,c. Κατά τα γνωστά ένα δείγμα εκπαίδευσης τροφοδοτείται στο αναδρομικό δίκτυο και πραγματοποιείται μια πρόβλεψη. Η πρόβλεψη αφορά το διάνυσμα $o(t)$ που έχει δημιουργηθεί στο τέλος του χρόνου. Στην συνέχεια με βάση τη συνάρτηση κόστους που έχει επιλεγεί, υπολογίζεται το σφάλμα και με τη χρήση του backpropagation ανανεώνονται τα βάρη (με βάση τα gradients). Η διαδικασία αυτή επαναλαμβάνεται για κάθε νέο δείγμα εκπαίδευσης (SGD) για μια ομαδα του συνόλου εκπαίδευση (Mini Batch Gradient Descent) ή για όλο το dataset σε μια εποχή (Gradient Descent).

3.7 Συνδυασμός Συνελικτικών - Αναδρομικών δικτύων

Η γενική αρχιτεκτονική του συνδυασμού CNN-RNN αναφέρεται στην εξαγωγή χαρακτηριστικών απο ένα CNN, στην ακολουθιακή επεξεργασία τους απο RNN δίκτυο και στην

τελική επιλογή κλάσης μέσω ενός softmax layer. Μεταξύ RNN και softmax layer, πιθανώς να παρεμβάλλονται πλήρως συνδεδεμένα επίπεδα. Η αρχιτεκτονική αυτή απεικονίζεται στην παρακάτω εικόνα:

Στην παραπάνω εικόνα, χρησιμοποιείται GRU που είναι ένας ειδικός τύπος RNN. Εκτός αυτών μπορεί να χρησιμοποιηθεί κάθε δομή που επιτυγχάνει ανάδραση και ακολουθιακή επεξεργασία της πληροφορίας, όπως RNN, GRU, LSTM:



Σχήμα 3.20: Συνδυασμός συνελικτικού (αριστερά) - αναδρομικού (δεξιά) δικτύου

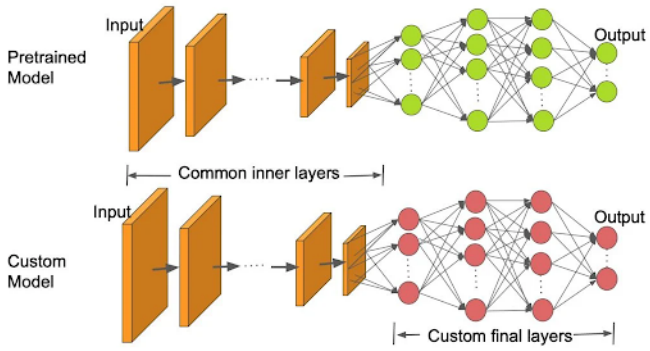
3.8 Μεταφορά Μάθησης

Η Μεταφορά μάθησης είναι ένα ερευνητικό πρόβλημα στο πεδίο της Μηχανικής Μάθησης που εστιάζει στην αποθήκευση γνώσης, η οποία αποκτήθηκε κατά την επίλυση ενός συγκεκριμένου προβλήματος, και στην εφαρμογή της σε ένα διαφορετικό αλλά παρόμοιο πρόβλημα. Για παράδειγμα, η γνώση που αποκτήθηκε κατά την αναγνώριση ποδηλάτων, μπορεί να χρησιμοποιηθεί για την αναγνώριση μοτοσυκλετών.

Τα παρακάτω πλέον, μπορούν να γίνουν κατανοητά, έχοντας θεμελιώσει βασικές έννοιες γύρω απο τα νευρωνικά δίκτυα και γενικότερα τα ευφυή υπολογιστικά συστήματα. Με τον όρο απόκτηση γνώσης αναφερόμαστε στην κατάλληλη ρύθμιση βαρών - bias των επιπέδων ενός μοντέλου, που προκύπτουν με τη διαδικασία της εκπαίδευσης. Με τον όρο 'μεταφορά γνώσης' αναφερόμαστε στην διατήρηση των τιμών αυτών των βαρών (κυρίως των αρχικών επιπέδων) και στην σύνθεση τους με ορισμένα άλλα επίπεδα, τα βάρη των οποίων θα εκπαιδευτούν εξ αρχής.

Η Μεταφορά μάθησης είναι ιδιαίτερα χρήσιμη όταν έχουμε στη διάθεση μας λιγα δεδομένα εκπαίδευσης για την επίλυση ενός προβλήματος. Όπως έχει αναφερθεί, τα πρώτα επίπεδα ενός δικτύου εξάχουν χαρακτηριστικά χαμηλού επιπέδου, όπως ακμές. Στα επόμενα επίπεδα, τα χαρακτηριστικά αυτά συντίθεται δημιουργώντας πιο σύνθετες αναπαραστάσεις. Για το λόγο, τα αρχικά επίπεδα μπορούν να χρησιμοποιηθούν για κάποιο άλλο (σχετικό) πρόβλημα. Ωστόσο τα τελικά επίπεδα, που είναι πιο εξειδικευμένα πάνω στο εκάστοτε πρόβλημα, θα πρέπει να εκπαιδευτούν εκ νέου.

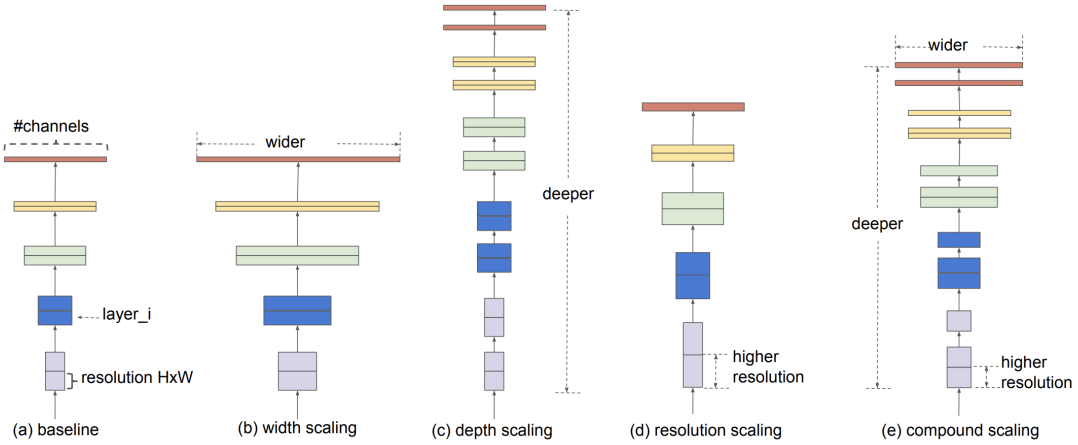
Στην συνέχεια περιγράφονται ορισμένες, εν γένει πολυεπίπεδες, αρχιτεκτονικές που εκπαιδεύτηκαν με τα χαρακτηριστικά του παρόντος συνόλου δεδομένων.



Σχήμα 3.21: Διαδικασία Μεταφοράς Γνώσης απο ένα προεκπαιδευμένο μοντέλο

3.8.1 EfficientNet

Τα συνελικτικά νευρωνικά δίκτυα αναπτύσσονται συνήθως κάτω απο ορισμένες απαιτήσεις τόσο σε πόρους και υπολογιστική ισχύ. Στην συνέχεια μπορούν να κλιμακωθούν, αν περισσότεροι πόροι είναι διαθέσιμοι. Για παράδειγμα, η επέκταση μιας αρχιτεκτονικής κατα *βάθος*, μπορεί να διακρίνει περισσότερα χαρακτηριστικά και ίσως πιο σημαντικά. Ωστόσο για αρκετά βαθιά μοντέλα υπάρχει ο κίνδυνος vanishing gradients, όπως επεξηγήθηκε στην Παράγραφο (3.4.9). Αντίστοιχα, αναφορικά με το *πλάτος*, όσο μεγαλύτερο είναι τόσο πιο εύκολη είναι η εκπαίδευση και τόσο πιο επιτυχής η διάκριση χαρακτηριστικών . Ωστόσο για μεγάλα πλάτη αντιμετωπίζουν δυσκολίες στην εξαγωγή χαρακτηριστικών υψηλού επιπέδου. Τέλος αναφορικά με την *ανάλυση* (resolution) της εικόνας, όσο μεγάλη είναι η ανάλυση, τόσο πιο εύκολα εντοπίζονται πρότυπα. Βέβαια για άψογη ανάλυση ελλοχεύει ο κίνδυνος υπερεκπαίδευσης σε στιγμιότυπα που πιθανώς να εντοπιστούν.

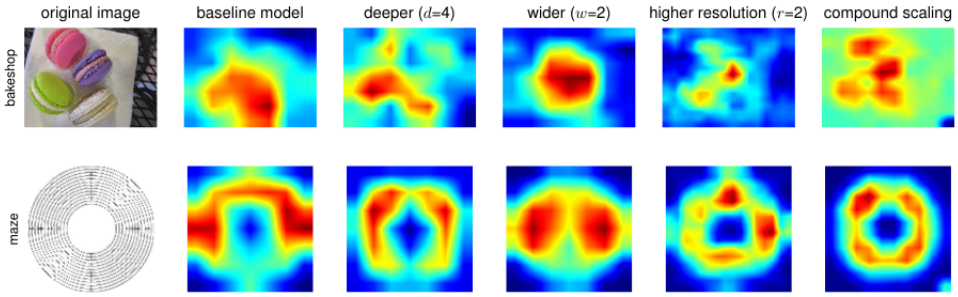


Σχήμα 3.22: Επέκταση Νευρωνικού δικτύου (a), ως προς το πλάτος (b), το μήκος (c), την ανάλυση (d) και βέλτιστα ως προς όλες (d)

Για όλους τους παραπάνω λόγους, η κλιμάκωση αυτή πρέπει να γίνεται ισορροπημένα ως προς το βάθος, το πλάτος του δικτύου καθώς και αναφορικά με την ανάλυση της εικόνας

εισόδου. Με βάση αυτή την παρατήρηση, το δίκτυο EfficientNet προτείνει μια μέθοδο ομοιόμορφης κλιμάκωσης (προς όλες τις κατευθύνσεις) χρησιμοποιώντας έναν αποτελεσματικό συντελεστή (compound coefficient). Η μέθοδος αυτή εφαρμόζεται σε αρχιτεκτονικές τύπου ResNet ή MobileNet (βλ. Παρακάτω 3.8.2).

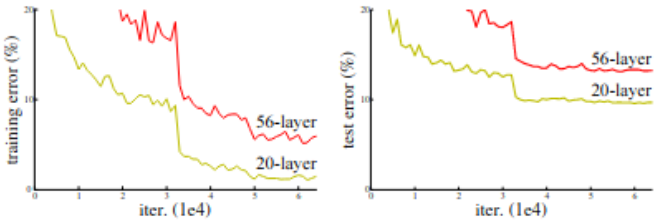
Στη παρακάτω εικόνα παρουσιάζεται η επίδραση της κλιμάκωσης με βάση τον συγκεκριμένο συντελεστή σε δύο διαφορετικές εικόνες. Οι εικόνες αυτές είναι στην ουσία image maps που παρουσιάζουν σε ποιές περιοχές εστίασε το συνελικτικό δίκτυο για να διακρίνει την κατηγορία που ανήκει η εικόνα εισόδου. Είναι προφανές ότι στην τελευταία περίπτωση περικλείεται περισσότερη πληροφορία και έτσι το νευρωνικό μπορεί να εστιάσει σε πιο σχετικές περιοχές [12].



Σχήμα 3.23: Επίδραση επέκτασης δικτύου σε δύο εικόνες

3.8.2 ResNet

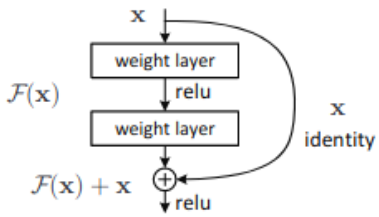
Η ιδέα της αρχιτεκτονικής ResNet, προέκυψε από την παρατήρηση ότι βαθύτερες αρχιτεκτονικές είναι πιο δύσκολο να εκπαιδευτούν. Το πρόβλημα αυτό είναι απόρροια των vanishing gradients που αναφέρθηκε και στην Παράγραφο 3.4.9. Έτσι λοιπόν, είναι αρκετά πιθανό πολυπεπίδες αρχιτεκτονικές να έχουν υψηλότερο training-test error σε σύγκριση με μικρότερες αρχιτεκτονικές. Κάτι τέτοιο φαίνεται και στην παρακάτω εικόνα. Η εκπαίδευση έχει πραγματοποιηθεί στο CIFAR-10.



