Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Ακουστική ανάλυση σκηνών σε έξυπνα περιβάλλοντα

Acoustic scene analysis in smart environments

Διπλωματική Εργασία του
Ιωάννη Αθανασίου

Επιβλέποντες: Γεράσιμος Ποταμιάνος, Αναπληρωτής Καθηγητής
Αθανάσιος Κοράκης, Αναπληρωτής Καθηγητής
Νικόλαος Μπέλλας, Αναπληρωτής Καθηγητής

2 Ιουλίου 2019

Περίληψη

Η διπλωματική αυτή εργασία ασχολείται με την ταξινόμηση πολλαπλών αρχείων ήχου, που έχουν καταγραφεί σε διάφορες ευρωπαϊκές μεγάλες πόλεις, σε διαφορετικούς χώρους, όπως μετρό, αεροδρόμιο, τραμ κλπ. Για το σκοπό αυτό χρησιμοποιήθηκαν ταξινομητές βασισμένοι σε Γκαουσιανά μοντέλα μίξης και σε συνελικτικά νευρωνικά δίκτυα. Για την υλοποίηση των ταξινομητών αυτών, χρησιμοποιήθηκε η βάση δεδομένων του πανεπιστημίου Τεχνολογίας του Τάμπερε (TUT), που αποτελείται από ηχογραφήσεις σε ποικίλες ακουστικές σκηνές.

Abstract

This diploma thesis deals with the classification of multiple audio files recorded in various European major cities in different locations such as metro, airport, tram, etc. For this purpose, classifiers based on Gaussian mixture models and convolutional neural networks were used. For the implementation of these classifiers, the TUT (Tampere University of Technology) database was used, consisting of recordings in a variety of acoustic scenes.

Ευχαριστίες

Κατ' αρχήν, θα ήθελα να ευχαριστήσω την οικογένειά μου, για την υποστήριξη και την αγάπη τους, διότι δε θα τα είχα καταφέρει χωρίς αυτούς. Κυρίως τον αδερφό μου, που με βοήθησε και ψυχολογικά και πρακτικά να φτάσω στο σημείο που είμαι. Στη συνέχεια, θα ήθελα να ευχαριστήσω τον επιβλέποντα της διπλωματικής μου εργασίας, τον κ. Γεράσιμο Ποταμιάνο. Οι γνώσεις του και η εμπειρία του ήταν κάτι περισσότερο από βοηθητικές, καθώς και η υποστήριξη και ο σεβασμός που μου έδειχνε όλον αυτόν τον καιρό. Όλα αυτά με ώθησαν να ασχοληθώ με το συγκεκριμένο αντικείμενο και να αποκτήσω αρκετές γνώσεις. Τέλος, θα ήθελα να ευχαριστήσω από καρδιάς τους φίλους μου και συναδέλφους μου, που με στήριζαν καθ' όλη τη φοιτητική μου ζωή. Χωρίς αυτούς δε θα τα κατάφερνα.

Ακουστική ανάλυση σκηνών σε έξυπνα περιβάλλοντα

Acoustic scene analysis in smart environments

Ιωάννης Αθανασίου
iathanasiou@inf.uth.gr
athanjohn757@gmail.com

2 Ιουλίου 2019

Περιεχόμενα

1 Εισαγωγή **6**
- 1.1 Ταξινόμηση ακουστικών σκηνών 6
- 1.2 Σκοπός της διπλωματικής 6
- 1.3 Βάση δεδομένων 7
- 1.4 Οργάνωση της διπλωματικής 8

2 Εξαγωγή χαρακτηριστικών / Ταξινομητές **10**
- 2.1 Αύξηση δεδομένων 11
- 2.2 Προ-επεξεργασία αρχείων ήχου 12
- 2.3 Φασματικοί συντελεστές συχνότητας Mel 13
- 2.4 Βαθιά μάθηση και νευρωνικά δίκτυα 15
  - 2.4.1 Θεωρητικό υπόβαθρο 15
  - 2.4.2 Συνελικτικά νευρωνικά δίκτυα 16
  - 2.4.3 Γκαουσιανό μοντέλο μίξης 17
- 2.5 Χαρακτηριστικά των μοντέλων 18
  - 2.5.1 Χαρακτηριστικά Συνελικτικού Νευρωνικού Δικτύου . 18
  - 2.5.2 Χαρακτηριστικά Γκαουσιανού Μοντέλου Μίξης . . . 23

3 Ανάλυση χαρακτηριστικών πρωτότυπου συστήματος **24**
- 3.1 Σύνολο δεδομένων 24
- 3.2 Διασταυρωμένη επικύρωση αποτελεσμάτων με τη μέθοδο K-fold 25
- 3.3 Πρωτότυπο σύστημα 25

4 Πειράματα κι αποτελέσματα **30**
- 4.1 Χαρακτηριστικά αρχείων ήχου και μοντέλων 30
  - 4.1.1 Συνελικτικό Νευρωνικό Δίκτυο 31
  - 4.1.2 Γκαουσιανό Μοντέλο μίξης 31
- 4.2 Πίνακες αποτελεσμάτων 32
  - 4.2.1 Συνελικτικό νευρωνικό δίκτυο με stereo σύνολο δεδομένων 32

4.2.2 Συνελικτικό νευρωνικό δίκτυο με stereo/pitch shift σύνολο δεδομένων 33
4.2.3 Συνελικτικό νευρωνικό δίκτυο με stereo/white noise σύνολο δεδομένων 34
4.2.4 Συνελικτικό νευρωνικό δίκτυο με σύνολο δεδομένων χωρισμένο σε αριστερό και δεξί κανάλι 34
4.2.5 Γκαουσιανό μοντέλο μίξης με stereo σύνολο δεδομένων 35
4.2.6 Γκαουσιανό μοντέλο μίξης με stereo/pitch shift σύνολο δεδομένων 35
4.2.7 Γκαουσιανό μοντέλο μίξης με stereo/white noise σύνολο δεδομένων 36
4.2.8 Γκαουσιανό μοντέλο μίξης με σύνολο δεδομένων χωρισμένο σε αριστερό και δεξί κανάλι 37

5 Συμπεράσματα και δυσκολίες **38**
5.1 Συμπεράσματα 38
5.2 Δυσκολίες 41

6 Μελλοντικές κατευθύνσεις έρευνας **43**
6.1 Ανασκόπηση της διπλωματικής 43
6.2 Μελλοντικές κατευθύνσεις έρευνας 44

Βιβλιογραφία **45**

Κατάλογος Σχημάτων

2.1 Μέθοδοι εξαγωγής χαρακτηριστικών αρχείων ήχου. Λήφθηκε από το [1] 11
2.2 Οπτική αναπαράσταση της μεθόδου μετατόπισης τόνου του αρχείου ήχου. Λήφθηκε από το [2] 11
2.3 Οπτική αναπαράσταση των βημάτων για τους φασματικούς συντελεστές συχνότητας Mel. Λήφθηκε από το [3] 14
2.4 Διαγράμματα πλάτους-συχνότητας για την κατανόηση των φίλτρων και των παραθύρων. (a) Ολόκληρο το φίλτρο, (b) παράδειγμα φάσματος ισχύος ενός πλαισίου ήχου, (c) φίλτρο 8 του φίλτρου, (d) φάσμα ισχύος με χρήση παραθύρου και του φίλτρου 8, (e) φίλτρο 20, (f) φάσμα ισχύος με χρήση παραθύρου και του φίλτρου 20. Λήφθηκε από το [4] 15
2.5 Σχεδιάγραμμα βιολογικού νευρώνα (αριστερά) και σχεδιάγραμμα τεχνητού νευρώνα (δεξιά). Φαίνεται η σχέση μεταξύ τους και το πώς προέκυψε ο τεχνητός νευρώνας από τον βιολογικό νευρώνα. Λήφθηκε από το [5] 16
2.6 Οπτική αναπαράσταση ενός συνελικτικού νευρωνικού δικτύου. Λήφθηκε από το [6] 17
2.7 Οπτική αναπαράσταση ενός γκαουσιανού μοντέλου μίξης. Λήφθηκε από το [7] 18
2.8 Οπτική αναπαράσταση ενός συνελικτικού επιπέδου. Εδώ έχουμε μία εικόνα 30 x 30 pixels και το φίλτρο 3 x 3 pixels. Πολλαπλασιάζουμε το συγκεκριμένο κομμάτι από pixels της εικόνας με το φίλτρο που έχουμε θέσει και στο τέλος προσθέτουμε όλα τα αποτελέσματα, ώστε να καταλήξουμε στην έξοδο (output volume). Λήφθηκε από το [8] 20
2.9 Οπτική αναπαράσταση των λειτουργιών των τύπων του επιπέδου Pooling. Λήφθηκε από το [9] 21
2.10 Βλέπουμε δύο δίκτυα. Ένα που δεν έχει χρησιμοποιηθεί το επίπεδο Dropout (αριστερά) και ένα που έχει χρησιμοποιηθεί (δεξιά). Λήφθηκε από το [10] 22

2.11 Οπτική αναπαράσταση της λειτουργίας του επιπέδου ισοπέδωσης. Λήφθηκε από το [11] 22
2.12 Η δομή ενός πυκνού / πλήρως συνδεδεμένου δικτύου (a) (b) και η εξάρτηση των δεδομένων (c). Λήφθηκε από το [12] 23

3.1 Διασταυρωμένη επικύρωση αποτελεσμάτων με 5-fold. Λήφθηκε από το [13] 25
3.2 Αρχιτεκτονική πρωτότυπου συστήματος. Λήφθηκε από το [14] 28

4.1 Αρχιτεκτονική συνελικτικού νευρωνικού δικτύου που χρησιμοποίησα. 31

5.1 Διάγραμμα αποτελεσμάτων μοντέλων / συνόλων δεδομένων. Δημιουργήθηκε από το [15] 39
5.2 Διάγραμμα αποτελεσμάτων ακουστικών σκηνών / συνόλων δεδομένων στο συνελικτικό νευρωνικό δίκτυο. Δημιουργήθηκε από το [15] 40
5.3 Διάγραμμα αποτελεσμάτων ακουστικών σκηνών / συνόλων δεδομένων στο γκαουσιανό μοντέλο μίξης. Δημιουργήθηκε από το [15] 40
5.4 Διάγραμμα αποτελεσμάτων μοντέλου / ακουστικών σκηνών στο σύνολο δεδομένων stereo. Δημιουργήθηκε από το [15] . 41

Κατάλογος Πινάκων

3.1 Διαχωρισμός τμημάτων ήχου στις ακουστικές σκηνές και στις πόλεις 24
3.2 Αποτελέσματα πειράματος του πρωτότυπου συστήματος . . 29

4.1 Χρόνοι εκπαίδευσης που χρειάστηκαν για να τρέξει το κάθε μοντέλο χρησιμοποιώντας το κάθε σύνολο δεδομένων 32
4.2 Συνελικτικό νευρωνικό δίκτυο / stereo 33
4.3 Συνελικτικό νευρωνικό δίκτυο / stereo - pitch shift 33
4.4 Συνελικτικό νευρωνικό δίκτυο / stereo - white noise 34
4.5 Συνελικτικό νευρωνικό δίκτυο / Left-Right channels 35
4.6 Γκαουσιανό μοντέλο μίξης / stereo 35
4.7 Γκαουσιανό μοντέλο μίξης / stereo-pitch shift 36
4.8 Γκαουσιανό μοντέλο μίξης / stereo - white noise 36
4.9 Γκαουσιανό μοντέλο μίξης / Left-Right channels 37

5.1 Τελικά ποσοστά για τα δύο μοντέλα για κάθε σύνολο δεδομένων 38

Κεφάλαιο 1

Εισαγωγή

1.1 Ταξινόμηση ακουστικών σκηνών

Η ταξινόμηση ακουστικών σκηνών αποτελεί σημαντικό θέμα στον τομέα τις περιβαλλοντικής ταξινόμησης και αναγνώρισης, ως ένα γενικό πρόβλημα κατάταξης που θέτει τα θεμέλια για την επίγνωση του περιβάλλοντος σε συσκευές, ρομπότ και πολλές άλλες εφαρμογές. Το πρόβλημα της ταξινόμησης ακουστικών σκηνών δεν είναι καινούργιο, αλλά έχει έρθει στο προσκήνιο μέσα στην τελευταία δεκαετία. Σε αυτό το διάστημα, οι προσεγγίσεις μηχανών μάθησης που χρησιμοποιήθηκαν ώστε να λυθεί το πρόβλημα αυτό έχουν αλλάξει δραματικά, με την βαθιά μάθηση να είναι αυτή τη στιγμή η πιο δημοφιλής προσέγγιση [16]. Ωστόσο, οι ήχοι του περιβάλλοντος αλλάζουν συνεχώς στο πέρασμα του χρόνου. Δηλαδή, ο ίδιος ήχος δεν θα συμβεί απαραίτητα ξανά. Οι άνθρωποι μπορούν να ανταποκρίνονται ευέλικτα σε μηδαμινές διαφορές των ήχων, ανάλογα με την εμπειρία τους, πράγμα που είναι τρομερά δύσκολο να αυτοματοποιηθεί με υπολογιστές [14]. Έχουν προταθεί και αξιολογηθεί διάφορα ακουστικά και οπτικά χαρακτηριστικά. Όμως, λίγες μελέτες έχουν διερευνήσει το σύνολο ακουστικών και οπτικών χαρακτηριστικών για την ταξινόμηση ακουστικών σκηνών [17].