Σχήμα 3.24: Σύκριση training-test error του δικτύου ResNet αναφορικά με το βάθος ενός δίκτύου

Σε μια αρχιτεκτονική ResNet, παρεμβάλλονται διασυνδέσεις (shortcut connections) μεταξύ επιπέδων, τα οποία δεν είναι κατά ανάγκη γειτονικά.. Έτσι λοιπόν μπορεί να αντιμετωπιστεί το πρόβλημα των vanishing gradients, καθώς κατά τη διαδικασία του back-propagation, η

τιμή του gradient μπορεί να διαδοθεί μέσω των shortcut connections. Έτσι είναι λιγότερο πιθανό να συρρικνωθεί στην τιμή μηδέν. Η δομή μιας μονάδας της αρχιτεκτονικής ResNet απεικονίζεται ακολούθως:



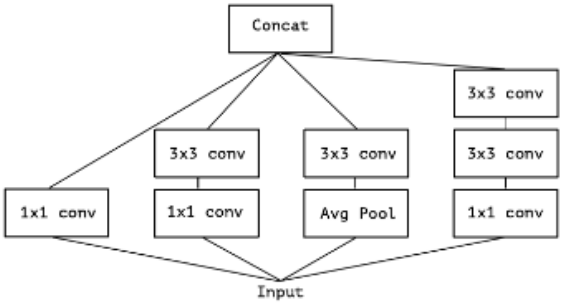
Σχήμα 3.25: Δομικό στοιχείο ResNet

Τα τμήματα αυτά συνδέονται και δημιουργούν την αρχιτεκτονική ResNet [11].

3.8.3 Xception

Η αρχιτεκτονική Xception, όπως υποδηλώνει και το όνομα της: "Extreme Inception", είναι εμπνευσμένη απο την αρχιτεκτονική Inception. Στην αρχιτεκτονική αυτή, τμήματα όπως αυτό που απεικονίζεται ακολούθως, διαδέχεται το ένα το άλλο.

Figure 1. A canonical Inception module (Inception V3).



Σχήμα 3.26: Δομικό στοιχείο Xception

Στην συγκεκριμένη περίπτωση, πρώτα εφαρμόζεται η συνέλιξη $1x1$ και μετά επέρχεται διαχωρισμός στις συνελίξεις με πυρήνες $3x3$. Η πράξη αυτή είναι γνωστή με το όνομα depthwise separable convolutions. Η αρχιτεκτονική Xception λοιπόν δεν είναι τίποτα άλλο από τμήματα που επιτελούν ξεχωριστές- ανεξάρτητες συνελίξεις, συνδεδεμένα μεταξύ τους τόσο γραμμικά όσο και με συνδέσεις που χρησιμοποιούνται στα ResNets (βλ Παράγραφο 3.8.2). [13]

Κεφάλαιο 4

Τεχνικές Λεπτομέρειες - Εργαλεία

Για την υλοποίηση της παρούσας διπλωματικής εργασίας έγινε χρήση της γλώσσας προγραμματισμού Python στο προγραμματιστικό περιβάλλον Google Colaboratory. Οι βασικές βιβλιοθήκες που συμπεριλήφθηκαν συνοψίζονται ακολούθως:

- numpy: Θεμελιώδες πακέτο για την εκτέλεση υπολογισμών, κυρίως σε πίνακες.
- matplotlib: Περιεκτική βιβλιοθήκη για την οπτικοποίηση των αποτελεσμάτων.
- sklearn: Βιβλιοθήκη για εφαρμογές data analysis machine learning (clustering, classification, regression κλπ).
- sktime: Βιβλιοθήκη για ανάλυση χρονοσειρών και εφαρμογή αλγορίθμων μηχανικής μάθησης σε δεδομένα που μεταβάλλονται με το χρόνο.
- tensorflow: Βιβλιοθήκη για εφαρμογές μηχανικής μάθησης, με εστίαση στην εκπαίδευση νευρωνικών δικτύων.
- keras: Υψηλού επιπέδου διεπαφή προγραμματισμού εφαρμογών (API) της πλατφόρμας tensorflow, για ανάπτυξη μοντέλων μηχανικής μάθησης.

Η επιτάχυνση της εκτέλεσης αριθμητικών πράξεων, και γενικότερα η εκπαίδευση των μοντέλων, έγινε εφικτή με τη χρήση κάρτας γραφικών GPU. Το περιβάλλον Google Colaboratory παρέχει διαφορετικούς τύπους GPU κάθε φορά, χωρίς ωστόσο να δίνεται η δυνατότητα επιλογής. Οι διαθέσιμες κάρτες γραφικών είναι: Nvidia K80s, T4s, P4s και P100s.

Κεφάλαιο 5

Περιγραφή Αρχιτεκτονικής

5.1 Καθορισμός Προβλήματος- Στόχου

Στόχο της παρούσας διπλωματικής εργασίας, αποτελεί η αναγνώριση βασικών φωνητικών εντολών με σκοπό την κατεύθυνση / οδήγηση ενός ρομποτικού οχήματος στο χώρο. Η εφαρμογή αυτή μπορεί βέβαια να γενικευθεί για τον έλεγχο οποιασδήποτε διάταξης η οποία έχει την ικανότητα κίνησης. Πιο ειδικά το σύστημα που θα σχεδιαστεί και θα αναπτυχθεί, θα μπορεί να αναγνωρίζει εντολές κίνησης ('stop', 'go'), κατεύθυνσης ('right', 'left', 'up', 'down'), αριθμούς ('one', 'two', 'three', 'four') καθώς και έναρξης - τερματισμού κάποιας διεργασίας ('on'-'off', 'yes' - 'no').

5.2 Επισκόπηση papers - Συγγενείς Εργασίες

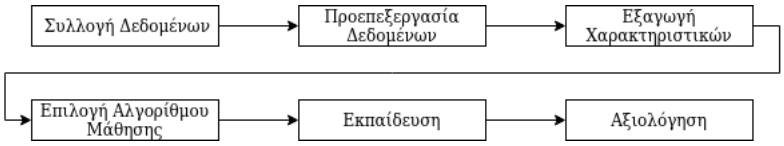
Όπως έγινε εμφανές και απο την Ιστορική Αναδρομή (βλ. Παράγραφο 1.1), η αναγνώριση φωνής απο ένα υπολογιστικό σύστημα, απασχόλησε απο νωρίς την επιστημονική κοινότητα. Μέχρι σήμερα, έχουν πραγματοποιηθεί αρκετές εργασίες χρησιμοποιώντας διαφορετικά εργαλεία αλλά και τεχνολογίες [7], [8], [9], [10]. Η διαφοροποίηση έγκειται τόσο στην εξαγωγή χαρακτηριστικών, όσο και στη χρήση αλγορίθμων μηχανικής μάθησης. Παράλληλα έχουν εξεταστεί και διάφορες αρχιτεκτονικές στα επιμέρους τεχνητά νευρωνικά δίκτυα. Παρακάτω εξηγούνται ορισμένες συγγενείς εργασίες με απώτερο σκοπό την αναγνώριση φωνητικών εντολών στο συγκεκριμένο dataset, με το οποίο εργαζόμαστε στην παρούσα εργασία.

Αρκετά σύγχρονα συστήματα βασίζονται στη χρήση Κρυφών Μαρκοβιανών Μοντέλων (HMMs). Πρόκειται για στατιστικά μοντέλα που περιλαμβάνουν καταστάσεις (states) με ορισμένες πιθανότητες μετάβασης μεταξύ αυτών. Οι πιθανότητες αυτές ορίζονται είτε για την μετάβαση στην επόμενη, είτε για την παραμονή στην τρέχουσα κατάσταση. Το ακουστικό σήμα αναλύεται σε μικρά χρονικά διαστήματα, ώστε σε κάθε διάστημα το σήμα να θεωρείται μια στάσιμη διαδικασία (stationary process). Κάθε κατάσταση αυτού του μοντέλου λοιπόν αντιπροσωπεύει ένα τμήμα του ακουστικού σήματος. Αναφορικά με το dataset, με το οποίο εργαζόμαστε (βλ. Παράγραφο 5.4) η χρήση HMM - GMM σημειώσε ποσοστά της τάξης του 65 %. [14]

Σε ποιο επίκαιρες εργασίες χρησιμοποιούνται μεταξύ άλλων βαθιά νευρωνικά δίκτυα. Οι πολυεπίπεδες αυτές αρχιτεκτονικές, επιτρέπουν τόσο την εκμάθηση χαρακτηριστικών όσο και τη σύνθεση τους, για την επιτυχή αναγνώριση φωνητικών εντολών. Η εξαγωγή χαρακτηριστικών που αποτυπώνονται σε εικόνες (όπως φασματογραφήματα), άνοιξε το δρόμο για τη χρήση συνελικτικών δικτύων CNN. Στην δεύτερη περίπτωση τα ποσοστά ακρίβειας ανέρχονται στο 90.7 % [15]. Ωστόσο η ανάγκη ακολουθιακής επεξεργασίας της ακουστικής πληροφορίας, έφερε στην επιφάνεια αρχιτεκτονικές που χρησιμοποιούν ανατροφοδότηση εισόδου - εξόδου. Τέτοιες αρχιτεκτονικές είναι για παράδειγμα τα αναδρομικά δίκτυα RNN. Ιδιαίτερο ενδιαφέρον παρουσιάζουν υλοποιήσεις που συνδυάζουν τις δύο αρχιτεκτονικές CNN - RNN, και έτσι αξιοποιούν στο έπακρο τα πλεονεκτήματα τους. Συνήθως οι εικόνες αναλύονται αρχικά με τη χρήση CNN και η πληροφορία απο τα τελικά στρώματα μεταβιβάζεται σε αναδρομικά δίκτυα. Η δομή αυτή επιτυγχάνει για το συγκεκριμένο dataset ποσοστά ακριβείας της τάξης του 87.8 % [15].

5.3 Δομή Παρούσας Εργασίας

Απο τη στιγμή που καθορίστηκε ο στόχος της παρούσας διπλωματικής, η διαδικασία που θα ακολουθηθεί για την επίλυση του προβλήματος, περιγράφεται στο ακόλουθο διάγραμμα ροής. Πρόκειται για την τυπική διαδικασία που χρησιμοποιείται σε κάθε πρόβλημα μηχανικής μάθησης. Κάθε ένα απο τα επιμέρους έξι στάδια πρόκειται να αναλυθεί με λεπτομέρεια ακολούθως:



Σχήμα 5.1: Διαχωρισμός και Οργάνωση Εργασίας σε 6 στάδια

5.4 Συλλογή δεδομένων

Τα δεδομένα που χρησιμοποιούνται, προέρχονται από το Google Speech Commands Dataset (συντ. GSC) [17]. Το συγκεκριμένο σύνολο, στην τελική του διαμόρφωση, περιλαμβάνει πάνω απο 100.000 ηχογραφημένα ακουστικά μηνύματα που αφορούν 35 λέξεις, διάρκειας 1 sec (ή λιγότερο). Η ηχογράφηση των λέξεων πραγματοποιήθηκε με ρυθμό δειγματοληψίας 16kHz, ενώ τα δεδομένα που προέκυψαν απο την δειγματοληψία, έχουν κωδικοποιηθεί χρησιμοποιώντας 16 bit PCM. Οι λέξεις αυτές είναι τα ψηφία 'zero' - 'nine', καθώς και εντολές όπως 'Yes', 'No', 'Up', 'Down', 'Left', 'Right', 'On', 'Off', 'Stop', 'Go', που μπορούν να χρησιμοποιηθούν σε αρκετές εφαρμογές IoT ή ρομποτικής. Πρόκειται λοιπόν για την πρώτη έκδοση του GSC, στην οποία περιλαμβάνονται οι πρώτες 20 εντολές. Σε ανανεωμένη έκδοση προστέθηκαν άλλες 10 εντολές, οι οποίες μπορούν να χρησιμοποιηθούν για έλεγχο της απόδοσης των μοντέλων. Οι λέξεις αυτές είναι: 'Bed', 'Bird', 'Cat', 'Dog', 'Happy', 'House',