1.2 Σκοπός της διπλωματικής

Στις μέρες μας, η ταξινόμηση ακουστικών σκηνών είναι ένα πολύ σημαντικό εργαλείο στον κλάδο της Αναγνώρισης Προτύπων. Σκοπός της διπλωματικής αυτής είναι να μελετηθούν διάφοροι ταξινομητές και διάφορες τεχνικές, ώστε να αυξήσουμε τα ποσοστά επιτυχίας της κατάταξης

αυτής. Καθώς ο κλάδος των νευρωνικών δικτύων είναι ακόμη καινούργιος και εξελίσσεται ραγδαία, στη διπλωματική αυτή έγκειται η ανάπτυξη ενός συστήματος για την κατάταξη αρχείων ήχου σε συγκεκριμένες ακουστικές σκηνές. Επομένως, δημιουργήθηκαν δύο συστήματα. Το ένα βασίζεται σε συνελικτικό νευρωνικό δίκτυο και το άλλο βασίζεται σε γκαουσιανό μοντέλο μίξης. Για την υλοποίηση του κώδικα, χρησιμοποιήθηκε η βιβλιοθήκη dcase_util, σε προγραμματιστική γλώσσα python.

1.3 Βάση δεδομένων

Η βάση δεδομένων που χρησιμοποιήθηκε σε αυτή τη διπλωματική είναι η TUT Urban Acoustic Scenes 2018, η οποία ηχογραφήθηκε σε 6 μεγάλες ευρωπαϊκές πόλεις. Για κάθε κατηγορία σκηνών, οι ηχογραφήσεις έγιναν σε διαφορετικές περιοχές. Για κάθε περιοχή ηχογραφήσεων, υπάρχουν 5-6 λεπτά ήχου. Οι αρχικές ηχογραφήσεις χωρίστηκαν σε τμήματα ήχου με διάρκεια 10 δευτερολέπτων, τα οποία παρέχονται ως ξεχωριστά αρχεία.

Η κύρια συσκευή ηχογράφησης αποτελείται από Soundman OKM I-I Klassik/studio A3, από electret binaural microphone και από ένα Zoom F8 μηχάνημα εγγραφής ήχου, τα οποία χρησιμοποιούν 48kHz ρυθμό δειγματοληψίας και 24 bit ανάλυση [18]. Τα μικρόφωνα είναι ειδικά σχεδιασμένα να μοιάζουν με ακουστικά, τα οποία τοποθετούνται στα αυτιά. Αυτό έχει σαν αποτέλεσμα, ο ήχος να μοιάζει αρκετά στον ήχο που φτάνει στο ανθρώπινο ακουστικό σύστημα του ατόμου που φοράει τον εξοπλισμό.

Η βάση αυτή, ηχογραφήθηκε από το Tampere University of Technology τη χρονική περίοδο 01/2018 - 03/2018. Η συλλογή δεδομένων χρηματοδοτήθηκε από το ευρωπαϊκό συμβούλιο έρευνας (European Research Council).

Χρησιμοποιήθηκε ένα σύνολο δεδομένων, το TUT Urban Acoustic Scenes 2018 development dataset, το οποίο αποτελείται από 24 ώρες ήχου, ισορροπημένες μεταξύ των κατηγοριών.

Στον διαγωνισμό του DCASE 2018, πραγματοποιήθηκαν αρκετά πειράματα. Ένα παράδειγμα, το οποίο είναι παρόμοιο με αυτή την πτυχιακή, είναι τα πειράματα που έγιναν από τους Matthias, Dorfer, Bernhard, Lehner, Hamid, Eghbalzadeh κ.α., οι οποίοι συνεργάζονταν με το Ινστι-

τούτο Υπολογιστικής Αντίληψης (CPJKU), Johannes Kepler University Linz, Austria. Η υλοποίησή τους αφορούσε ένα σύνολο από συνελικτικά νευρωνικά δίκτυα και τα ποσοστά επιτυχίας τους ήταν 80.5% [19].

1.4 Οργάνωση της διπλωματικής

Η διπλωματική χωρίζεται σε 6 κεφάλαια σχετικά με το θέμα της ταξινόμησης αρχείων ήχου σε συγκεκριμένες ακουστικές σκηνές, όπου σε κάθε κεφάλαιο αναλύεται ένα συγκεκριμένο θέμα. Πιο συγκεκριμένα:

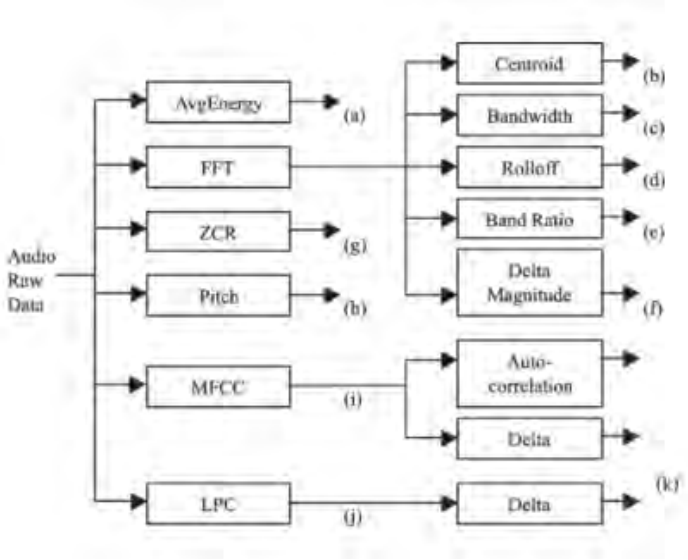
- **Κεφάλαιο 2:** Περιέχει όλη τη θεωρία και την πληροφορία που υπάρχει στη συγκεκριμένη διπλωματική, σχετικά με τον τρόπο λύσης του προβλήματος της κατάταξης ακουστικών σκηνών. Σε αυτό το κεφάλαιο αναλύεται η προ-επεξεργασία των αρχείων ήχου και η θεωρία της εξαγωγής χαρακτηριστικών φασματικών συντελεστών συχνότητας Μελ (MFCCs). Επιπροσθέτως, δίνονται επιπλέον πληροφορίες για τα μοντέλα που χρησιμοποιούνται, δηλαδή για το συνελικτικό νευρωνικό δίκτυο και το γκαουσιανό μοντέλο μίξης, καθώς και περισσότερο μαθηματικό υπόβαθρο, για την καλύτερη κατανόησή τους.
- **Κεφάλαιο 3:** Γίνεται αναφορά στο πρωτότυπο σύστημα του DCASE 2018 – Task1A, πάνω στο οποίο στηρίζεται αυτή η διπλωματική. Θα αναφερθούν οι τεχνικές που χρησιμοποιεί και τα χαρακτηριστικά των ηχητικών σημάτων. Ακόμα, θα υπάρχουν σχετικοί πίνακες με τα ποσοστά επιτυχίας του πρωτότυπου συστήματος σε κάθε κατηγορία, αλλά και κατά μέσο όρο.
- **Κεφάλαιο 4:** Εδώ παρουσιάζονται εκτενέστερα όλα τα εργαλεία και όλες οι υλοποιήσεις για την εκτέλεση των πειραμάτων. Περιγράφονται όλες οι βιβλιοθήκες που χρησιμοποιήθηκαν για τα αρχεία python, που βοήθησαν στην προ-επεξεργασία του ήχου, στους ταξινομητές, στο πρωτότυπο σύστημα και γενικότερα σε ό,τι αρχείο python δούλεψα. Στη συνέχεια, παρουσιάζονται οι πίνακες αποτελεσμάτων.
- **Κεφάλαιο 5:** Παρουσιάζονται σε διαγράμματα οι συγκρίσεις μεταξύ των μοντέλων που χρησιμοποιήθηκαν κατά την υλοποίηση του πειράματος. Στη συνέχεια, σχολιάζονται και οι δυσκολίες που αντιμετώπισα για τις πιο βέλτιστες λύσεις.
- **Κεφάλαιο 6:** Στο τελευταίο κεφάλαιο, γίνεται μία ανασκόπηση της

διπλωματικής εργασίας και παρουσιάζονται υλοποιήσεις που θα φέρουν καλύτερα αποτελέσματα στο μέλλον.

Κεφάλαιο 2

Εξαγωγή χαρακτηριστικών / Ταξινομητές

Υπάρχουν αρκετοί τρόποι για να γίνει η εξαγωγή των χαρακτηριστικών των αρχείων ήχου, όπως log-mel ενέργειες, φασματικοί συντελεστές συχνότητας Mel, νοηματικό φάσμα ισχύος, μετασχηματισμός constant-Q (CQT) κ.α. Κάθε τρόπος εξαγωγής έχει τα δικά του χαρακτηριστικά. Η επιλογή των κατάλληλων χαρακτηριστικών είναι περίπλοκη, διότι θέλουμε να βρούμε την καλύτερη δυνατή λύση. Από τη στιγμή που έχουμε να κάνουμε με αρχεία ήχου, στην υλοποίησή μου χρησιμοποίησα φασματικούς συντελεστές συχνότητας Mel. Όμως, πριν από την εξαγωγή των χαρακτηριστικών, γίνεται η προ-επεξεργασία των αρχείων ήχου και η αύξηση των δεδομένων. Με τον τρόπο αυτό, βοηθάμε το μοντέλο μας να γίνει πιο ανθεκτικό σε λάθη και πιο ευέλικτο. Σημειώνεται πως η αύξηση αυτή γίνεται με διάφορους χειρισμούς του υπάρχοντος συνόλου δεδομένων και δεν χρησιμοποιούνται εξωτερικά σύνολα.

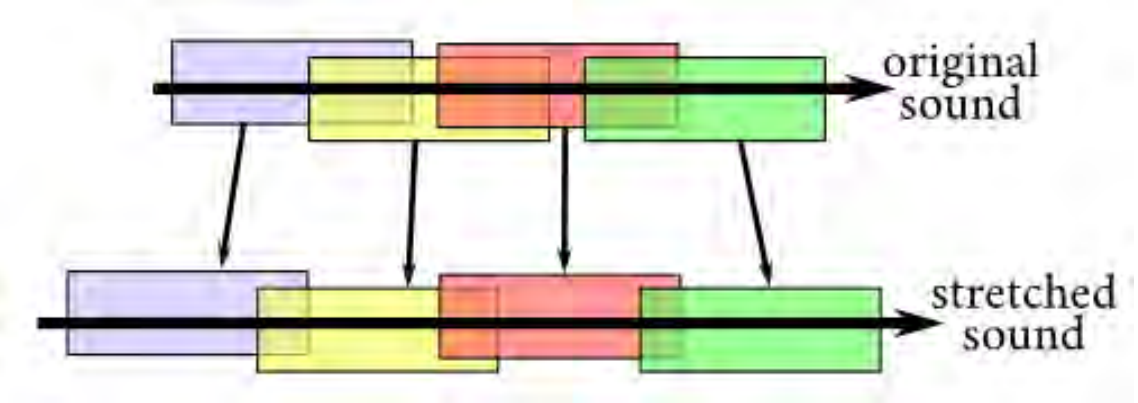


Σχήμα 2.1: Μέθοδοι εξαγωγής χαρακτηριστικών αρχείων ήχου. Λήφθηκε από το [1]

2.1 Αύξηση δεδομένων

Το πρώτο βήμα που ακολούθησα, πριν την προεπεξεργασία είναι η αύξηση των δεδομένων μου. Για κάθε αρχείο ήχου του συνόλου δεδομένων, προσπάθησα να του αλλάξω συγκεκριμένα χαρακτηριστικά, όπως η μετατόπιση του τόνου και η προσθήκη λευκού θορύβου.

- **Μετατόπιση τόνου:** Η μετατόπιση του τόνου του ήχου, μπορεί να γίνει με δύο βήματα. Το πρώτο είναι το τέντωμα του ήχου, κατά το οποίο «σπάμε» το αρχείο ήχου σε επικαλυπτόμενα bits και τα διαμορφώνουμε ώστε να είναι ακόμη πιο επικαλυπτόμενα, αν θέλουμε να μικρύνουμε το μήκος του αρχείου ήχου ή να είναι λιγότερο επικαλυπτόμενα, αν θέλουμε να το μεγαλώσουμε.



Σχήμα 2.2: Οπτική αναπαράσταση της μεθόδου μετατόπισης τόνου του αρχείου ήχου. Λήφθηκε από το [2]

Το δεύτερο βήμα είναι η μεταβολή της ταχύτητας του ήχου. Σε περίπτωση που θέλουμε υψηλότερο τόνο, τότε αυξάνουμε την ταχύτητα του ήχου ώστε το τελικό αρχείο να έχει ίδια διάρκεια με το κανονικό, αλλά υψηλότερο τόνο λόγω της μεγαλύτερης ταχύτητας.

- **Προσθήκη λευκού θορύβου:** Ο λευκός θόρυβος είναι ένας τύπος θορύβου που παράγεται αν συνδυάσουμε τους ήχους όλων των διαφορετικών συχνοτήτων μαζί. Παρόλο που στις περισσότερες περιπτώσεις θέλουμε να μειώσουμε όσο γίνεται περισσότερο τον θόρυβο των ήχων, το να προσθέσουμε λευκό θόρυβο μας βοηθάει αρκετά σε περιπτώσεις αστικών ήχων. Για παράδειγμα, ας πάρουμε δύο κατηγορίες από το DCASE. Μία είναι το metro-station και η άλλη είναι το airport. Και στις δύο περιπτώσεις, συναντάμε θόρυβο στο βάθος, όπως οι συνομιλίες των ανθρώπων, τα βήματά τους κλπ. Εμείς, επειδή θέλουμε να ταξινομήσουμε τα αρχεία ήχου στις κατηγορίες αυτές, είναι εμφανές, πως όταν έχουμε κοινούς ήχους σε διαφορετικές κατηγορίες, υπάρχει περίπτωση αποτυχίας. Επομένως, καθαρίζουμε τον θόρυβο στο βάθος και κρατάμε τον βασικό ήχο, που στο metro-station είναι το μετρό και στο airport είναι το αεροπλάνο. Επομένως, προσθέτουμε ελάχιστο λευκό θόρυβο, ώστε να καλύψουμε τον θόρυβο στο βάθος, αλλά να μην καλύψουμε και τον βασικό ήχο.

2.2 Προ-επεξεργασία αρχείων ήχου

Στη συνέχεια, πριν την εξαγωγή χαρακτηριστικών, είναι η προ-επεξεργασία ήχου. Γίνεται δηλαδή μια αλλαγή στα χαρακτηριστικά που φέρει το ηχητικό σήμα. Αυτές οι αλλαγές γίνονται κυρίως για την αφαίρεση ανεπιθύμητων τμημάτων, ανεπιθύμητων συχνοτήτων και οτιδήποτε μας είναι αδιάφορο και αρνητικό για την εξαγωγή των αποτελεσμάτων.

Αρχικά, γίνεται υποδειγματοληψία του ρυθμού δειγματοληψίας του αρχείου. Από τα χαρακτηριστικά του μικροφώνου, το αρχείο ήχου έχει 48kHz ρυθμό δειγματοληψίας και κατεβαίνει στα 44.1kHz. Αυτό γίνεται, διότι στις υψηλές συχνότητες υπάρχει σχεδόν μηδενική πληροφορία, κάτι που δεν μας χρησιμεύει.

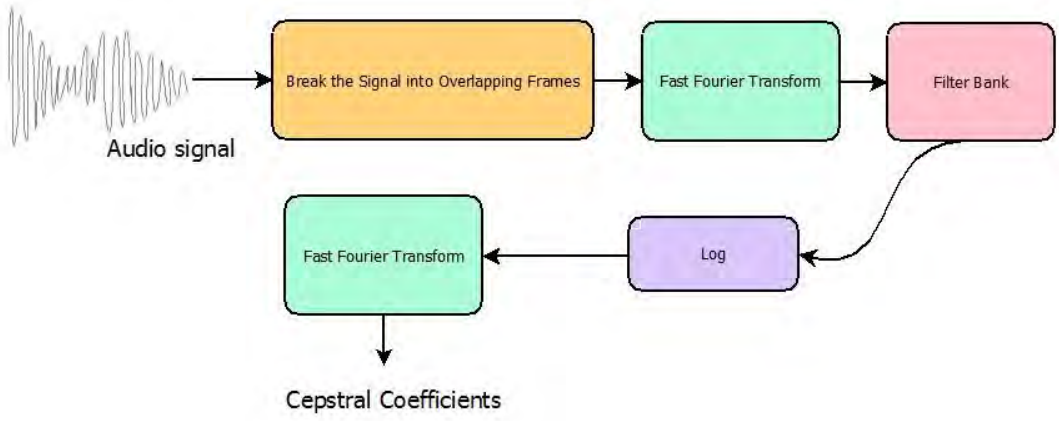
Επίσης, υπάρχουν δύο παράμετροι που πρέπει να θέσουμε. Η μία είναι το μέγεθος του παραθύρου. Με αυτό το παράθυρο, «σπάμε» το σήμα σε μικρά πλαίσια, δηλαδή μικρά χρονικά διαστήματα. Θέτω το μέγεθος παραθύρου σε 40ms. Αυτό σημαίνει ότι σε ένα αρχείο ήχου με ρυθμό

δειγματοληψίας 44.1kHz και μέγεθος παραθύρου 40ms, το μέγεθος των πλαισίων είναι $0.04 \times 44100 = 1764$ δείγματα. Η επόμενη παράμετρος είναι το μέγεθος του βήματος. Ένα βήμα ενός πλαισίου είναι συνήθως 10ms, αλλά στη συγκεκριμένη διπλωματική αλλάζει σε 20ms. Αυτό επιτρέπει να υπάρχει μία επικάλυψη σε κάθε πλαίσιο. Η εξαγωγή των χαρακτηριστικών γίνεται, στη συνέχεια, σε κάθε πλαίσιο του αρχείου ήχου.

2.3 Φασματικοί συντελεστές συχνότητας Mel

Μετά την αύξηση των δεδομένων και την προ-επεξεργασία των αρχείων ήχου, σειρά έχει η εξαγωγή των χαρακτηριστικών. Οι φασματικοί συντελεστές συχνότητας Mel είναι ευρέως γνωστοί και χρησιμοποιούνται για την αυτόματη ομιλία και την αναγνώριση φωνής. Παρουσιάστηκαν από τους Davis και Mermelstein τη δεκαετία του '80 και είναι αρκετά χρήσιμοι από τότε. Συνοπτικά, τα βήματα που ακολουθούνται είναι τα εξής:

1. Πλαισίωση του σήματος σε μικρότερα πλαίσια
2. Για κάθε πλαίσιο υπολογίζεται η εκτίμηση του περιοδογράμματος του φάσματος ισχύος
3. Εφαρμογή των mel φίλτρων στο φάσμα ισχύος, πρόσθεση της ενέργειας σε κάθε φίλτρο
4. Παίρνουμε τον λογάριθμο όλων των ενεργειών των φίλτρων
5. Παίρνουμε τον διακριτό μετασχηματισμό συνημιτόνου (DCT).
6. Κρατάμε τους συντελεστές του διακριτού μετασχηματισμού συνημιτόνου 2-13 και τους υπόλοιπους τους διαγράφουμε.



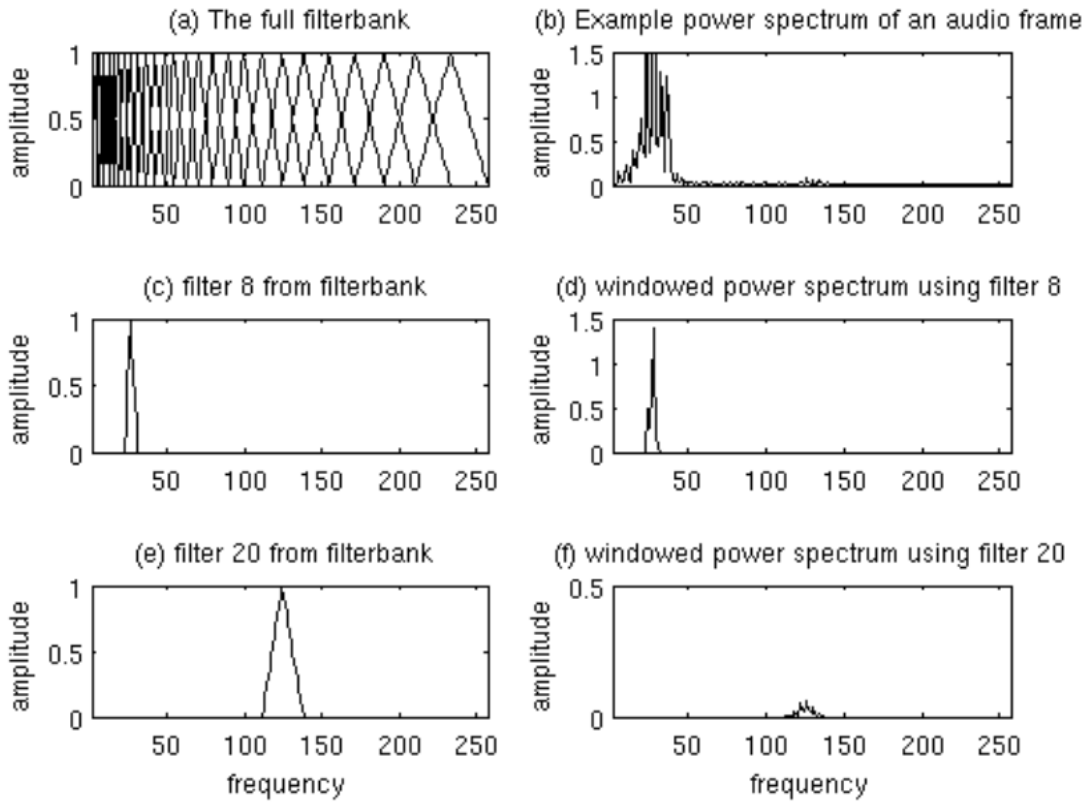
Σχήμα 2.3: Οπτική αναπαράσταση των βημάτων για τους φασματικούς συντελεστές συχνότητας Mel. Λήφθηκε από το [3]

Ο τύπος για να μετατρέψουμε την συχνότητα σε κλίμακα Mel είναι:

$$M(f) = 1125 \times \ln(1 + \frac{f}{100})$$

Ο τύπος για να μετατρέψουμε από κλίμακα Mel σε συχνότητα είναι:

$$M^{-1}(m) = 700 \times e^{\frac{m}{1125}-1}$$



Σχήμα 2.4: Διαγράμματα πλάτους-συχνότητας για την κατανόηση των φίλτρων και των παραθύρων. (a) Ολόκληρο το φίλτρο, (b) παράδειγμα φάσματος ισχύος ενός πλαισίου ήχου, (c) φίλτρο 8 του φίλτρου, (d) φάσμα ισχύος με χρήση παραθύρου και του φίλτρου 8, (e) φίλτρο 20, (f) φάσμα ισχύος με χρήση παραθύρου και του φίλτρου 20. Λήφθηκε από το [4]

Στη συγκεκριμένη υλοποίηση, ο τύπος παραθύρου που χρησιμοποιείται είναι ασύμμετρο Hamming παράθυρο. Ο τύπος για το Hamming παράθυρο είναι:

$$w[n] = 0.54 - 0.46 \times \cos(2\pi \times \frac{n}{N}), 0 \leq n \leq N$$

Το μέγεθος του παραθύρου δίνεται από τη σχέση: $L = N + 1$

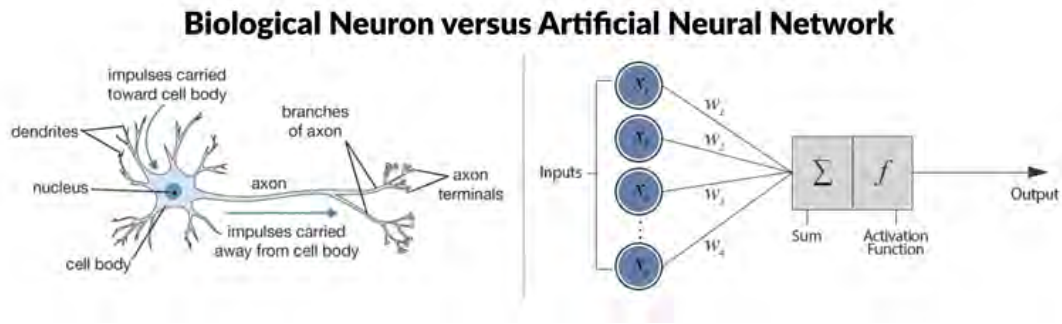
2.4 Βαθιά μάθηση και νευρωνικά δίκτυα

2.4.1 Θεωρητικό υπόβαθρο

Νευρωνικό δίκτυο ονομάζεται ένα κύκλωμα διασυνδεδεμένων νευρώνων. Πρόκειται για ένα αφηρημένο αλγοριθμικό κατασκεύασμα το οποίο ε-

μπίπτει στον τομέα της υπολογιστικής νοημοσύνης, όταν στόχος του νευρωνικού δικτύου είναι η επίλυση κάποιου υπολογιστικού προβλήματος [20].