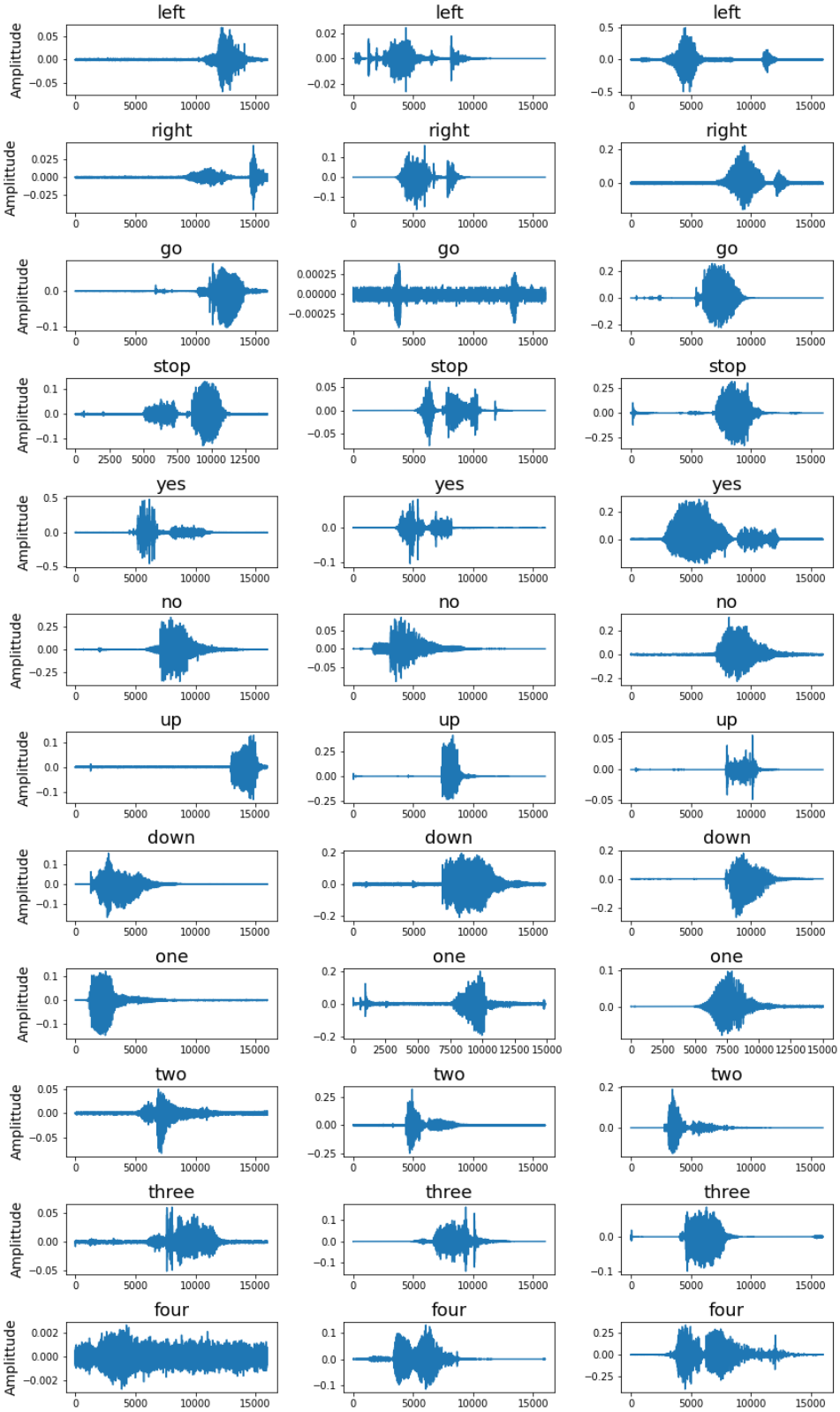
Εντολή	# recordings	Ποσοστό %
go	2372	8.3
stop	2380	8.4
right	2367	8.3
left	2353	8.3
yes	2377	8.4
no	2375	8.3
up	2375	8.4
down	2359	8.3
one	2370	8.3
two	2373	8.4
three	2356	8.3
four	2372	8.3

Πίνακας 5.1: Πλήθος recordings για κάθε εντολή

'Marvin', 'Sheila', 'Tree', 'Wow', 'Learn', 'Follow', 'Forward', 'Backward', 'Visual'. Ως παράδειγμα η λέξη 'Tree' που ακούγεται όπως η λέξη 'Three' ενώ οι άλλες λέξεις προστέθηκαν για να καλύψουν ενα ευρύ φάσμα φωνημάτων. Οι ανάγκες της συγκεκριμένης εργασίας, μπορούσαν να καλυφθούν απο την πρώτη έκδοση (v1) του GSC. Πιο ειδικά απο τις 20 λέξεις, επιλέχθηκαν 12 οι οποίες φαίνονται στον Πίνακα 5.1. Είναι αξιοσημείωτος επίσης, ο τρόπος με τον οποίο δημιουργήθηκε το συγκεκριμένο dataset. Όλα τα μηνύματα συγκεντρώθηκαν απο απλούς χρήστες μέσω μιας εφαρμογής που μπορούσαν να εγκαταστήσουν είτε στο κινητό (σε συσκευές Android) είτε διαδικτυακά (σε web browser). Με τον τρόπο αυτό τα ηχογραφημένα μηνύματα απείχαν αρκετά απο την ποιότητα ηχογραφήσεων με υψηλής ποιότητας μικρόφωνα, που πραγματοποιούνται σε ένα studio κάτω απο αυστηρές συνθήκες. Ωστόσο τα επιτυχημένα μοντέλα αναγνώρισης φωνής, οφείλουν να αντιμετωπίζουν θορυβώδη περιβάλλοντα και ανθρώπους που μιλούν με φυσικό τρόπο.

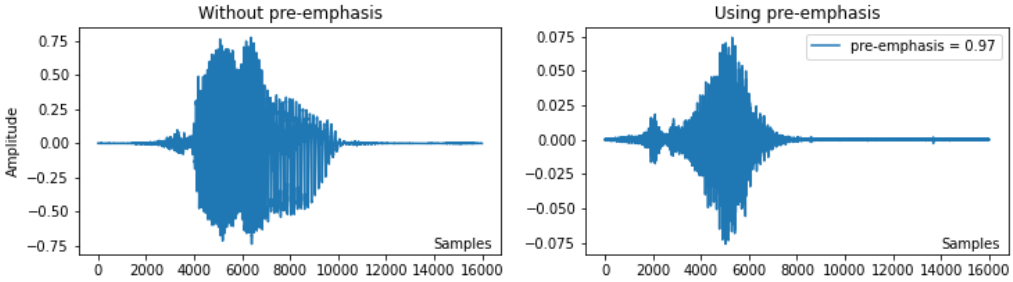
5.5 Προεπεξεργασία Δεδομένων στο πεδίο του Χρόνου

Στην συνέχεια (Εικόνα 5.2) απεικονίζονται 3 κυματομορφές (οριζόντια) στο πεδίο του χρόνου για κάθε μια απο τις 12 εντολές (κάθετα). Όπως είναι αναμενόμενο, δεν είναι δυνατόν να αναδειχτούν συσχετίσεις μεταξύ των κυματομορφών που αναφέρονται στην ίδια εντολή, αλλά και ούτε διαφοροποιήσεις μεταξύ διαφορετικών εντολών. Μετά την συγκέντρωση των δεδομένων σειρά έχει η προεπεξεργασία τους, ώστε να διαμορφωθούν κατάλληλα πριν την εξαγωγή των χαρακτηριστικών. Πρώτον εφαρμόζεται pre-emphasis (σύμφωνα με ό,τι έχει αναφερθεί στην Παράγραφο 2.2.3. Τα αποτελέσματα εφαρμογής του φίλτρου, στο πεδίο του χρόνου, φαίνονται στην Εικόνα 5.3, ενώ αναφορικά με το πεδίο της συχνότητας παρουσιάζονται στην Εικόνα 5.4. Αναφορικά με το πεδίο της συχνότητας, παρατηρεί κανείς ότι ενισχύθηκαν οι υψηλές συχνότητες, όπως ήταν αναμενόμενο:

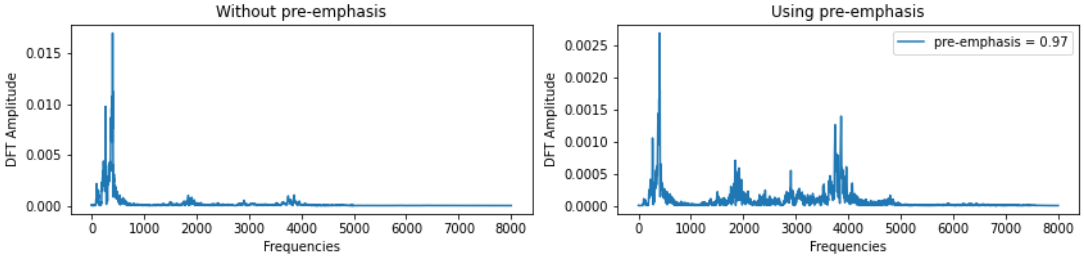


Σχήμα 5.2: Παρουσιάση τριών κυματομορφών στο Πεδίο του χρόνου για όλες τις εντολές

Στο σημείο αυτό να σημειωθεί ότι θα μπορούσαμε στα ακουστικά σήματα να προσθέσουμε περεταίρω θόρυβο. Με τον τρόπο αυτό, ενδέχεται να μειωθεί το σφάλμα γενίκευσης και να βελτιωθεί η δομή του προβλήματος αντιστοίχισης. Αυτό συμβαίνει καθώς το μοντέλο εκπαίδευσης, θα έχει μικρότερη ικανότητα μάθησης πάνω στο σύνολο εκπαίδευσης με αποτέλεσμα να αυξάνεται η απόδοση στη γενίκευση. Ωστόσο, όπως αναφέρθηκε και στην παράγραφο Συλλογή Δεδομένων 5.4, οι συνθήκες κάτω απο τις οποίες πραγματοποιήθηκαν οι ηχογραφήσεις, επέφεραν κάποια αλλοίωση στο τελικό αποτέλεσμα. Για το λόγο αυτό δεν κρίνεται αναγκαία η προσθήκη θορύβου.



Σχήμα 5.3: Επίδραση εφαρμογής pre-emphasis στο πεδίο του χρόνου



Σχήμα 5.4: Επίδραση εφαρμογής pre-emphasis στο πεδίο της συχνότητας

5.6 Εξαγωγή Χαρακτηριστικών

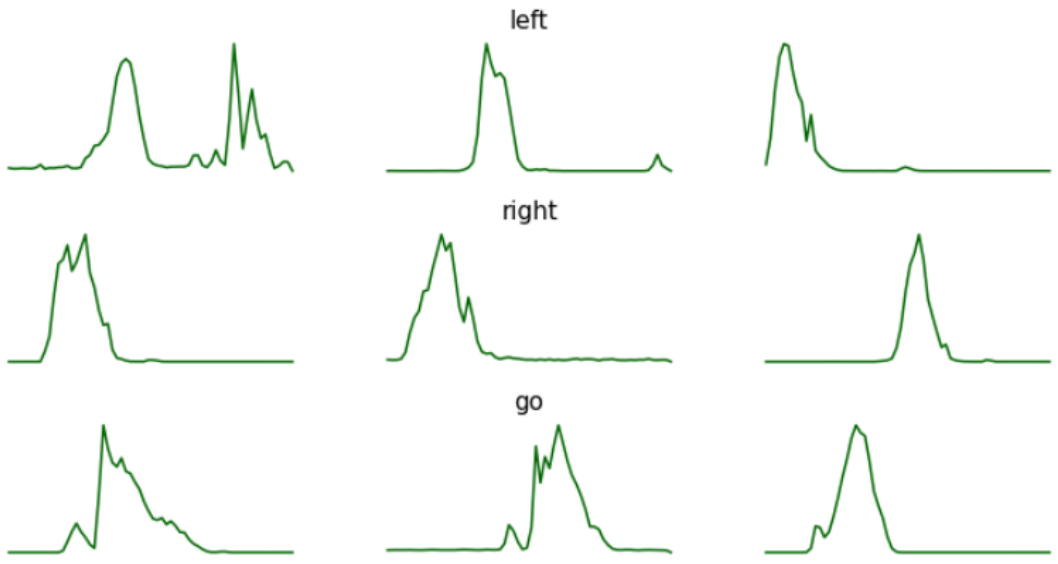
5.6.1 Πεδίο του Χρόνου

RMS

Μετά τη δειγματοληψία του σήματος, υπολογίζονται σε κυλιόμενα παράθυρα οι RMS τιμές σύμφωνα με την θεωρία που έχει περιγραφεί στην Παράγραφο 2.2.2. Πλέον τα ακουστικά σήματα, θα μελετηθούν σαν χρονοσειρές, δείγμα των οποίων παρουσιάζεται ακολούθως:

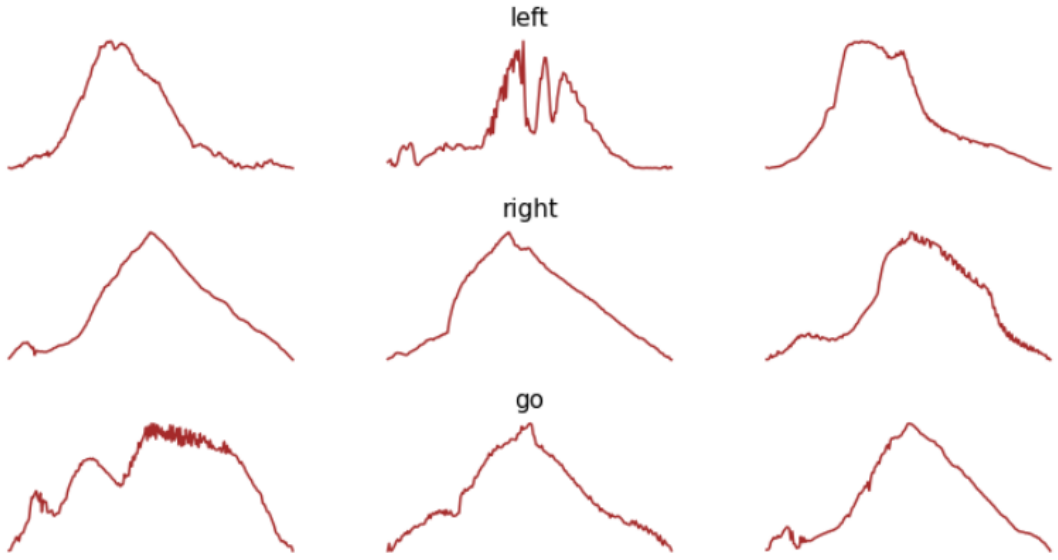
ZCR

Μετά τη δειγματοληψία του σήματος, υπολογίζονται οι ZCR τιμές σύμφωνα με την θεωρία που έχει περιγραφεί στην Παράγραφο 2.2.2. Πλέον τα ακουστικά σήματα, θα μελετηθούν σαν