Οι Warren McCulloch και Walter Pitts (1943) δημιούργησαν ένα υπολογιστικό μοντέλο για νευρωνικά δίκτυα που βασίζεται σε μαθηματικά και αλγόριθμους και ονομάζεται λογική κατωφλίου. Αυτό το μοντέλο άνοιξε το δρόμο για την έρευνα νευρωνικών δικτύων για να χωριστεί σε δύο προσεγγίσεις. Μια προσέγγιση επικεντρώθηκε στις βιολογικές διεργασίες στον εγκέφαλο, ενώ η άλλη επικεντρώθηκε στην εφαρμογή των νευρωνικών δικτύων στην τεχνητή νοημοσύνη. Αυτό το έργο οδήγησε στην εργασία σε δίκτυα νυερώνων και τη σύνδεσή τους με μαθηματικά μοντέλα υπολογισμού [21].

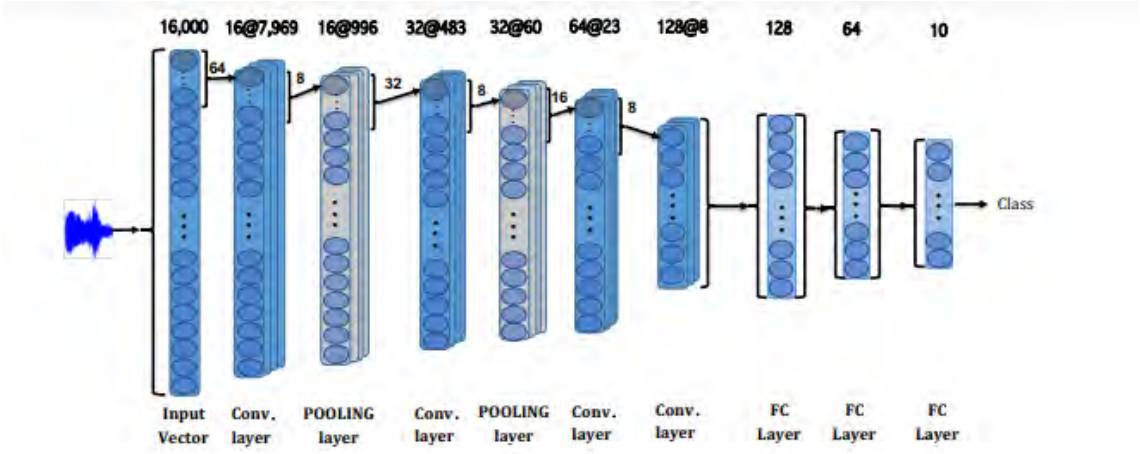


Σχήμα 2.5: Σχεδιάγραμμα βιολογικού νευρώνα (αριστερά) και σχεδιάγραμμα τεχνητού νευρώνα (δεξιά). Φαίνεται η σχέση μεταξύ τους και το πώς προέκυψε ο τεχνητός νευρώνας από τον βιολογικό νευρώνα. Λήφθηκε από το [5]

2.4.2 Συνελικτικά νευρωνικά δίκτυα

Ένα συνελικτικό νευρωνικό δίκτυο αποτελείται από ένα επίπεδο εισόδου και ένα επίπεδο εξόδου. Ενδιάμεσα, μπορούν να υπάρχουν πολλαπλά κρυφά επίπεδα. Τα κρυφά επίπεδα ενός συνελικτικού νευρωνικού δικτύου τυπικά αποτελούνται από συνελικτικά επίπεδα, επίπεδα διορθωμένης γραμμικής μονάδας (ReLU), όπως συνάρτηση ενεργοποίησης, επίπεδα pooling, πλήρως συνδεδεμένα επίπεδα και επίπεδα κανονικοποίησης.

Μπορούν να χρησιμοποιηθούν για αναγνώριση εικόνας, ανάλυση βίντεο, επεξεργασία φυσικής γλώσσας, ανακάλυψη φαρμάκων κ.α. [22]



Σχήμα 2.6: Οπτική αναπαράσταση ενός συνελικτικού νευρωνικού δικτύου. Λήφθηκε από το [6]

2.4.3 Γκαουσιανό μοντέλο μίξης

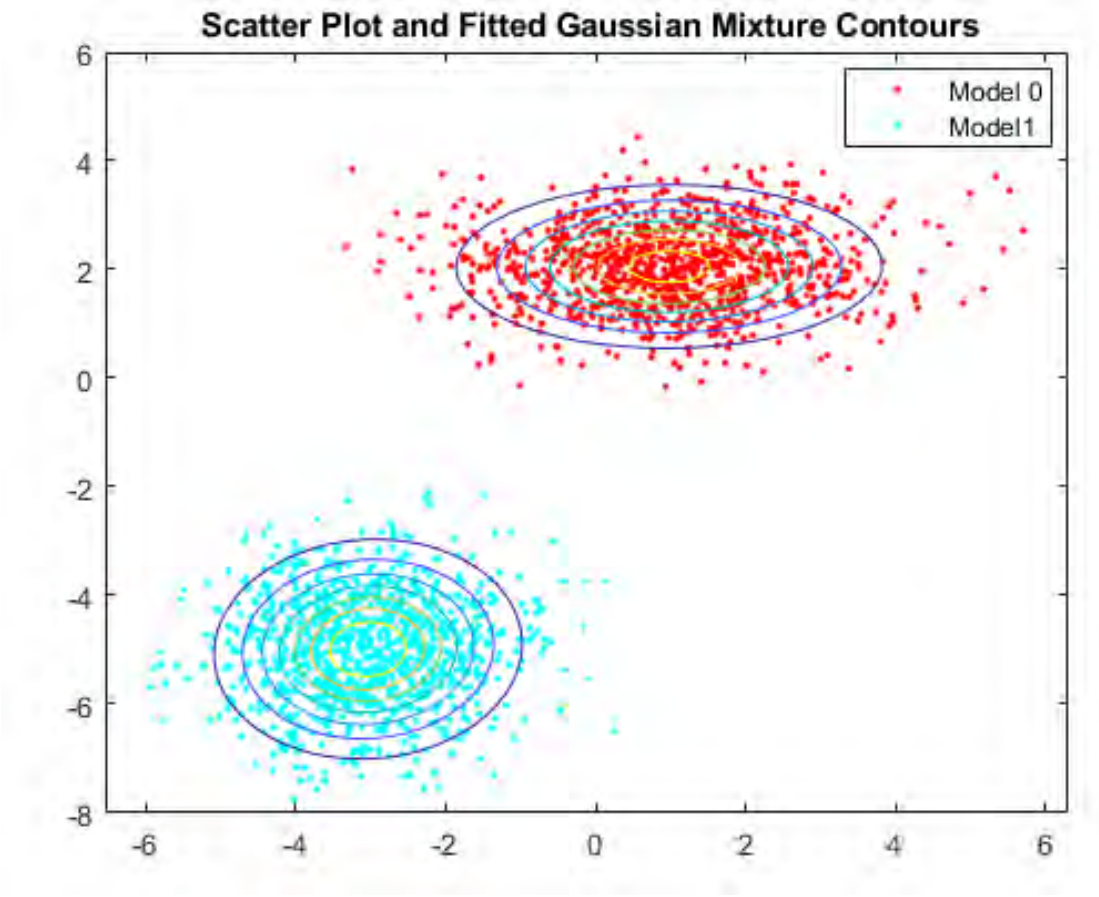
Στην πραγματική ζωή, πολλές βάσεις δεδομένων μπορούν να μοντελοποιηθούν με γκαουσιανή κατανομή. Επομένως, είναι φυσικό να υποθέσουμε ότι οι συστάδες προέρχονται από διαφορετικές γκαουσιανές κατανομές. Ή διαφορετικά, γίνεται προσπάθεια να μοντελοποιηθεί η βάση δεδομένων ως ένα μείγμα διαφόρων γκαουσιανών κατανομών.

Σε μία διάσταση, η συνάρτηση πυκνότητας πιθανότητας μιας γκαουσιανής κατανομής δίνεται από τον τύπο:

$$G(X|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{-x-\mu^2}{2\sigma^2}}$$

όπου μ και σ^2 είναι η μέση τιμή και η διασπορά της κατανομής αντίστοιχα [23].

Τα γκαουσιανά μοντέλα μίξης είναι ουσιαστικά πιθανολογικά μοντέλα για την αντιπροσώπευση των κατανεμημένων υποπληθυσμών σε έναν συνολικό πληθυσμό. Τα μοντέλα μίξης δεν απαιτούν να μάθουμε σε ποιον υποπληθυσμό ανήκει ένα σημείο δεδομένων. Αυτό επιτρέπει στο μοντέλο να μαθαίνει αυτόματα τους υποπληθυσμούς. Δεδομένου ότι η εκχώρηση υποπληθυσμού δεν είναι γνωστή, αυτό αποτελεί μία μορφή μη εποπτευόμενης μάθησης [24].



Σχήμα 2.7: Οπτική αναπαράσταση ενός γκαουσιανού μοντέλου μίξης. Λήφθηκε από το [7]