Σχήμα 5.5: Εξαγωγή RMS τιμών στο πεδίο του χρόνου

χρονοσειρές, δείγμα των οποίων παρουσιάζεται ακολούθως:

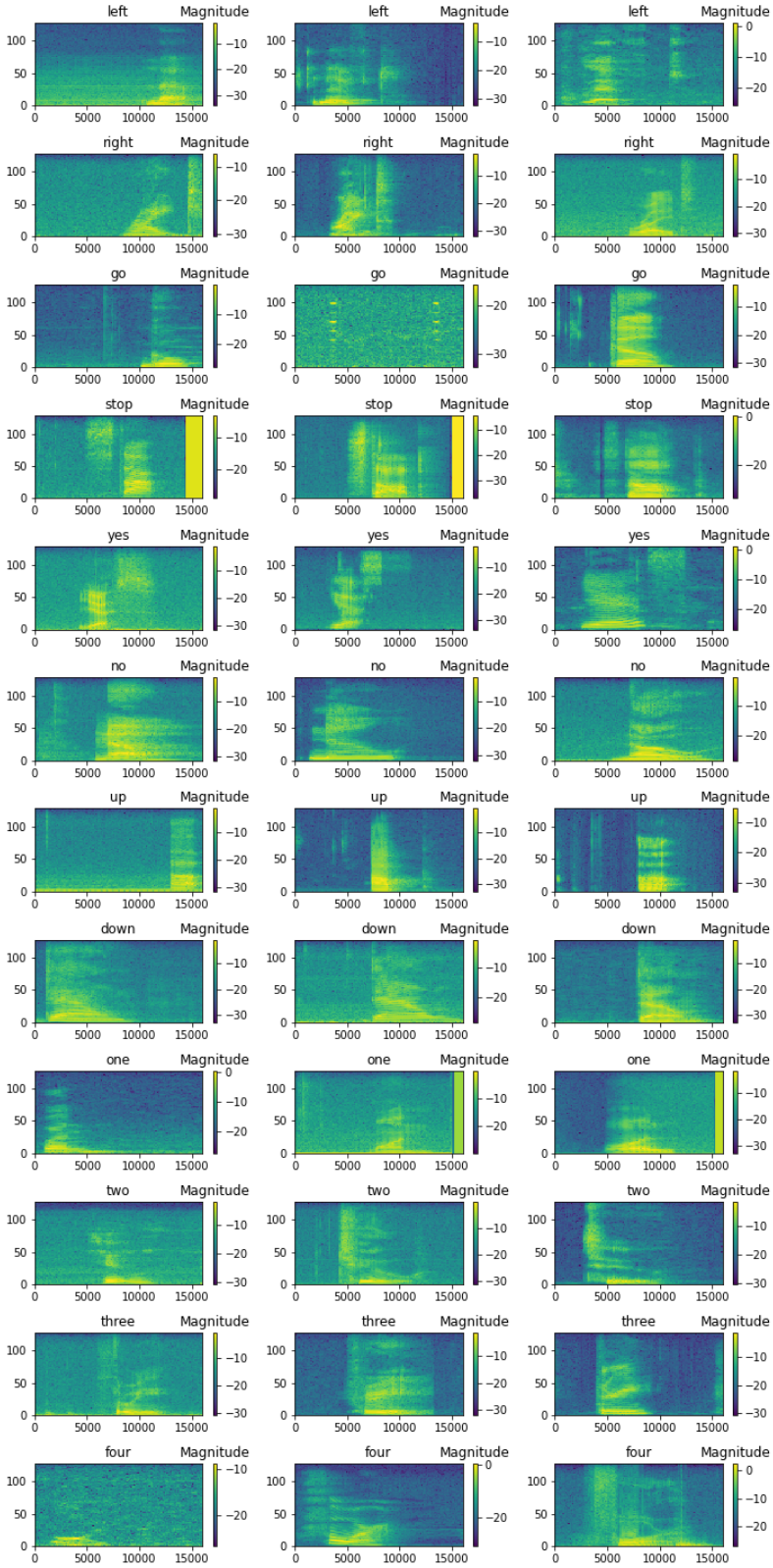


Σχήμα 5.6: Εξαγωγή ZCR τιμών στο πεδίο του χρόνου

5.6.2 Πεδίο της Συχνότητας

Log Spectrograms

Σύμφωνα με όσα έχουν εξηγηθεί στην Παράγραφο 2.3.2, εξάγονται τα log spectrograms για κάθε ακουστικό σήμα. (Βλ . Εικόνα 5.7)



Σχήμα 5.7: Εξαγωγή log spectrograms για τα ακουστικά σήματα των εντολών

MFCCs

Σύμφωνα με όσα έχουν εξηγηθεί στην Παράγραφο 2.3.4, εξάγονται οι συντελεστές mfcc για κάθε ακουστικό σήμα. (Βλ . Εικόνα 5.8)

delta MFCCs

Σύμφωνα με όσα έχουν εξηγηθεί στην Παράγραφο 2.3.4, εξάγονται οι συντελεστές delta mfcc για κάθε ακουστικό σήμα. (Βλ . Εικόνα 5.9)

5.7 Επιλογή Αλγορίθμου Μάθησης - Αρχιτεκτονικής

5.7.1 Πεδίο του Χρόνου

Στο Πεδίο του χρόνου (όπου τα χαρακτηριστικά είναι οι τιμές ZCR / RMS), εξετάζονται τρεις αλγόριθμοι μάθησης, καθένας με διαφορετική λειτουργία:

- **Distance based**: Στην περίπτωση αυτή, ο αλγόριθμος που χρησιμοποιείται είναι ο kNN - DTW, ο οποίος έχει εξηγηθεί στην Παράγραφο 3.3.2.
- **Interval based**: Στην περίπτωση αυτή, ο αλγόριθμος που χρησιμοποιείται είναι ο Random Forest, ο οποίος έχει εξηγηθεί στην Παράγραφο 3.3.2.
- **Frequency based**: Στην περίπτωση αυτή, ο αλγόριθμος που χρησιμοποιείται είναι ο RISE, ο οποίος έχει εξηγηθεί στην Παράγραφο 3.3.2.

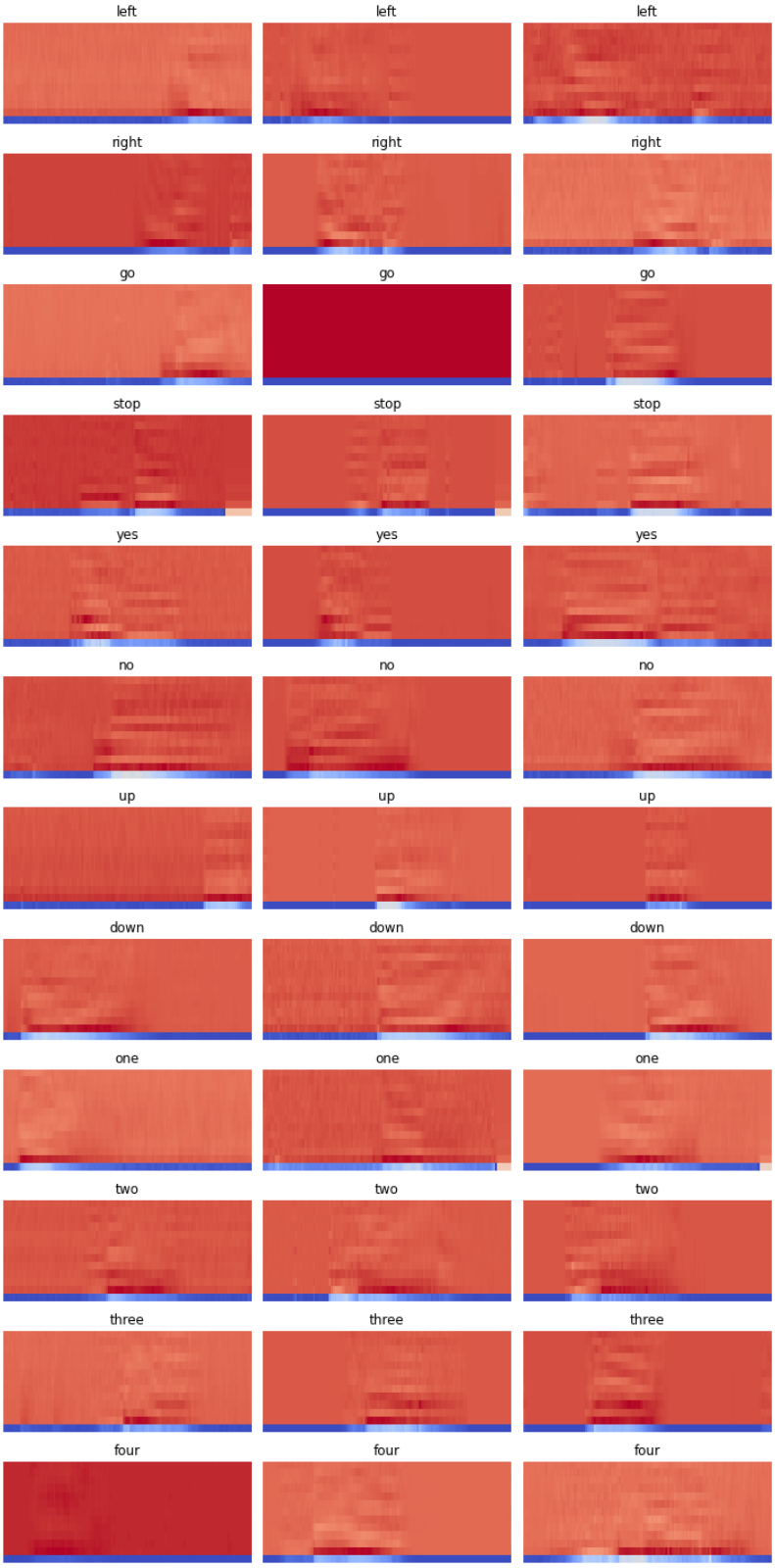
5.7.2 Πεδίο της Συχνότητας

- **CNN**: Στην περίπτωση αυτή, εξετάζεται η αρχιτεκτονική ενός συνελικτικού δικτύου CNN, όπως έχει περιγραφεί στην Παράγραφο 3.5
- **CNN - RNN**: Στην περίπτωση αυτή, εξετάζεται η αρχιτεκτονική ενός συνελικτικού δικτύου, όπως έχει περιγραφεί στην παράγραφο 3.5 ακολουθούμενο απο ένα αναδρομικό δίκτυο RNN, όπως έχει περιγραφεί στην Παράγραφο 3.7
- **Transfer Learning Models**: Στην περίπτωση αυτή, πραγματοποιείται μεταφορά μάθησης, χρησιμοποιώντας ορισμένα προεκπαιδευμένα μοντέλα, όπως έχει περιγραφεί στην Παράγραφο 3.8

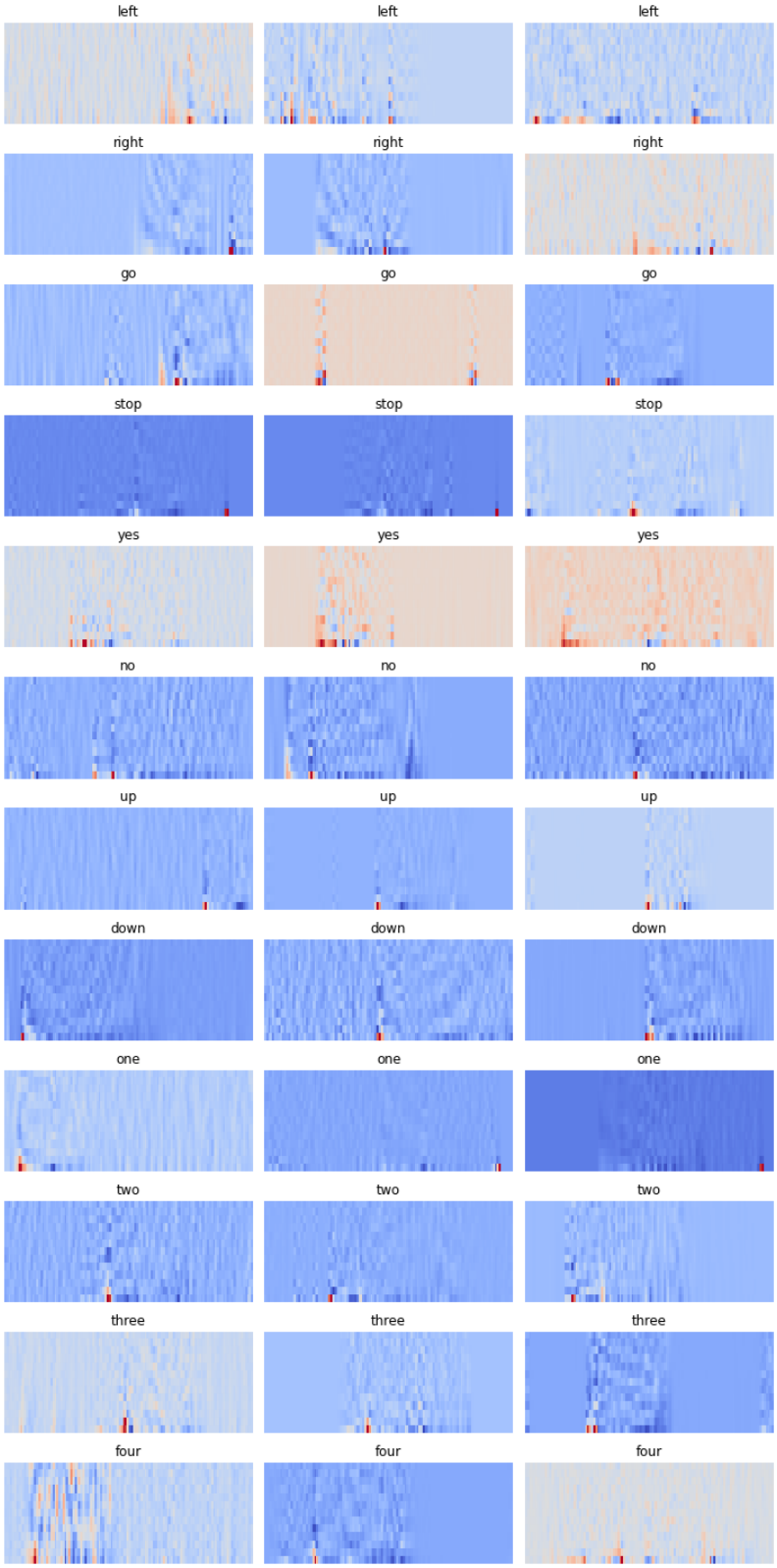
5.8 Εκπαίδευση

5.8.1 Πεδίο του Χρόνου

Στο πεδίο του χρόνου, τα χαρακτηριστικά που χρησιμοποιούνται για την εκπαίδευση αλγορίθμων και αρχιτεκτονικών νευρωνικών δικτύων, είναι οι RMS τιμές των ακουστικών σημάτων.



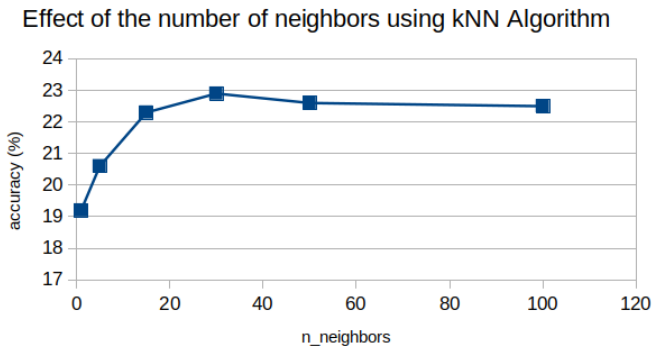
Σχήμα 5.8: Εξαγωγή MFCCs για τα ακουστικά σήματα των εντολών



Σχήμα 5.9: Εξαγωγή delta-MFCCs για τα ακουστικά σήματα των εντολών

Οι αλγόριθμοι που χρησιμοποιήθηκαν μπορούν να διακριθούν σε 3 βασικές κατηγορίες: α) Αλγόριθμος βασισμένος σε Αποστάσεις (Model 1), β) Αλγόριθμος βασισμένος σε Διαστήματα (Model 2), γ) Αλγόριθμος βασισμένος σε συχνότητες (Model 3).

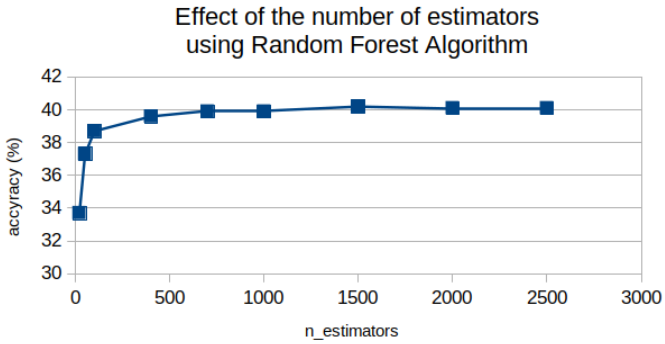
Model 1: Distance based (kNN -DTW) on RMS values



Σχήμα 5.10: Επίδραση του αριθμού των γειτόνων του αλγορίθμου KNN, στο test accuracy

Παρατηρεί κανείς ότι απο ένα σημείο και μετά ($\#neighbors > 40$), το ποσοστό ορθότητας στο σύνολο ελέγχου σταθεροποείται και δεν επιδέχεται βελτίωσης. Το μεγαλύτερο ποσοστό σημειώθηκε για πλήθος γειτόνων ίσο με 30.

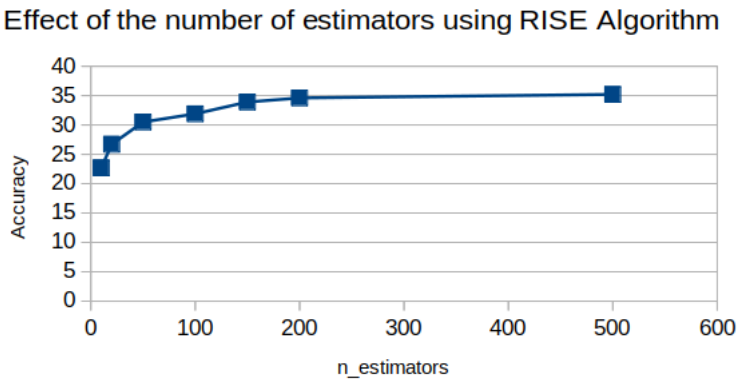
Model 2: Interval based (Random Forest) on RMS values



Σχήμα 5.11: Επίδραση του αριθμού των εκτιμητών του αλγορίθμου Random Forest, στο test accuracy

Παρατηρεί κανείς και σε αυτή την περίπτωση, ότι απο ένα σημείο και μετά ($\#estimators > 500$), το ποσοστό ορθότητας στο σύνολο ελέγχου σταθεροποείται και δεν επιδέχεται βελτίωσης. Το μεγαλύτερο ποσοστό σημειώθηκε για πλήθος εκτιμητών ίσο με 1500 (χωρίς ωστόσο σημαντική διαφορά απο το προηγούμενο πείραμα).

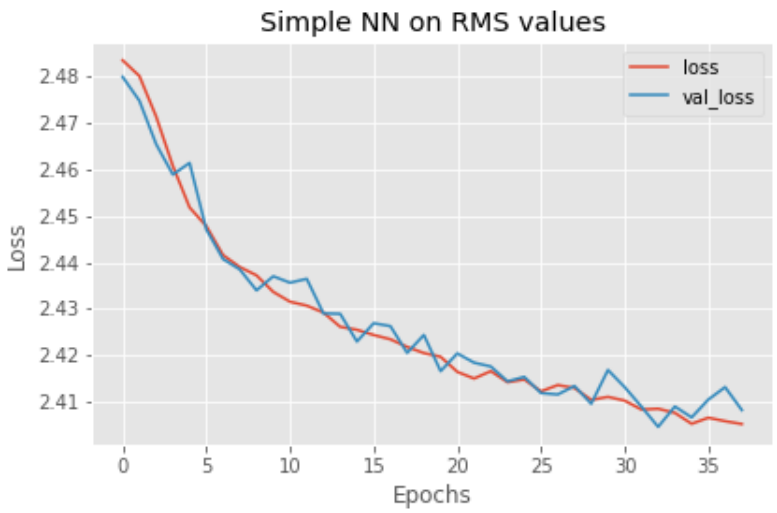
Model 3: Frequency Based (RISE) on RMS values



Σχήμα 5.12: Επίδραση του αριθμού των εκτιμητών του αλγορίθμου RISE, στο test accuracy

Και στην περίπτωση του αλγορίθμου RISE, το ποσοστό σταθεροποιείται σε μια τιμή. Το μεγαλύτερο ποσοστό σημειώθηκε για πλήθος εκτιμητών ίσο με 200.

Model 3: Simple NN model on RMS values



Σχήμα 5.13: Καμπύλη εκπαίδευσης απλού νευρωνικού δικτύου χρησιμοποιώντας τις RMS τιμές των ακουστικών σημάτων

Η αρχιτεκτονική του δικτύου η οποία διαμορφώθηκε ήταν ένα πλήρως συνδεδεμένο δίκτυο τριών επιπέδων. Το πρώτο επίπεδο είχε 64 κόμβους το δεύτερο 48 και το τρίτο και τελευταίο 12 (όσες και οι κλάσεις). Στα δύο πρώτα επίπεδα χρησιμοποιείται συνάρτηση ενεργοποίησης ReLU ενώ στο τελευταίο επίπεδο softmax. Επίσης μετά τα δυο πρώτα επίπεδα, χρησιμοποιείται dropout με τιμή 0.25. Έισοδο αποτέλεσαν οι RMS τιμές. Παρόλο που το σφάλμα στο σύνολο επικύρωσης ακολουθεί το σφάλμα κατά την διάρκεια της εκπαίδευσης, εντούτοις φαίνεται να

σταθεροποείται σε μια τιμή και μάλιστα μεγάλη. Τα συμπεράσματα στο πεδίο του χρόνου, συνοψίζονται στην Παράγραφο της Αξιολόγησης (5.9).

5.8.2 Πεδίο της Συχνότητας

Πειραματισμός με Χαρακτηριστικά

Αρχικά ορίζονται κάποιες απλές αρχιτεκτονικές: (1) Συνελικτικό δίκτυο (2) Συνελικτικό δίκτυο ακολουθούμενο απο Αναδρομικό δίκτυο (3) Μοντέλο, στο οποίο αξιοποιείται η τεχνική της μεταφοράς μάθησης. Σε κάθε μια απο αυτές τις αρχιτεκτονικές, εξετάζεται η επίδραση διαφορετικών χαρακτηριστικών. Η εκπαίδευση πραγματοποιήθηκε αρχικά για μικρό αριθμό εποχών (10). Αυτό πραγματοποιήθηκε για δύο λόγους. Πρώτον, για να εκτιμηθεί απο νωρίς η επίδοση και κατ' επέκταση η συμπεριφορά των επιμέρους μοντέλων και δεύτερον, γιατί για baseline αρχιτεκτονικές και μεγάλο αριθμό εποχών, θα ήταν σχεδόν βέβαιη η υπερεκπαιδευση.

	spectrogram	mfccs	delta mfccs
CNN	77	**87**	62
CNN -RNN	44	**78**	53
ResNet	53	90	**92**

Πίνακας 5.2: Accuracy (%) για Αρχιτεκτονικές- Χαρακτηριστικά, χωρίς βλετιστοποιήση

Με βάση τον παραπάνω πίνακα (Πίνακας 5.3) επιλέγονται και οι συνδυασμοί αρχιτεκτονικών - χαρακτηριστικών οι οποίοι θα βλετιστοποιηθούν στην συνέχεια. Πιο ειδικά για κάθε μοντέλο (CNN, CNN-RNN, Transfer Learning Model), σημειώνεται με bold η επιλογή χαρακτηριστικού που πέτυχε το μεγαλύτερο ποσοστό ορθότητας. Συνοπτικά οι αρχιτεκτονικές που πρόκειται να μελετηθούν είναι:

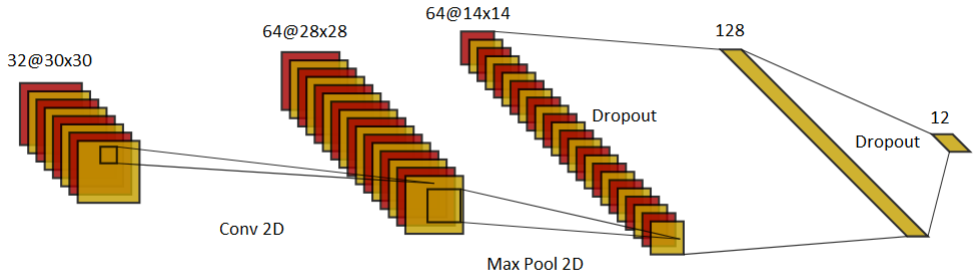
	Architecture	Feature	Acc (%)
Model 1	CNN	MFCCs	87
Model 2	CNN-RNN	MFCCs	78
Model 3	ResNet-50	delta-MFCCs	92

Πίνακας 5.3: Σύνοψη αρχιτεκτονικών (baseline) που πρόκειται να βελτιστοποιηθούν

Το μοντέλο που χρησιμοποιήθηκε αρχικά, οπτικοποιείται στο ακόλουθο σχήμα (Σχήμα 5.14). Να σημειωθεί ότι, προσθέτοντας επίπεδο κανονικοποίησης, η συγκεκριμένη αρχιτεκτονική απέδωσε ποσοστό ορθότητας 87 %.