2.5 Χαρακτηριστικά των μοντέλων

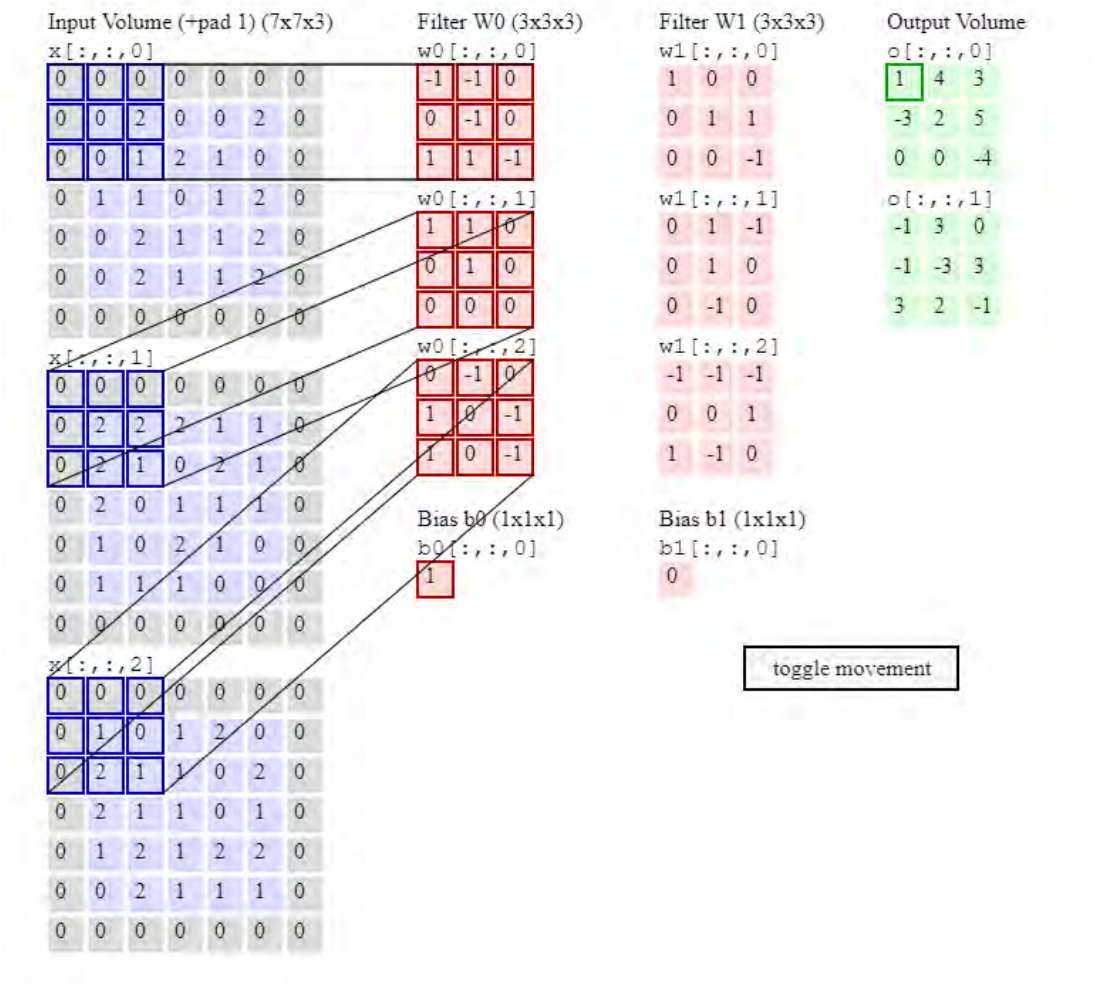
Σε αυτή την ενότητα, παρουσιάζονται τα κύρια χαρακτηριστικά των δύο μοντέλων που χρησιμοποίησα, δηλαδή του συνελικτικού νευρωνικού δικτύου, και του γκαουσιανιού μοντέλου μίξης, καθώς και μερικές επιπλέον λεπτομέρειες.

2.5.1 Χαρακτηριστικά Συνελικτικού Νευρωνικού Δικτύου

Ένα συνελικτικό νευρωνικό δίκτυο, όπως προαναφέραμε, αποτελείται από μία σειρά από διαφορετικά επίπεδα. Το κάθε επίπεδο έχει και μια ξεχωριστή δουλειά που του ανατίθεται, ώστε να κάνει τα αρχεία ήχου πιο κατανοητά και πιο αναγνωρίσιμα. Σαν είσοδο δέχεται την ¨αναπαράσταση¨ του ηχητικού αρχείου, που είναι το φασματογράφημά του. Αυτό προέρχεται από την εξαγωγή των χαρακτηριστικών και συγκεκρι-

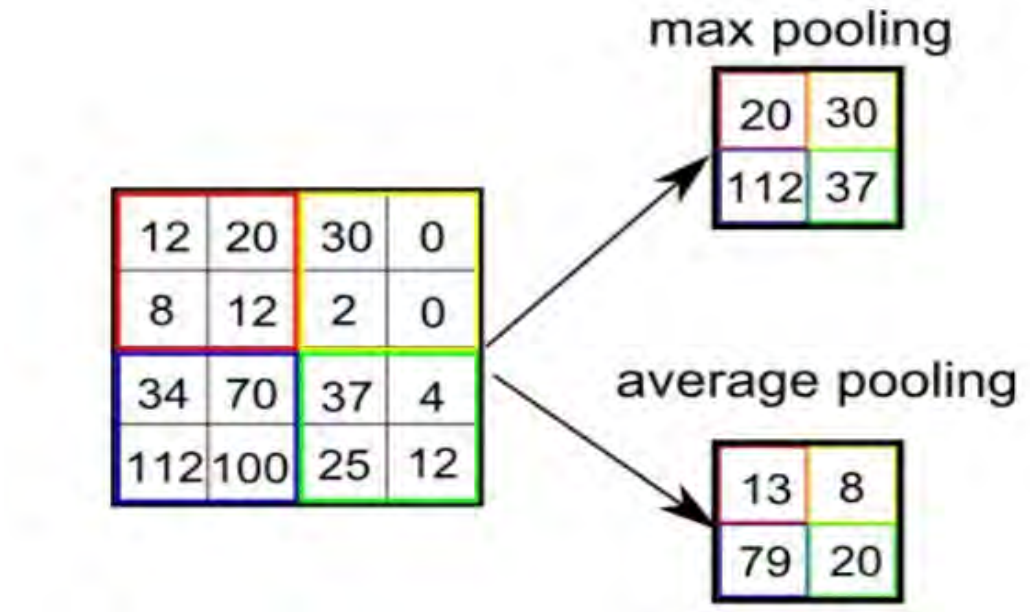
μένα από τους φασματικούς συντελεστές συχνότητας Mel. Ως έξοδο, το συνελικτικό νευρωνικό δίκτυο δίνει ένα συγκεκριμένο αποτέλεσμα, μία συγκεκριμένη κλάση. Αυτή η κλάση προκύπτει από την πρόβλεψη της σειράς των επιπέδων, από τα οποία περνάει το φασματογράφημα. Το μέγεθος του συνόλου δεδομένων, το μέγεθος των αρχείων, ο θόρυβος που υπάρχει μέσα σε αυτό, ακόμα και οι συχνότητες που αναγράφονται, μπορούν να επηρεάσουν αισθητά το τελικό αποτέλεσμα. Μπορούμε όμως να οδηγηθούμε σε όσο το δυνατόν καλύτερο ποσοστό επιτυχίας ανάλογα με τα επίπεδα που χρησιμοποιούμε στο συνελικτικό νευρωνικό δίκτυο. Αυτό απαιτεί αρκετή εμπειρία και καλή γνώση των επιπέδων αυτών. Συγκεκριμένα, τα επίπεδα αυτά είναι τα εξής:

Συνελικτικό επίπεδο: Αντί να εστιάζουμε σε ένα pixel του φασματογραφήματος κάθε φορά, το συνελικτικό επίπεδο παίρνει τετραγωνικούς πίνακες από pixels και τα περνάει μέσα από ένα φίλτρο. Το μέγεθος του φίλτρου είναι στην κρίση του προγραμματιστή. Το φίλτρο αυτό είναι επίσης τετραγωνικός πίνακας, μικρότερο από το φασματόγραμμα και ίσο με το σύνολο των pixels που εστιάζουμε κάθε φορά.



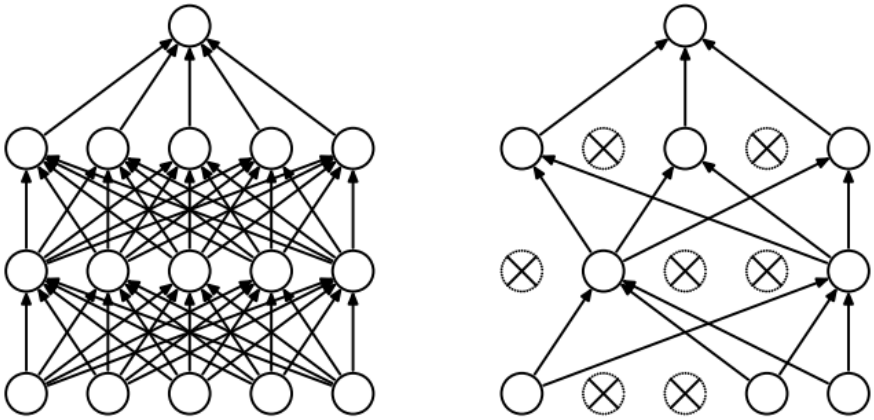
Σχήμα 2.8: Οπτική αναπαράσταση ενός συνελικτικού επιπέδου. Εδώ έχουμε μία εικόνα 30 x 30 pixels και το φίλτρο 3 x 3 pixels. Πολλαπλασιάζουμε το συγκεκριμένο κομμάτι από pixels της εικόνας με το φίλτρο που έχουμε θέσει και στο τέλος προσθέτουμε όλα τα αποτελέσματα, ώστε να καταλήξουμε στην έξοδο (output volume). Λήφθηκε από το [8]

Επίπεδο Pooling: Το επίπεδο Pooling είναι υπεύθυνο για τη μείωση του χωρικού μεγέθους των συνελιγμένων χαρακτηριστικών. Αυτό σημαίνει ότι μειώνει την υπολογιστική ενέργεια που απαιτείται για την επεξεργασία των δεδομένων, μέσω της μείωσης των διαστάσεων. Υπάρχουν δύο ήδη επιπέδων Pooling. Το πρώτο είναι το Max Pooling, το οποίο επιστρέφει τη μέγιστη τιμή από το τμήμα της εικόνας που καλύπτεται από τον πυρήνα. Το δεύτερο είναι το Average Pooling, το οποίο επιστρέφει τη μέση τιμή όλων των τιμών από το τμήμα της εικόνας που καλύπτεται από τον πυρήνα.



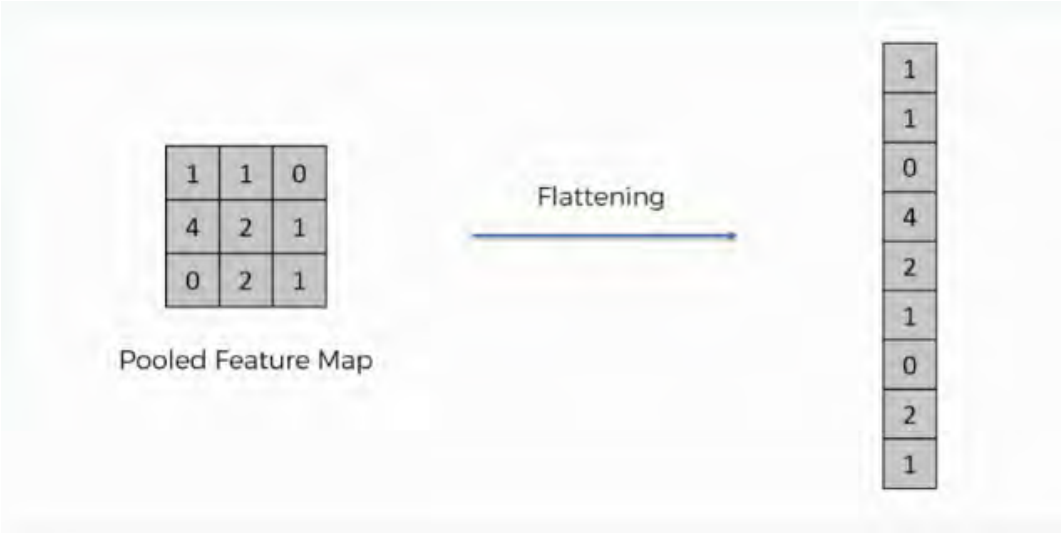
Σχήμα 2.9: Οπτική αναπαράσταση των λειτουργιών των τύπων του επιπέδου Pooling. Λήφθηκε από το [9]

Επίπεδο Dropout: Αυτό το επίπεδο είναι υπεύθυνο για να αποφεύγεται όσο το δυνατόν περισσότερο η υπερφόρτωση του δικτύου μας. Σε κάθε στάδιο εκπαίδευσης, ατομικοί κόμβοι είτε αποσύρονται από το δίκτυο με πιθανότητα 1-p, είτε μένουν με πιθανότητα p, έτσι ώστε να παραμείνει ένα μειωμένο δίκτυο. Το επίπεδο Dropout αναγκάζει ένα νευρωνικό δίκτυο να μάθει πιο ανθεκτικά χαρακτηριστικά που είναι χρήσιμα, σε συνδυασμό με πολλά διαφορετικά τυχαία υποσύνολα των άλλων νευρώνων. Σχεδόν διπλασιάζει τον αριθμό των επαναλήψεων που απαιτούνται για την σύγκλιση. Ωστόσο, ο χρόνος εκπαίδευσης για κάθε στάδιο είναι μικρότερος.