Model 1: Συνελικτικό Δίκτυο (CNN) εφαρμοσμένο σε χαρακτηριστικά MFCC

Τα επιμέρους βήματα που πραγματοποιήθηκαν για την βελτιστοποίηση της αρχιτεκτονικής οργανώνονται σε τρία στάδια. Στο πρώτο στάδιο (A), πραγματοποείται μια διερεύνηση

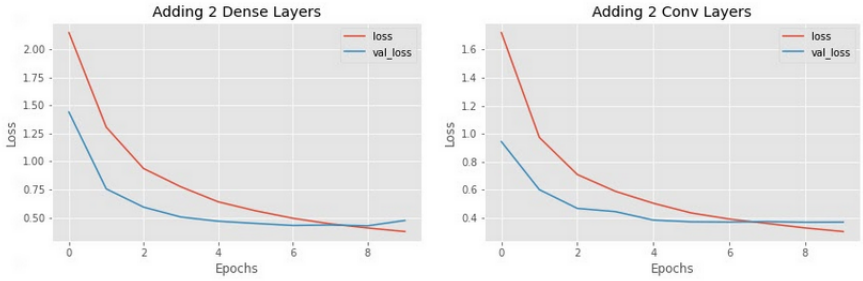


Σχήμα 5.14: Συνελικτικό Δίκτυο (CNN) εφαρμοσμένο σε χαρακτηριστικά MFCC

ως προς την αρχικοποίηση της αρχιτεκτονικής (Βάθος Δικτύου, Μέγεθος Εικόνας Εισόδου, Επίπεδα Pooling, Συναρτήσεις Ενεργοποίησης). Στο δεύτερο στάδιο (Β), εξετάζονται διάφοροι παράμετροι που αφορούν την διαμόρφωση της εκπαίδευσης (Βελτιστοποιητές, learning rate). Το τρίτο στάδιο (Γ), αφορά τον έλεγχο αλλά και την πρόληψη της υπερεκπαίδευσης και εξετάζεται η λειτουργία dropout.

A1 Βάθος Δικτύων

Αρχικά ο πειραματισμός αφορούσε το βάθος του συγκεκριμένου δικτύου. Πιο ειδικά δοκιμάστηκε τόσο η επέκταση των συνελικτικών επιπέδων (Εικόνα 5.15, αριστερά), όσο και των πλήρων συνδεδεμένων δικτύων στο τέλος (Εικόνα 5.15, δεξιά).

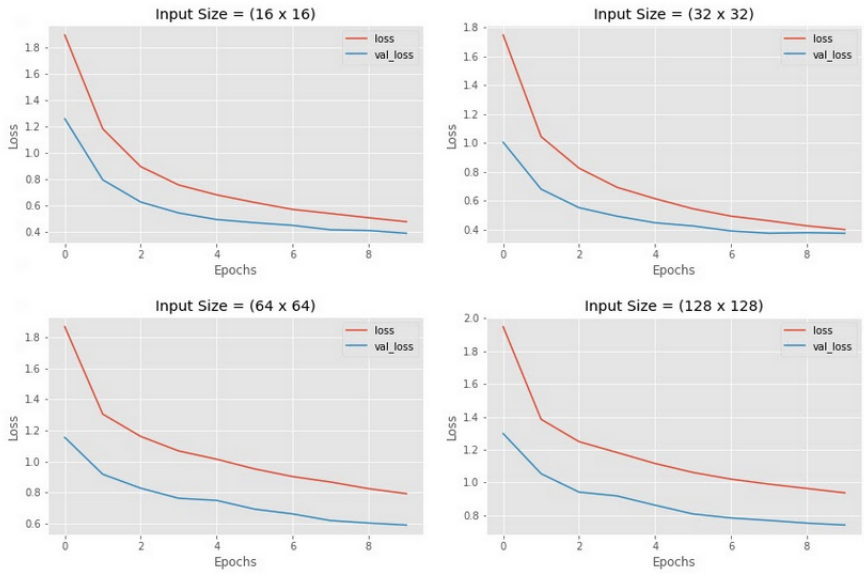


Σχήμα 5.15: Επίδραση αύξησης βάθους του συνελικτικού δικτύου

Και στις δύο περιπτώσεις, με την χρήση περισσότερων επιπέδων, αυξήθηκαν οι παράμετροι προς εκπαίδευση. Αυτό είχε ως αποτέλεσμα την υπερεκπαίδευση. Για το λόγο αυτό διατηρούμε τον αρχικό αριθμό επιπέδων, όπως φαίνεται και στο Σχήμα 5.14. Στην συνέχεια πειραματιζόμαστε και με τα υπόλοιπα επίπεδα.

A2 Μέγεθος Εικόνας Εισόδου (input layer)

Στο σημείο αυτό, ο πειραματισμος αφορούσε το μέγεθος εικόνας εισόδου. Τα μεγέθη που εξετάστηκαν ήταν (16×16), (32×32), (64×64), (128×128). Οι καμπύλες εκπαίδευσης για τις 4 αυτές περιπτώσεις παρουσιάζονται ακολούθως στην Εικόνα 5.16.



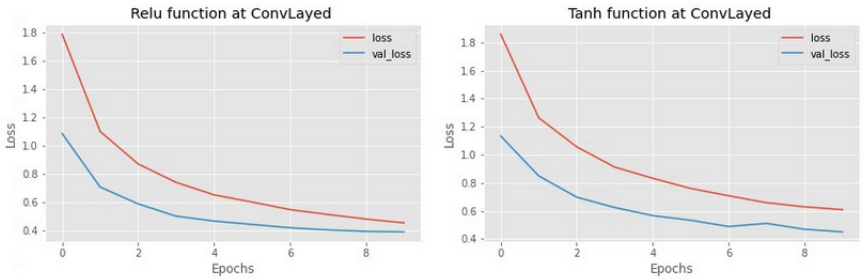
Σχήμα 5.16: Επίδραση του μεγέθους της εικόνας εισόδου

Απο τις παρακάτω γραφικές, καλές προοπτικές φαίνεται να έχει το μέγεθος εισόδου (32×32), καθώς οι δύο καμπύλες val loss train loss συγκλίνουν σε μια σχετικά μικρή τιμή. Στις άλλες περιπτώσεις για μεγέθη εισόδου (64×64) και (128×128), η συνάρτηση απωλειών φαίνεται να σταθεροποιείται σε τιμές οι οποίες ειναι μεγαλυτερες απο αυτή που επιτυγχάνεται με μέγεθος εισόδου (32×32).

A3 Επίπεδα Pooling

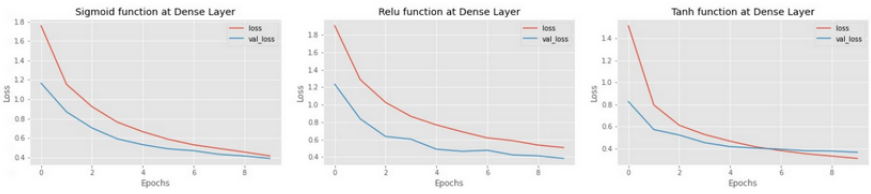
Στην αρχιτεκτονική που φαίνεται στο Σχήμα 0, δοκιμάστηκε η αντικατάσταση max pooling απο average pooling, χωρίς ωστόσο να παρατηρείται κάποια αξιοσημείωτη αλλαγή.

A4 Συναρτήσεις Ενεργοποίησης



Σχήμα 5.17: Πειραματισμός με συναρτήσεις ενεργοποίησης στα Συνελικτικά επίπεδα

Στα συνελικτικά επίπεδα, χρησιμοποιήθηκε η συνάρτηση ενεργοποίησης ReLU. Η αντικατάστασή της, με σιγμοειδή ή tanh (στα επίπεδα αυτά) δεν επέφερε αξιοσημείωτες αλλαγές. Ωστόσο διατηρείται η συνάρτηση ενεργοποίησης relu, καθώς η tanh φαίνεται να οδηγεί σε υπερεκπαίδευση (οι καμπύλες train - val loss φαίνεται να αποκλίνουν).

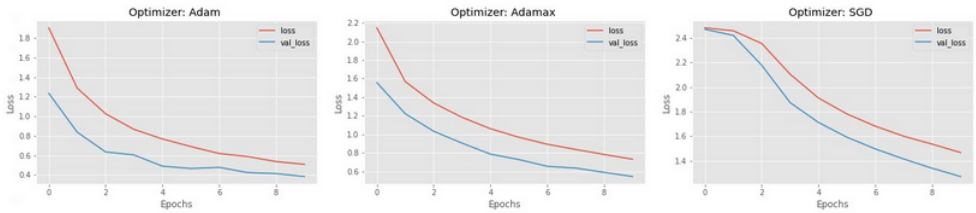


Σχήμα 5.18: Πειραματισμός με συναρτήσεις ενεργοποίησης στα υπόλοιπα επίπεδα

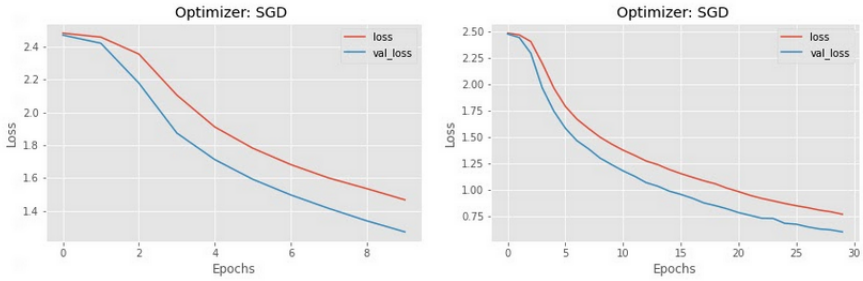
Στα πλήρως συνδεδεμένα επίπεδα (βλ. 5.18) , εξετάστηκε η χρήση και των υπολοίπων συναρτήσεων ενεργοποίησης. Όπως παρατηρεί κανείς, κατάλληλη επιλογή συνιστά η χρήση συνάρτησης relu καθώς οι δυο καμπύλες train val loss φαίνεται να συγκλίνουν ικανοποιητικά για μεγαλύτερο αριθμό εποχών.

B1 Πειραματισμός με βελτιστοποιητές

Οι βελτιστοποιητές οι οποίοι εξετάστηκαν ήταν Adam, Adamax και SGD. Οι διαφορές ανάμεσα σε Adam - Adamax δεν ήταν αισθητές. Ενδιαφέρον παρουσιάζει η τελευταία περίπτωση , του αλγορίθμου SGD. Στην τελευταία περίπτωση αυξάνουμε περαιτέρω τον αριθμό των εποχών (βλ Εικόνα 5.20) μιας και η συνάρτηση loss φαίνεται να έχει καθοδική τάση και όχι να σταθεροποιείται όπως στην περίπτωση Adam Adamax. Ωστόσο με τη χρήση του αλγορίθμου SGD ακόμα και για μεγάλο αριθμό εποχώς η τιμή loss φαίνεται να σταθεροποιείται γύρω απο την τιμή 0.7, που είναι μεγαλύτερη απο την τιμή 0.4 που επετεύχθη με τους Adam -Adamax βελτιστοποιητές.

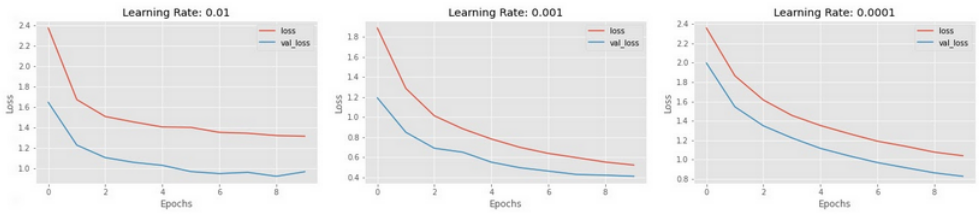


Σχήμα 5.19: Πειραματισμός με βελτιστοποιητές



Σχήμα 5.20: Μελέτη sgd για περισσότερες εποχές

B2 Πειραματισμός με learning rate



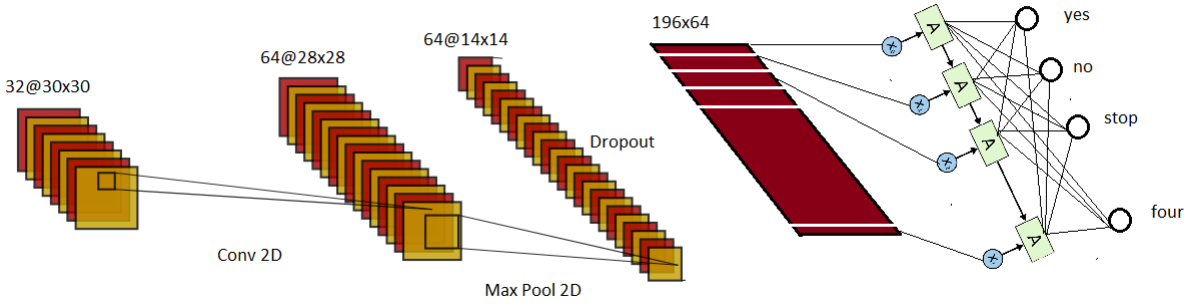
Σχήμα 5.21: Πειραματισμός με learning rate

Αρχικά δοκιμάστηκε και η επίδραση συγκεκριμένων τιμών του ρυθμού εκμάθησης. Η τεχνική που χρησιμοποιήθηκε ήταν η εκπαίδευση του μοντέλου με σχετικά μικρό learning rate στην αρχή, και στην συνέχεια η μείωση του στις τελευταίες εποχές εκπαίδευσης.

Γ1 Πειραματισμός με dropout

Στο σημείο αυτό, έχοντας αποφασίσει την τελική διάταξη, αυξάνουμε τον αριθμό των εποχών. Για την πρόληψη της υπερεκπαιδευσης χρησιμοποιούμε την τεχνική Early Stopping ενώ για την αντιμετώπιση της πραγματοποιείται Κανονικοποίηση L1 - L2 στα συνελικτικά επίπεδα, και προσθήκη Dropout Layers. Ωστόσο η κανονικοποίηση L1 - L2, συγκρατούσε το val accuracy σε μικρές τιμές και έτσι απορρίφθηκε. Τέλος, για το ενδεχόμενο της διακοπής της εκπαίδευσης, τα βάρη αποθηκεύονται και διατηρουνται κάθε φορά που υπάρχει βελτίωση στο val accuracy. Με τη χρήση των παραπάνω το ποσοστό επιτυχίας που σημειώθηκε ήταν 90.3%. Ο πίνακας σύγχυσης για την συγκεκριμένη διάταξη παρουσιάζεται στην Εικόνα 5.26.

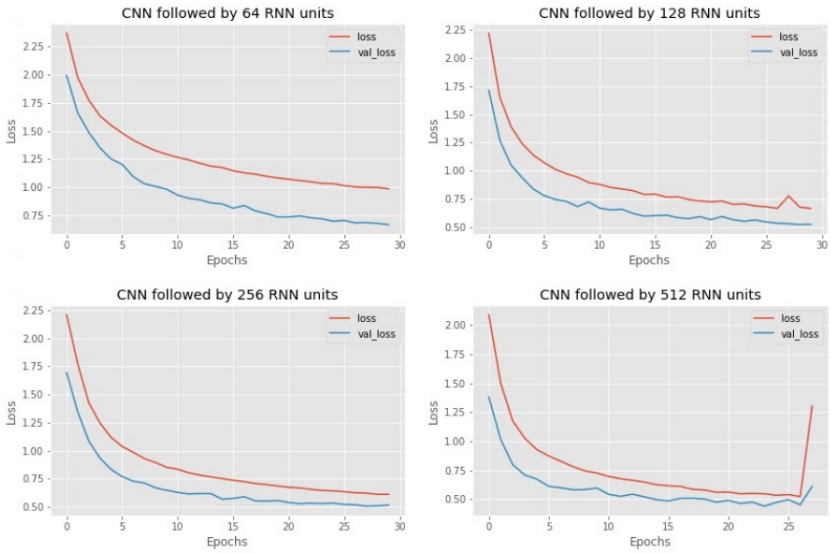
Model 2: Συνελικτικό Δίκτυο (CNN) ακολουθούμενο από αναδρομικό δίκτυο (RNN) εφαρμοσμένο σε χαρακτηριστικά MFCC



Σχήμα 5.22: Συνελικτικό Δίκτυο (CNN) ακολουθούμενο από αναδρομικό δίκτυο (RNN) σε MFCC

Στην συνέχεια, η δομή του συνελικτικου δικτύου στην οποία καταλήξαμε μετά από πειραματισμό, χρησιμοποιήθηκε ώς το πρώτο στάδιο μιας νέας αρχιτεκτονικής συνελικτικού δικτύου CNN ακολουθούμενο απο μονάδες RNN (combined CNN-RNN). Ο πειραματισμός αφορούσε το πλήθος των RNN επιπέδων. Γίνεται αντιληπτό ότι όσο αυξάνεται το πλήθος των μονάδων

RNN, τόσο μειώνεται και το σφάλμα στα σύνολα εκπαίδευσης - επικύρωσης. Στην τελευταία περίπτωση, παρατηρείται μια απότομη αύξηση στο σφάλμα, γεγονός που εξηγεί την διακοπή της εκπαίδευσης πριν την ολοκλήρωση 30 εποχών. Τα ποσοστά ορθότητας, συνοψίζονται στον Πίνακα 5.4:



Σχήμα 5.23: Πειραματισμός με πλήθος μονάδων RNN

RNN units	Accuracy (%)
64	78.6
128	83.8
256	83.6
512	**84.9**
1024	55.1

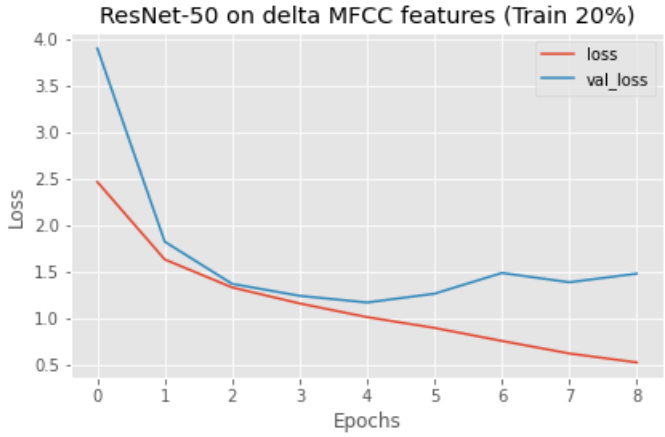
Πίνακας 5.4: Επιδραση του πλήθους των μονάδων RNN στο test accuracy

Model 3 : Μοντέλο μηχανικής μάθησης σε χαρακτηριστικά delta- MFCC

Οι πειραματισμοί ολοκληρώθηκαν με την εισαγωγή - εκπαίδευση συγκεκριμένων αρχιτεκτονικών, που χρησιμοποιούνται ευρέως σε ορισμένα προβλήματα μηχανικής μάθησης. Όπως ειπώθηκε και στο θεωρητικό κομμάτι (βλ. Παράγραφο 3.8), αν το πρόβλημα που πρόκειται να επιλυθεί είναι παρόμοιο με αυτό για το οποίο έχει σχεδιαστεί μια αρχιτεκτονική, τότε αρκεί η εκπαίδευση μερικών επιπέδων. Σε αντίθετη περίπτωση, κρίνεται αναγκαία η εκπαίδευση μεγαλύτερου αριθμού επιπέδων. Σε πρώτο στάδιο λοιπόν, εξετάστηκε το πλήθος των επιπέδων που πρέπει να εκπαιδευτούν, ενώ σε δεύτερο στάδιο δοκιμάστηκαν και άλλες αρχιτεκτονικές.

- Δοκιμή πλήθους επιπέδων προς εκπαίδευση Για την αρχιτεκτονική ResNet50, δοκιμάστηκε η διατήρηση των βαρών των πρώτων επιπέδων και η εκπαίδευση των τελευταίων

(20 %) επιπέδων. Απο την καμπύλη εκμάθησης εύκολα παρατηρεί κανείς απόκλιση μεταξύ της καμπύλης val loss και train loss (βλ. Εικόνα 5.24), γεγονός που σηματοδοτεί υπερεκπαίδευση. Στην συγκεκριμένη περίπτωση το ποσοστό επιτυχίας που σημειώθηκε ήταν 61.7%



Σχήμα 5.24: Επαίδευση 20 % των επιπέδων του ResNet

Έπειτα αυξάνεται το πλήθος των επιπέδων προς εκπαίδευση. Στην περίπτωση εκπαίδευσης του συνόλου των επιπέδων του ResNet-50 το ποσοστό ορθότητας κυμαίνεται στο 93.5 %. Παράλληλα εξετάζονται και άλλες εκδοχές της αρχιτεκτονικής ResNet. Τα αποτελέσματα συνοψίζονται στον ακόλουθο πίνακα. Σε κάθε περίπτωση εκπαιδεύεται το σύνολο των βαρών.

- Δοκιμή διαφορετικών Αρχιτεκτονικών

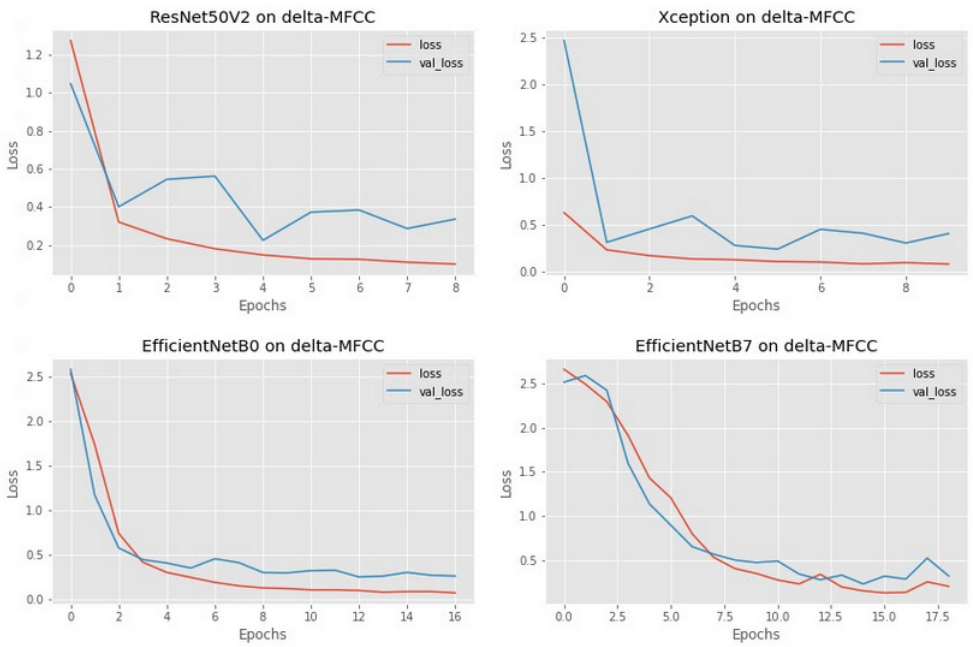
  Εκτός απο την αρχιτεκτονική ResNet, δοκιμάστηκε η εκπαίδευση του συνόλου των βαρών των αρχιτεκτονικών EfficientNet και Xception. Τα αποτελέσματα στο σύνολο εκπαίδευσης συνοψίζονται στον ακόλουθο πίνακα:

Architecture	Accuracy (%)
Res-Net 50	93.5
Res-Net101	94.2
Res-Net152	**94.9**
Res-Net 50V2	93.5
Res-Net101V2	94.7
Res-Net152V2	94.3

Architecture	Accuracy (%)
EfficientNetB0	**94.7**
EfficientNetB3	93.9
EfficientNetB5	93.6
EfficientNetB7	93.7
Xception	94.4

Πίνακας 5.5: Διερεύνηση επεκτάσεων ορισμένων αρχιτεκτονικών

Τέλος ενδιαφέρον παρουσιάζουν και οι καμπύλες εκπαίδευσης για διαφορετικές αρχιτεκτονικές. Καλύτερη συμπεριφορά στην εκπαίδευση μεταξύ των τριών αρχιτεκτονικών ResNet50, Xception και EfficientNet παρουσιάζει το μοντέλο EfficientNet. Στην περίπτωση αυτή οι δυο καμπύλες (train - val loss) συγκλίνουν ικανοποιητικά.



Σχήμα 5.25: Καμπύλες εκπαίδευσης για διάφορες αρχιτεκτονικές

5.9 Αξιολόγηση

Τελευταίο βήμα στο συγκεκριμένο κεφάλαιο αποτελεί η αξιολόγηση των μοντέλων που υλοποιήθηκαν, Η αξιολόγηση πραγματοποείται χρησιμοποιώντας Πίνακες Σύγχησης που επεξηγήθηκαν στην Παράγραφο 3.4.12.

5.9.1 Πεδίο του Χρόνου

Τα ποσοστά ορθότητας που προέκυψαν, χωρίς επιπλέον πειραματισμούς ήταν: Distance Based (kNN - DTW): 19.2%, Interval Based (Random Forest): 33.7 %, Frequency Based (RISE): 22.7%. Με πειραματισμούς ωστόσο, αναφορικά με τις παραμέτρους των επιμέρους αλγορίθμων, τα ποσοστά αυτά παρουσίασαν μια ανοδική τάση. Συγκεκριμένα, Distance Based (kNN - DTW): 22.5 % (+3.3), Interval Based (Random Forest): 40.1 % (+6.4) , Frequency Based (RISE): 35.2 % (+12.5).

5.9.2 Πεδίο της Συχνότητας

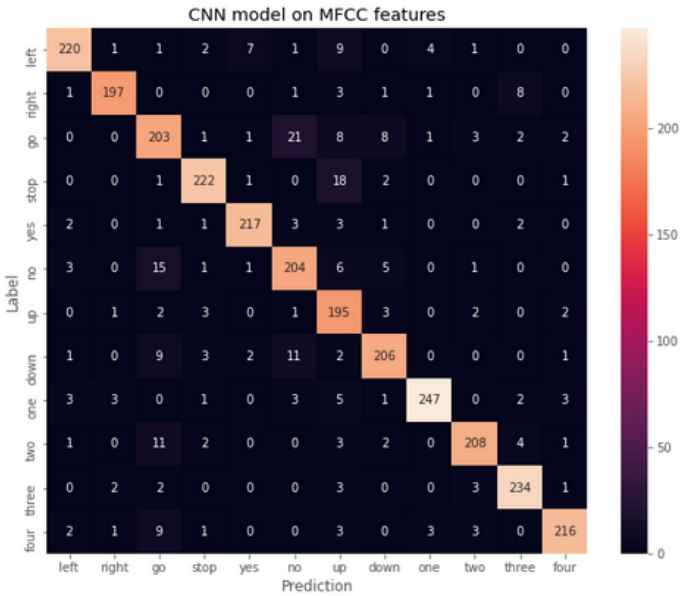
Στο πεδίο της συχνότητας, τα τελικά ποσοστά συγκεντρώνονται στον Πίνακα 5.6. Με τους παραπάνω πειραματισμούς, προέκυψε μια βελτίωση +3% για CNN, +7% για CNN-RNN και +3% για ResNet.