Σχήμα 2.10: Βλέπουμε δύο δίκτυα. Ένα που δεν έχει χρησιμοποιηθεί το επίπεδο Dropout (αριστερά) και ένα που έχει χρησιμοποιηθεί (δεξιά). Λήφθηκε από το [10]

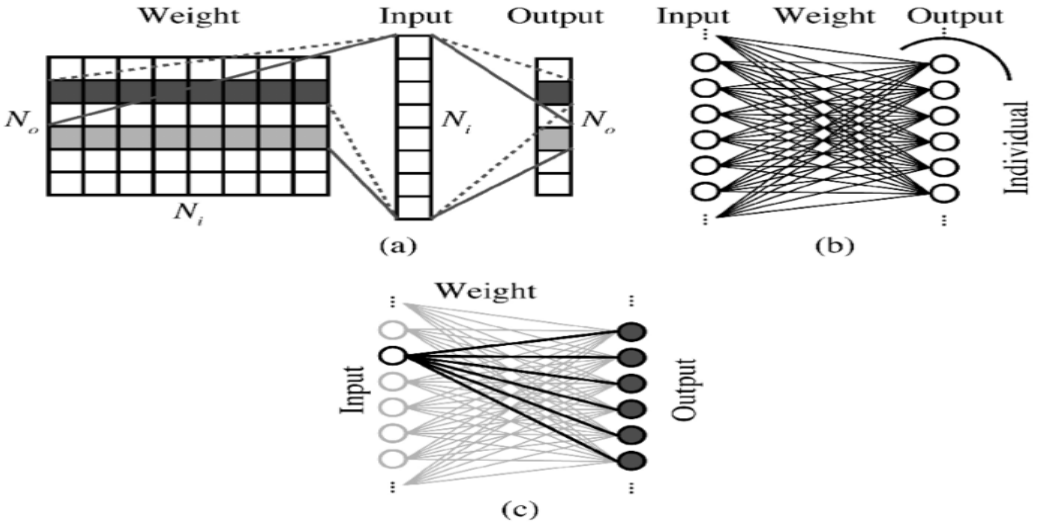
Επίπεδο ισοπέδωσης: Εδώ, έχοντας τον πίνακα εξόδου του σταδίου Pooling, τον μετατρέπουμε σε μία στήλη, δηλαδή τον κάνουμε επίπεδο. Ο κύριος λόγος που γίνεται αυτό, είναι επειδή θα χρειαστεί να εισάγουμε αυτά τα δεδομένα στο νευρωνικό δίκτυο ξανά.



Σχήμα 2.11: Οπτική αναπαράσταση της λειτουργίας του επιπέδου ισοπέδωσης. Λήφθηκε από το [11]

Πυκνό επίπεδο / Πλήρως συνδεδεμένο επίπεδο: Παρόλο που υπάρχουν δύο ονόματα για το συγκεκριμένο επίπεδο, η δουλειά τους είναι η ίδια. Το επίπεδο αυτό αφορά μία γραμμική λειτουργία, στην οποία κάθε είσοδος συνδέεται με κάθε έξοδο. Γενικά ακολουθείται από μία μη γραμμική συνάρτηση ενεργοποίησης (όπως η ανορθωμένη γραμμική μονάδα - ReLU). Ένα πλήρως συνδεδεμένο επίπεδο,

χρησιμοποιείται για να αλλάξει τις διαστάσεις του διανύσματος εισόδου. Μαθηματικά μιλώντας, εφαρμόζει περιστροφή, κλιμάκωση και μετασχηματισμό μετατόπισης στο διάνυσμα αυτό.



Σχήμα 2.12: Η δομή ενός πυκνού / πλήρως συνδεδεμένου δικτύου (a) (b) και η εξάρτηση των δεδομένων (c). Λήφθηκε από το [12]

2.5.2 Χαρακτηριστικά Γκαουσιανού Μοντέλου Μίξης

Ένα γκαουσιανό μοντέλο μίξης είναι ένα πιθανοτικό μοντέλο που υποθέτει ότι όλα τα σημεία δεδομένων παράγονται από ένα μείγμα ενός πεπερασμένου αριθμού γκαουσιανών κατανομών με άγνωστες παραμέτρους. Μπορούμε να σκεφτούμε τα μοντέλα μίξης ως γενικευμένη ομαδοποίηση για να ενσωματώσουμε πληροφορίες σχετικά με τη δομή συν-διασποράς των δεδομένων καθώς και τα κέντρα των κρυφών Γκαουσιανών.

Κεφάλαιο 3

Ανάλυση χαρακτηριστικών πρωτότυπου συστήματος

3.1 Σύνολο δεδομένων

Το σύνολο δεδομένων που χρησιμοποίησα, είναι το TUT Urban Acoustic Scenes 2018 development dataset. Αποτελείται από 24 ώρες ήχου, χωρισμένο σε 8640 τμήματα των 10 δευτερολέπτων. Σε κάθε ακουστική σκηνή αντιστοιχούν 864 τμήματα ήχου, δηλαδή 144 λεπτά. Πιο συγκεκριμένα, τα τμήματα ήχου για κάθε ακουστική σκηνή σε κάθε πόλη, φαίνονται στον παρακάτω πίνακα.

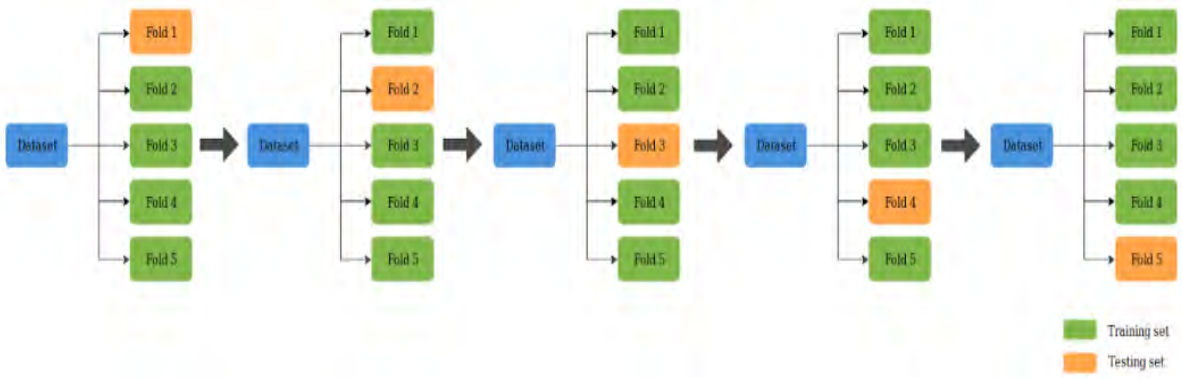
	Barcelona	Helsinki	London	Paris	Stockholm	Vienna	Total
Airport	128	149	145	156	158	128	864
Bus	144	144	144	144	144	144	864
Metro	144	144	144	144	144	144	864
Metro station	141	144	146	144	145	144	864
Park	144	144	144	144	144	144	864
Public square	144	144	144	144	144	144	864
Shopping mall	144	144	144	144	144	144	864
Street pedestrian	145	145	145	144	145	140	864
Street traffic	144	144	144	144	144	144	864
Tram	143	145	144	144	144	144	864

Πίνακας 3.1: Διαχωρισμός τμημάτων ήχου στις ακουστικές σκηνές και στις πόλεις

Σημειώνεται ότι όλες οι πληροφορίες για τα αρχεία ήχου, όπως σε ποια ακουστική σκηνή ανήκουν, ποια είναι η πηγή τους κ.α., είναι αποθηκευμένες σε αρχεία *.csv. Κάθε αλλαγή και χειρισμός των ήχων περνάει μέσα από αυτά τα αρχεία.

3.2 Διασταυρωμένη επικύρωση αποτελεσμάτων με τη μέθοδο K-fold

Η συγκεκριμένη τεχνική, χρησιμοποιείται για καλύτερο χειρισμό του συνόλου δεδομένων. Πιο συγκεκριμένα, η διασταυρωμένη επικύρωση αποτελεσμάτων με τη μέθοδο K-fold χωρίζει ένα δοσμένο σύνολο δεδομένων σε K διαφορετικά υποσύνολα, όπου το κάθε υποσύνολο θα χρησιμοποιηθεί κάποια στιγμή για εξέταση. Για παράδειγμα, αν έχουμε 5 υποσύνολα (1,2,3,4,5), τότε στην πρώτη επανάληψη θα χρησιμοποιηθούν τα (1,2,3,4) ως σύνολο μάθησης και το 5 ως σύνολο εξέτασης.



Σχήμα 3.1: Διασταυρωμένη επικύρωση αποτελεσμάτων με 5-fold. Λήφθηκε από το [13]

3.3 Πρωτότυπο σύστημα

Το πρωτότυπο σύστημα έχει υλοποιηθεί από τους δημιουργούς του DCASE 2018 challenge και συγκεκριμένα από τον Toni Heittola. Βασίζεται κατά κύριο λόγο στη βιβλιοθήκη dcase_util της γλώσσας python, η οποία περιέχει όλα τα χαρακτηριστικά για την επεξεργασία ήχου, εξαγωγή δεδομένων, τα χαρακτηριστικά ήχου, τη βιβλιοθήκη Keras, η οποία αφορά τα μοντέλα μάθησης κ.α. Το πρωτότυπο σύστημα χωρίζει το σύνολο σε 2 υποσύνολα. Το ένα είναι το υποσύνολο εκπαίδευσης του μοντέλου, το οποίο αποτελείται από το 70% του αρχικού συνόλου δεδομένων. Το άλλο είναι το υποσύνολο εκτίμησης και αποτελείται από το υπόλοιπο 30%, το οποίο επιλέχτηκε με τέτοιο τρόπο, ώστε τα δύο υποσύνολα να μην έχουν τμήματα ήχου από την ίδια τοποθεσία, αλλά να έχουν δεδομένα από κάθε πόλη. Η επίδοση του μοντέλου εκτιμάται στο υποσύνολο εκτίμησης μετά από κάθε epoch και η καλύτερη επίδοση επιλέγεται στο τέλος. Όλες οι παράμετροι και οι τεχνικές υπάρχουν σε ένα αρχείο .yaml, από το οποίο

τις παίρνει το πρωτότυπο σύστημα.

Αρχικά, ελέγχει αν υπάρχουν τα δεδομένα και αν δεν υπάρχουν, τα κατεβάζει. Στη συνέχεια, περνάει όλα τα αρχεία ήχου από τις εξής συναρτήσεις:

- **do_feature _extraction:** Η συνάρτηση αυτή παίρνει όλες τις παραμέτρους που έχουν δοθεί στο αρχείο .yaml, οι οποίες αφορούν την εξαγωγή των χαρακτηριστικών. Μέσα σε αυτά, είναι η υποδειγματοληψία, οι φασματικοί συντελεστές συχνότητας Mel, το μέγεθος βήματος, το μέγεθος και ο τύπος παραθύρου και ο τύπος του φασματογραφήματος. Υπάρχουν πολλές ακόμη παράμετροι που μπορούν να χρησιμοποιηθούν, ωστόσο το πρωτότυπο σύστημα, δεν τις χρησιμοποιεί.
- **do_normalization:** Σε αυτή τη συνάρτηση, όλα τα δεδομένα περνάνε από ένα επίπεδο κανονικοποίησης. Αυτό σημαίνει ότι γίνεται κανονικοποίηση στις συχνότητες, στο πλάτος κλπ. των αρχείων ήχου. Ο λόγος που το κάνουμε αυτό, είναι ώστε να μη διαφέρουν αρκετά αυτά τα χαρακτηριστικά των ήχων, για να είναι πιο κατανοητά στο μοντέλο μάθησης που χρησιμοποιούμε. Δεν θέλουμε να χαλάμε χρόνο και υπολογισμούς σε τέτοιες αποκλίσεις.
- **do_learning:** Η πιο σημαντική συνάρτηση, είναι αυτή της εκμάθησης του μοντέλου. Εδώ γίνονται αλλαγές στους τύπους μοντέλων, στα χαρακτηριστικά τους, στις επαναλήψεις και γενικά σε οτιδήποτε χρειάζεται να αλλάξει ώστε να έχουμε καλύτερα αποτελέσματα. Σε αυτό το σημείο χρησιμοποιείται το υποσύνολο εκπαίδευσης. Το πρωτότυπο σύστημα χρησιμοποιεί ένα συνελικτικό νευρωνικό δίκτυο, το οποίο αποτελείται αρχικά από δύο επίπεδα. Το πρώτο χρησιμοποιεί:

  1. Συνελικτικό επίπεδο, με πυρήνα 7 x 7 και αριθμό φίλτρων 32. Το συγκεκριμένο επίπεδο, επειδή είναι το πρώτο, πρέπει να περιέχει και τις διαστάσεις της εισόδου (στη συγκεκριμένη περίπτωση 40 x 500)
  2. Κανονικοποίηση παρτίδας
  3. Συνάρτηση ενεργοποίησης, με την τεχνική της διορθωμένης γραμμικής μονάδας
  4. 2D Max Pooling, μεγέθους 5 x 5

5. Dropout, με μείωση κατά 30%

Το δεύτερο χρησιμοποιεί:

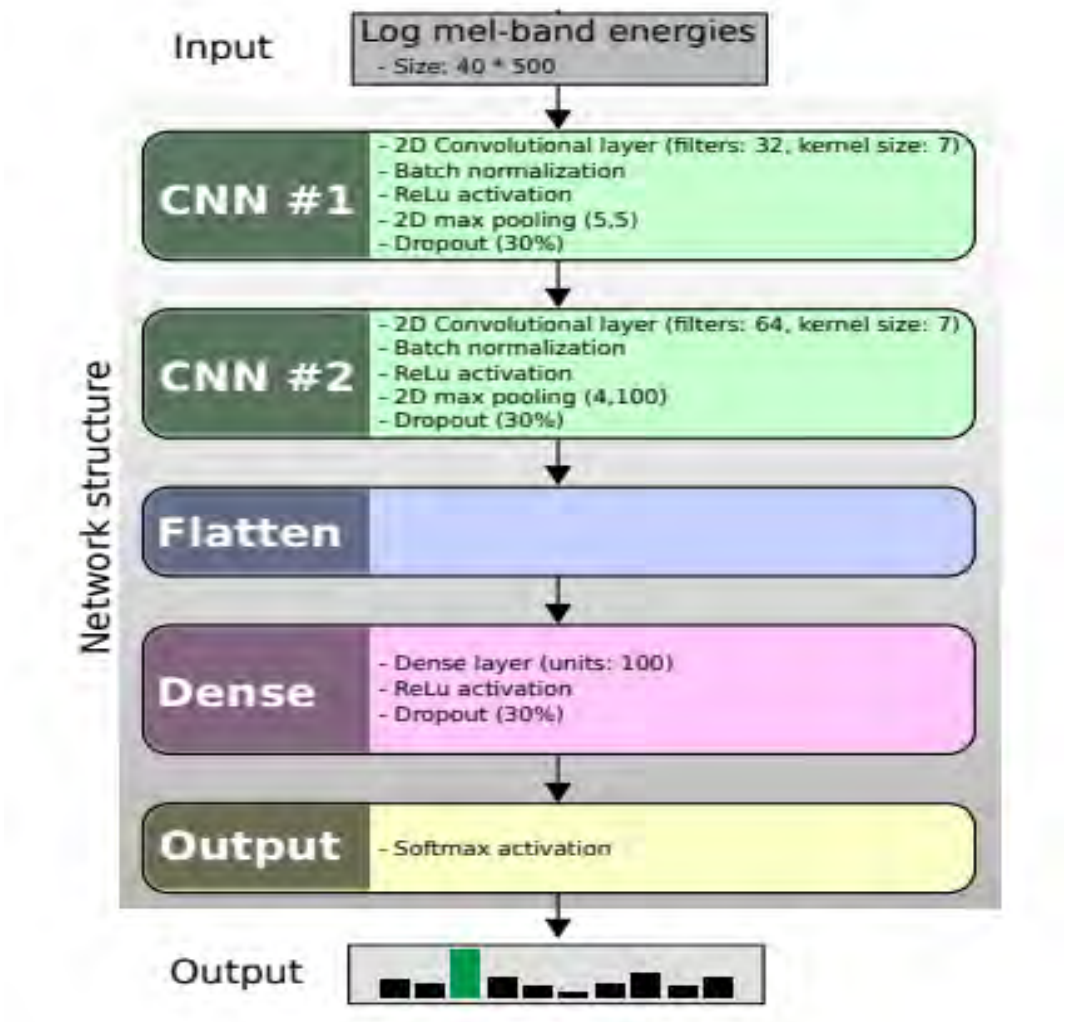
1. Συνελικτικό επίπεδο, με πυρήνα 7 x 7 και αριθμό φίλτρων 64
2. Κανονικοποίηση παρτίδας
3. Συνάρτηση ενεργοποίησης, με την τεχνική της διορθωμένης γραμμικής μονάδας
4. 2D Max Pooling, μεγέθους 4 x 100
5. Dropout, με μείωση κατά 30%

Στη συνέχεια, υπάρχει το επίπεδο Flatten, το πλήρως συνδεδεμένο επίπεδο, το οποίο αποτελείται από:

1. Πυκνό δίκτυο, με αριθμό μονάδων 100
2. Συνάρτηση ενεργοποίησης ReLU
3. Επίπεδο Dropout, με ποσοστό μείωσης 30%

Τέλος, το συνελικτικό δίκτυο του πρωτότυπου συστήματος χρησιμοποιεί συνάρτηση ενεργοποίησης με την τεχνική softmax. Αυτή η μέθοδος χρησιμοποιείται πάντα στο τέλος και μας δίνει τις πιθανότητες να υπάρχει το αρχείο ήχου σε κάθε μία από τις ακουστικές σκηνές. Πιο συγκεκριμένα, έχει ως είσοδο έναν πίνακα με αριθμούς, μετατρέπει αυτούς τους αριθμούς σε πιθανότητες με άθροισμα 1 και δίνει σαν έξοδο ένα διάνυσμα που αντιπροσωπεύει τις κατανομές πιθανοτήτων μιας λίστας πιθανών αποτελεσμάτων.

Η αρχιτεκτονική του συνελικτικού νευωρικού δικτύου, φαίνεται στο παρακάτω σχήμα.



Σχήμα 3.2: Αρχιτεκτονική πρωτότυπου συστήματος. Λήφθηκε από το [14]

- **do_testing:** Σε αυτό το στάδιο, η εκπαίδευση του μοντέλου έχει τελειώσει και περνάμε στο σημείο της εξέτασης. Σε αυτό το σημείο, χρησιμοποιείται το υποσύνολο εκτίμησης. Η διαφορά με την εκτίμηση μετά από κάθε epoch, είναι ότι τα αρχεία ήχου δε φέρουν καμία πληροφορία σχετικά με την ακουστική σκηνή στην οποία ηχογραφήθηκαν. Είναι κατά κάποιο τρόπο άγνωστα στο μοντέλο. Τα αποτελέσματα αποθηκεύονται ώστε να περάσουν στο επόμενο στάδιο.
- **do_evaluation:** Η τελευταία συνάρτηση είναι και αυτή που μας δείχνει οπτικά τα ποσοστά επιτυχίας του μοντέλου μας. Παίρνει ώς είσοδο τα αποτελέσματα από την συνάρτηση do_testing και τα αποτυπώνει στην οθόνη. Χρησιμοποιείται το εργαλείο sed_eval, το οποίο είναι βιβλιοθήκη της python. Μας δίνει το ποσοστό επιτυχίας σε κάθε ακουστική σκηνή και στο τέλος το μέσο όρο των ποσοστών αυτών.

Το σύστημα εκτελέστηκε 10 φορές και εξήγαγε τα αποτελέσματα που φαίνονται στον παρακάτω πίνακα.

Acoustic Scene	Development set
Airport	72.9%
Bus	62.9%
Metro	51.2%
Metro station	55.4%
Park	79.1%
Public square	40.1%
Shopping mall	49.6%
Street pedestrian	50.0%
Street traffic	80.5%
Tram	55.1%
Average	**59.7**(±0.7)%

Πίνακας 3.2: Αποτελέσματα πειράματος του πρωτότυπου συστήματος

Κεφάλαιο 4

Πειράματα κι αποτελέσματα

Σε αυτό το κεφάλαιο παραθέτονται όλα τα χαρακτηριστικά που χρησιμοποίησα, δηλαδή όλες οι αλλαγές που έκανα στο πρωτότυπο σύστημα. Αρχικά, ξεκίνησα με την αύξηση δεδομένων. Στη συνέχεια, έγιναν οι εξαγωγές χαρακτηριστικών. Στο πιο σημαντικό στάδιο, αυτό της εκμάθησης, έφτιαξα δύο διαφορετικά μοντέλα, ένα συνελικτικό νευρωνικό δίκτυο και ένα γκαουσιανό μοντέλο μίξης. Χρησιμοποίησα διασταυρωμένη επικύρωση αποτελεσμάτων με 5-fold.

4.1 Χαρακτηριστικά αρχείων ήχου και μοντέλων

Πιο συγκεκριμένα, τα αρχεία είναι τύπου *.wav και έχουν ρυθμό δειγματοληψίας 44.1kHz και ανάλυση 24 bits. Για τους φασματικούς συντελεστές συχνότητας Mel, χρησιμοποιώ 128 ζώνες Mel, μέγεθος FFT παραθύρου 2048, ελάχιστη συχνότητα στις ζώνες Mel 0Hz και μέγιστη 22050Hz. Στη συνέχεια, χώρισα τα αρχεία ήχου από ένα κανάλι stereo σε δύο κανάλια mono. Το πρώτο κανάλι είναι το αριστερό και το δεύτερο το δεξί. Στη συνέχεια, πρόσθεσα λευκό θόρυβο στα αρχεία και μετατόπισα τον τόνο τους. Σημειώνεται ότι δημιουργήθηκαν διαφορετικοί φάκελοι για κάθε αλλαγή. Δηλαδή, υπάρχουν συνολικά 4 φάκελοι:

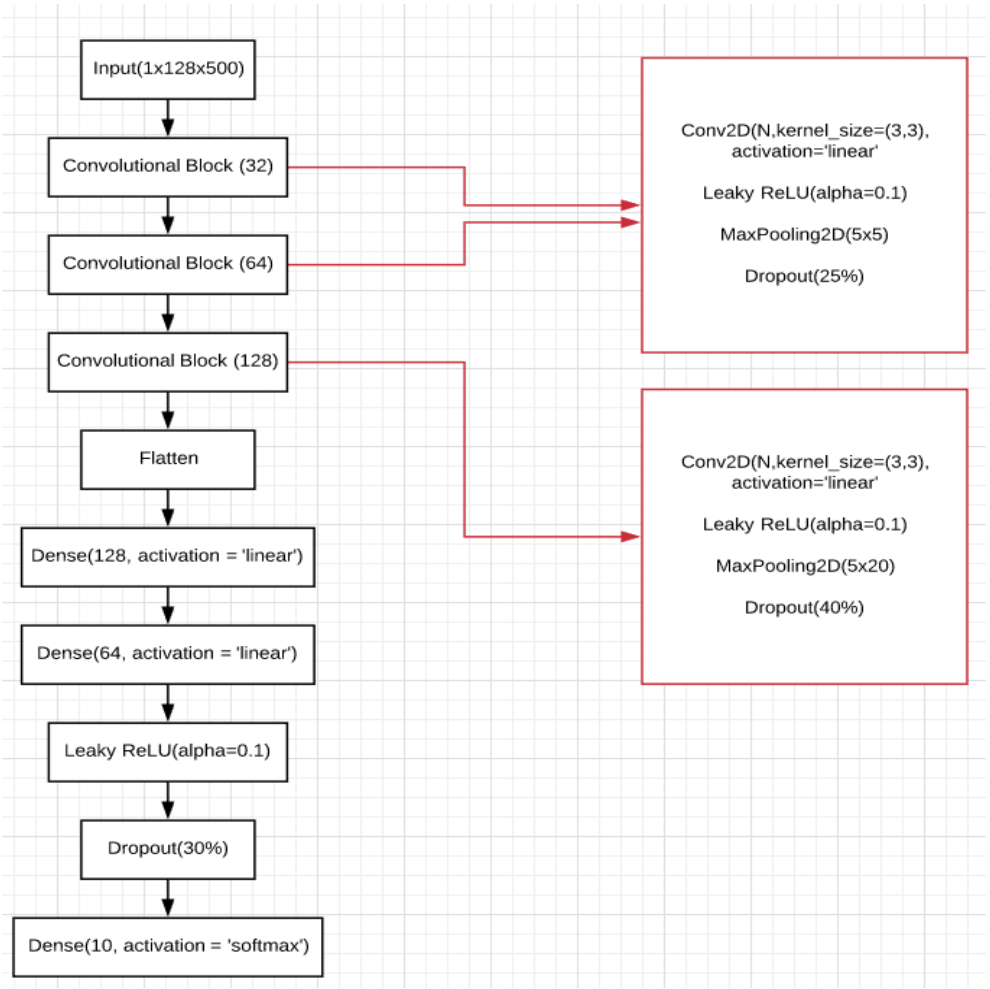
1. Αρχεία stereo
2. Αρχεία stereo με μετατόπιση τόνου.
3. Αρχεία stereo με λευκό θόρυβο.
4. Αρχεία mono χωρισμένα σε αριστερό και δεξί κανάλι

Παρακάτω, παρουσιάζονται πιο αναλυτικά τα βήματα που ακολούθησα

για τα πειράματά μου. Φαίνονται οι διαχωρισμοί, η αύξηση των δεδομένων, καθώς και η αρχιτεκτονική των μοντέλων που χρησιμοποίησα.