Πίνακες Σύγχησης

Απο τον παρακάνω πίνακα, φαίνεται ότι τα περισσότερα λάθη πραγματοποιούνται μεταξύ των εντολών go και no, καθώς και μεταξύ των εντολών stop και up.Η προσέγγιση CNN-

Accuracy (%) μετά τη Βελτιστοποίηση			
	Architecture	Feature	Acc (%)
Model 1	CNN	mfccs	90.3
Model 2	CNN-RNN	mfccs	84.9
Model 3	ResNet-50	delta-mfccs	94.9

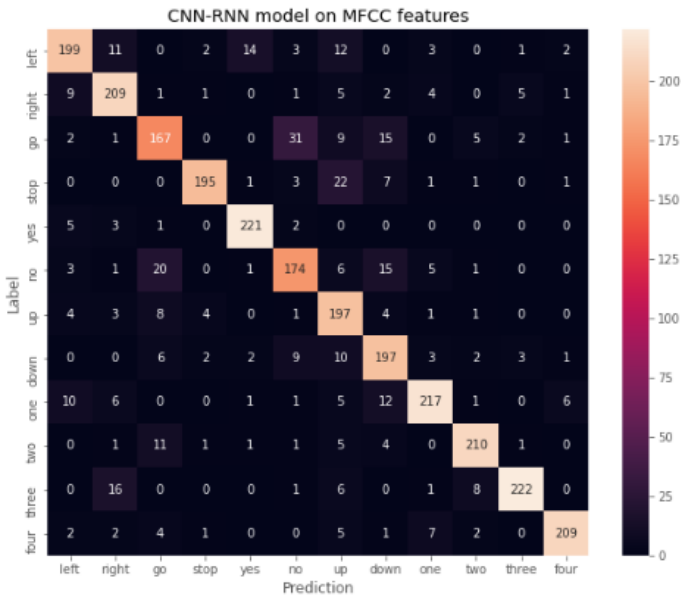
Πίνακας 5.6: Accuracy (%) για Αρχιτεκτονικές- Χαρακτηριστικά, μετά την βλετιστοποιήση



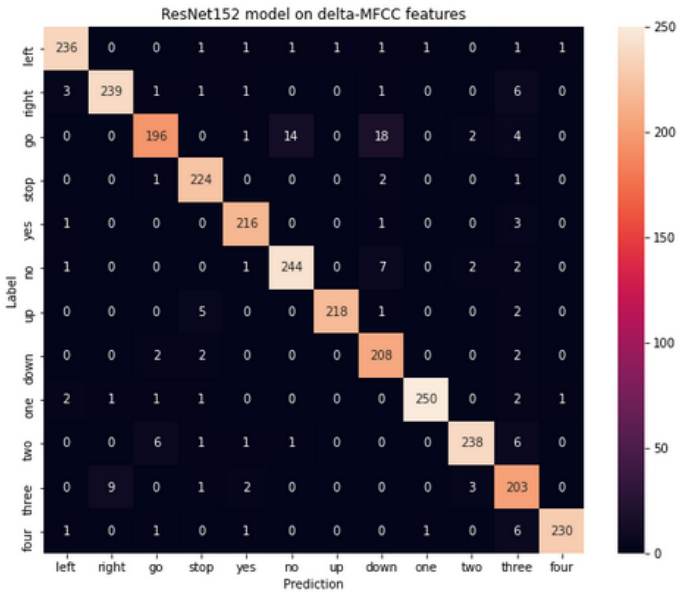
Σχήμα 5.26: Πίνακας Σύγχησης για την αριχτεκτονική CNN

RNN (5.27 δεν μείωσε τον αριθμό των λαθών μεταξύ των εντολών “go”- “no” και “stop” - “up”. Απο την άλλη, παρατηρήθηκε και κάποια αύξηση στα λάθη μεταξύ των κατηγοριών “down”-”no” και three”-”right”

Η εκπαίδευση του μοντέλου ResNet152 μείωσε σημαντικά τα λάθη μεταξύ των εντολών (βλ. Εικόνα 5.28). Είναι αξιοσημείωτο το γεγονός ότι κανένα ακουστικό σήμα που αφορούσε την εντολή “no” δεν ταξινομήθηκε στην κατηγορία “go” . Το αντίθετο ωστόσο παρατηρήθηκε 14 φορές (ποσοστό 0.06 %). Επίσης παρατηρήθηκαν και ορισμένα λάθη μεταξύ των σημάτων που ήταν “go” αλλά ταξινομήθηκαν ως “down”, ποσοστό 0.08 %).



Σχήμα 5.27: Πίνακας Σύγχησης για την αριχτεκτονική CNN-RNN



Σχήμα 5.28: Πίνακας Σύγχησης για την αριχτεκτονική ResNet-152

Κεφάλαιο 6

Σύνοψη-Μελλοντικές Επεκτάσεις

6.1 Σύνοψη

Στην συγκεκριμένη διπλωματική εξετάστηκε η επίδοση αλγορίθμων μηχανικής μάθησης, τόσο στο πεδίο του χρόνου, οσο και στο πεδίο της συχνότητας. Στο πεδίο του χρόνου, αν και τα χαρακτηριστικά (rms τιμές ακουστικού σήματος) δεν ήταν αρκετά αντιπροσωπευτικά της λέξης που προφέρονταν, εντούτοις με τη χρήση Random Forest σημειώθηκε ποσοστό λίγο πάνω απο 40 %. Αναφορικά με το πεδίο της συχνότητας, τα μοντέλα που χρησιμοποιήθηκαν ήταν απλό συνελικτικό δίκτυο (CNN), συνελικτικό δίκτυο ακολουθούμενο απο ανατροφοδοτούμενο δίκτυο (CNN-RNN), καθώς και συγκεκριμένες αρχιτεκτονικές (ResNet, Xception, Efficient Net). Το συνελικτικό δίκτυο, είτε ακολουθούμενο από μονάδες RNN είτε όχι, σημείωσε το καλύτερο ποσοστό σε χαρακτηριστικά mfcc (84.9 %, 90.3 %, αντίστοιχα). Απο την άλλη οι αρχιτεκτονικές ResNet, Xception, Efficient σημείωσαν υψηλά ποσοστά στα χαρακτηριστικά delta- mfccs. Στην καλύτερη περίπτωση, εισάγοντας το μοντέλο ResNet 152 και εκπαιδεύοντας εκ νέου τα βάρη του, το ποσοστό που σημειώθηκε ήταν της τάξης του 94.9 %.

6.2 Μελλοντικές Επεκτάσεις

Η παρούσα εργασία θα μπορούσε να χρησιμοποιηθεί για να καλύψει διάφορες απαιτήσεις ή να προσδώσει ευκολίες σε διάφορες εργασίες. Χρήση ασανσέρ με φωνητικές εντολές, πλοήγηση αυτόνομων ρομποτικών οχημάτων σε χώρους εργασίας, έλεγχος της κατεύθυνσης αμαξιδίων για άτομα με ειδικές ανάγκες και ειδικότερα δυσλειτουργίες στα άκρα, αποτελούν μόνο κάποιες απο τις εφαρμογές του συστήματος που εξετάστηκε και αναπτύχθηκε. Τέλος να σημειωθεί ότι το μοντέλο που κατασκευάστηκε έχει απόδοση 94.9 % στο σύνολο δοκιμής και πραγματοποιούνται ελάχιστα λάθη ως προς κάποιες κατηγορίες κυρίως “no”, “go”. Μια λύση σε αυτό θα ήταν η δημιουργία ενός συνόλου ακολουθιακών καταστάσεων που θα αποδίδουν μια πιθανότητα σε μια λέξη, δεδομένης της προηγούμενης που ακούστηκε. Για παράδειγμα, αν η προηγουμενη λεξη που προφέρθηκε ήταν “stop” πιο πιθανό η λέξη που προφέρεται να είναι “go” παρά “stop”. Αντίστοιχη εργασία μπορεί να πραγματοποιηθεί πάνω σε ένα διαφορετι-

κό σύνολο ακουστικών δεδομένων και έτσι να επεκταθεί το σύνολο των εφαρμογών. Κάτι τέτοιο βέβαια, μπορεί να πραγματοποιηθεί εισάγοντας το μοντέλο που αναπτύχθηκε στην παρούσα διπλωματική, διατηρώντας τα βάρη των αρχικών επιπέδων, και εκπαιδεύοντας τα επίπεδα της κεφαλής. Με λίγα λόγια μπορεί να χρησιμοποιηθεί σαν μοντέλο για μεταφοράς μάθησης (transfer learning). Τέλος, παρουσιάζει ερευνητικά ενδιαφέρον, η ενσωμάτωση του εκπαιδεμένου μοντέλου σε κάποιο άλλο σύστημα ή πλακέτα και η μελέτη του χρόνου απόκρισης, η κατανάλωση ενέργειας και οι απαιτήσεις σε μνήμη.

Βιβλιογραφία

[1] L. R. Rabiner, R. W. Schafer. *Theory and Applications of Digital Speech Processing.* Pearson, 2011.

[2] T. Hastie, R. Tibshirani, J. H. Friedman. *The Elements of Statistical Learning.* Springer, 2001.

[3] A. Oppenheim, R. Schafer. *Discrete-Time Signal Processing.* Pearson, 2009.

[4] S. Haykin *Νευρωνικά Δίκτυα και Μηχανική Μάθηση.* Παπασωτηρίου, 2010.

[5] S. Theodoridis. *Αναγνώριση Προτύπων.* Εκδόσεις Πασχαλίδης, 2011.

[6] Καραγιάννης Γιώργος, Μαραγκός Πέτρος. *Βασικές Αρχές Σημάτων και Συστημάτων.* Παπασωτηρίου, 2011.

[7] Berdibayeva, Gulmira K. and Spirkin, Andrey N. and Bodin, Oleg N. and Bezborodova, Oksana E. *Features of Speech Commands Recognition Using an Artificial Neural Network*, 2021.

[8] Sokolov, Artem and Savchenko, Andrey V. *Voice command recognition in intelligent systems using deep neural networks*, 2019.

[9] Masato Nishimori and Takeshi Saitoh and Ryosuke Konishi. *Voice Controlled Intelligent Wheelchair*, 2007.

[10] Xiaoling Lv and Minglu Zhang and Hui Li *Robot control based on voice command*, 2008.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. *Deep Residual Learning for Image Recognition.* `arXiv:1512.03385`, 2015. Submitted on 10 Dec 2015.

[12] Mingxing Tan, Quoc V. Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* `arXiv:1905.11946`, 2020. Last revised 11 Sep 2020 .

[13] Francois Chollet. *Xception: Deep Learning with Depthwise Separable Convolutions.* `https://arxiv.org/abs/1610.02357`, 2017. Last revised 4 Apr 2017.

[14] Amita C. Patil and Hyung Lee. *Speech to Text Translation Using Google Speech Commands.* 2019.

[15] Yundong Zhang, Naveen Suda1, Liangzhen Lai1,Vikas Chandra. *Hello Edge: Keyword Spotting on Microcontrollers*. 2019.

[16] Katia Moskvitch. The machines that learned to listen. `https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen`, 2017. Last accessed on 15/02/2017.

[17] Pete Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. `https://arxiv.org/pdf/1804.03209.pdf`, 2018.

[18] IBM Shoebox. `https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html`, 2017. Last accessed on 15/02/2017.

[19] Stephanie. The Current Challenges Of Speech Recognition. `https://onlim.com/en/the-current-challenges-of-speech-recognition/` 2017. Last accessed on 15/10/2019.

[20] Neha Kulshreshtha, Avantika Mishra . 15 Uses of Voice Recognition Software Today. `https://fireflies.ai/blog/uses-of-voice-recognition-software`, 2020. Last accessed on 20/01/2020.

[21] Makenzie Holland. Broad use of EHR voice assistants still years away. `https://searchenterpriseai.techtarget.com/feature/Broad-use-of-EHR-voice-assistants-still-years-away`, 2021. Last accessed on 09/03/2021.

[22] Projects: Planetary Microphones . `https://web.archive.org/web/20120127161038/http://www.planetary.org/programs/projects/planetary_microphones/mars_microphone.html`, 2012. Last accessed on 27/01/2012.

[23] Air Force Wright Aeronautical Labs Wright-Patterson Afb Oh AFTI (Advanced Fighter Technology Integration)/F-16. `https://apps.dtic.mil/sti/citations/ADP002863`, 1983.