4.1.1 Συνελικτικό Νευρωνικό Δίκτυο

Το πρώτο μοντέλο που χρησιμοποιώ, είναι ένα συνελικτικό νευρωνικό δίκτυο. Αποτελείται από 3 συνελικτικά δίκτυα, και από τα πλήρως συνδεδεμένα επίπεδα. Η αρχιτεκτονική φαίνεται στο σχήμα παρακάτω.



Σχήμα 4.1: Αρχιτεκτονική συνελικτικού νευρωνικού δικτύου που χρησιμοποίησα.

4.1.2 Γκαουσιανό Μοντέλο μίξης

Το δεύτερο μοντέλο είναι το γκαουσιανό μοντέλο μίξης. Σε αυτό, χρησιμοποιώ 8 συνιστώσες.

4.2 Πίνακες αποτελεσμάτων

Τα αποτελέσματα κάθε μοντέλου για κάθε διαφορετικό σύνολο δεδομένων, φαίνονται στους παρακάτω πίνακες. Χρησιμοποίησα υπολογιστή με κάρτα γραφικών Gigabyte GTX 1060 6GB και μνήμη RAM 16GB. Οι χρόνοι για κάθε διαφορετικό πείραμα ήταν οι εξής:

	stereo	stereo/pitch shift	stereo/white noise	Left/Right channels
CNN	$\sim 5h$	$\sim 10h$	$\sim 10h$	$\sim 7h30m$
GMM	$\sim 7h$	$\sim 13h$	$\sim 13h$	$\sim 12h$

Πίνακας 4.1: Χρόνοι εκπαίδευσης που χρειάστηκαν για να τρέξει το κάθε μοντέλο χρησιμοποιώντας το κάθε σύνολο δεδομένων

4.2.1 Συνελικτικό νευρωνικό δίκτυο με stereo σύνολο δεδομένων

Αυτός ο συνδυασμός ήταν ο αρχικός, σε αυτόν δηλαδή που έκανα όλες τις αλλαγές στα αρχεία ήχου και στην αρχιτεκτονική του συνελικτικού νευρωνικού δικτύου μου. Μετά από αρκετές αλλαγές, κατέληξα στο σύστημα που αναφέρεται προηγουμένως. Το τελικό αποτέλεσμα είναι 68.4% επιτυχία. Το καλύτερο Fold ήταν το δεύτερο με ποσοστό 70.5%. Η ακουστική σκηνή με τη μεγαλύτερη επιτυχία ήταν η park με 86.1%, ενώ η σκηνή με τη μικρότερη επιτυχία ήταν η metro με 52.3%.

Scene	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Airport	60.2%	66.3%	69.8%	83.1%	69.8%	69.8%
Bus	75.6%	81.6%	66.7%	80.5%	94.8%	79.8%
Metro	57.2%	44.5%	46.8%	49.7%	63.4%	52.3%
Metro station	46.4%	63.8%	46.6%	59.2%	54.6%	54.1%
Park	74.4%	81.6%	90.8%	93.7%	90.2%	86.1%
Public square	63.1%	73.0%	72.4%	50.0%	44.8%	60.7%
Shopping mall	86.3%	67.2%	72.4%	78.7%	47.1%	70.4%
Street pedestrian	39.0%	79.2%	73.4%	52.6%	61.8%	61.2%
Street traffic	85.1%	82.2%	89.7%	92.5%	75.9%	85.1%
Tram	62.5%	66.1%	75.3%	61.5%	58.0%	64.7%
Average	65.0%	70.5%	70.4%	70.2%	66.1%	68.4%

Πίνακας 4.2: Συνελικτικό νευρωνικό δίκτυο / stereo

4.2.2 Συνελικτικό νευρωνικό δίκτυο με stereo/pitch shift σύνολο δεδομένων

Σε αυτόν τον συνδυασμό, είχα 63.0% επιτυχία. Το καλύτερο Fold ήταν το τρίτο με ποσοστό 66.0%. Η ακουστική σκηνή με τη μεγαλύτερη επιτυχία ήταν η street_traffic με 81.6%, ενώ η σκηνή με τη μικρότερη επιτυχία ήταν η metro με 45.4%. Παρατηρούμε αύξηση στη σκηνή airport, ωστόσο, στις υπόλοιπες υπάρχει μείωση.

Scene	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Airport	70.5%	70.3%	85.2%	69.5%	69.2%	72.9%
Bus	69.6%	75.9%	71.8%	78.2%	89.7%	77.0%
Metro	47.4%	51.7%	42.2%	38.2%	47.7%	45.4%
Metro station	51.8%	66.7%	44.5%	47.7%	48.6%	51.9%
Park	72.9%	74.1%	91.1%	88.8%	73.9%	80.2%
Public square	52.1%	53.4%	50.9%	39.1%	48.6%	48.8%
Shopping mall	89.9%	64.7%	66.7%	82.2%	40.2%	68.7%
Street pedestrian	41.0%	60.7%	56.1%	35.0%	42.5%	47.0%
Street traffic	78.0%	82.8%	83.0%	92.5%	71.8%	81.6%
Tram	62.5%	51.1%	69.0%	37.6%	61.8%	56.4%
Average	63.6%	65.1%	66.0%	60.9%	59.4%	63.0%

Πίνακας 4.3: Συνελικτικό νευρωνικό δίκτυο / stereo - pitch shift

4.2.3 Συνελικτικό νευρωνικό δίκτυο με stereo/white noise σύνολο δεδομένων

Σε αυτόν τον συνδυασμό παρατηρούμε τα μικρότερα ποσοστά σε σχέση με τα υπόλοιπα. Είναι χαμηλότερο και από το πρωτότυπο. Εδώ παρουσιάζει 54.2% επιτυχία με καλύτερο Fold το τέταρτο με 56.6%. Η ακουστική σκηνή με τη μεγαλύτερη επιτυχία ήταν η airport με 85.1% με σημαντική αύξηση από τις δύο προηγούμενες υλοποιήσεις, ενώ η σκηνή με τη μικρότερη επιτυχία ήταν η street_pedestrian με 21.5%.

Scene	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Airport	83.2%	80.5%	94.2%	83.1%	84.3%	85.1%
Bus	68.5%	62.9%	70.1%	77.3%	89.1%	73.6%
Metro	41.3%	45.1%	40.8%	34.1%	45.9%	41.4%
Metro station	46.7%	41.1%	41.4%	31.3%	39.9%	40.1%
Park	62.5%	60.1%	78.4%	87.9%	80.7%	73.9%
Public square	30.7%	37.6%	29.9%	19.8%	15.5%	26.7%
Shopping mall	70.5%	47.4%	52.0%	62.4%	37.9%	54.0%
Street pedestrian	14.8%	18.2%	24.0%	20.5%	30.1%	21.5%
Street traffic	80.1%	74.1%	77.3%	88.8%	76.7%	79.4%
Tram	42.0%	31.3%	48.6%	60.3%	47.7%	46.0%
Average	54.0%	49.8%	55.7%	56.6%	54.8%	54.2%

Πίνακας 4.4: Συνελικτικό νευρωνικό δίκτυο / stereo - white noise

4.2.4 Συνελικτικό νευρωνικό δίκτυο με σύνολο δεδομένων χωρισμένο σε αριστερό και δεξί κανάλι

Εδώ τα ποσοστά είναι καλύτερα απο τα δύο προηγούμενα συστήματα, αλλά χειρότερα από το stereo. Ο συνδυασμός αυτός έχει ποσοστό επιτυχίας 65.2% με καλύτερο Fold το τρίτο με 67.7%. Η ακουστική σκηνή με τη μεγαλύτερη επιτυχία ήταν η street_traffic με 85.2%, ενώ η σκηνή με τη μικρότερη επιτυχία ήταν η metro_station με 50.6%.

Scene	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Airport	66.5%	73.8%	85.2%	83.1%	76.7%	77.1%
Bus	51.5%	83.3%	74.4%	68.4%	92.5%	74.0%
Metro	61.6%	40.5%	51.2%	57.2%	74.1%	56.9%
Metro station	47.6%	64.4%	46.0%	49.7%	45.1%	50.6%
Park	69.9%	75.6%	87.9%	83.9%	77.9%	79.0%
Public square	46.1%	59.8%	54.3%	41.1%	59.8%	52.2%
Shopping mall	75.3%	67.2%	62.9%	75.0%	61.5%	68.4%
Street pedestrian	51.7%	44.2%	54.3%	55.5%	48.0%	50.8%
Street traffic	83.9%	89.1%	88.5%	91.1%	73.6%	85.2%
Tram	62.2%	48.6%	72.7%	52.6%	54.3%	58.1%
Average	61.6%	64.6%	67.7%	65.8%	66.4%	65.2%

Πίνακας 4.5: Συνελικτικό νευρωνικό δίκτυο / Left-Right channels

4.2.5 Γκαουσιανό μοντέλο μίξης με stereo σύνολο δεδομένων

Το επόμενο μοντέλο είναι το γκαουσιανό μοντέλο μίξης. Σε αυτόν τον συνδυασμό έχουμε επιτυχία 78.36% με καλύτερο Fold το δεύτερο με 82.09%. Η ακουστική σκηνή με τη μεγαλύτερη επιτυχία ήταν η street_traffic με 91.29%, ενώ η σκηνή με τη μικρότερη επιτυχία ήταν η metro_station με 58.27%.

Scene	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Airport	82.39%	81.40%	83.14%	88.37%	76.74%	82.41%
Bus	85.71%	93.10%	82.18%	81.03%	95.40%	87.49%
Metro	83.82%	83.82%	74.57%	75.72%	83.72%	80.33%
Metro station	66.07%	55.17%	63.79%	55.75%	50.57%	58.27%
Park	76.79%	91.95%	94.83%	87.36%	93.68%	88.92%
Public square	66.67%	79.31%	74.14%	49.43%	38.51%	61.61%
Shopping mall	79.76%	73.56%	81.03%	86.21%	70.11%	78.14%
Street pedestrian	54.07%	81.50%	81.50%	87.86%	74.57%	75.90%
Street traffic	87.50%	94.25%	92.53%	95.40%	86.78%	91.29%
Tram	68.45%	86.78%	85.06%	85.06%	70.69%	79.21%
Average	75.12%	82.09%	81.28%	79.22%	74.08%	78.36%

Πίνακας 4.6: Γκαουσιανό μοντέλο μίξης / stereo

4.2.6 Γκαουσιανό μοντέλο μίξης με stereo/pitch shift σύνολο δεδομένων

Ο συνδυασμός αυτός παρουσιάζει ποσοστά μικρότερα απο τον προηγούμενο. Το ποσοστό επιτυχίας του είναι 65.62%. Το Fold με το μεγα-

λύτερο ποσοστό είναι το δεύτερο με 66.69%. Η ακουστική σκηνή με τη μεγαλύτερη επιτυχία ήταν η park με 87.19%, ενώ η σκηνή με τη μικρότερη επιτυχία ήταν η metro_station με 46.30%.

Scene	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Airport	59.94%	63.95%	66.28%	61.92%	59.59%	62.34%
Bus	73.21%	76.15%	58.33%	66.95%	84.77%	71.88%
Metro	69.65%	67.05%	54.62%	60.40%	62.79%	62.90%
Metro station	47.32%	48.85%	45.40%	41.09%	48.85%	46.30%
Park	76.49%	89.94%	89.66%	85.63%	94.25%	87.19%
Public square	55.06%	59.48%	59.77%	35.06%	28.45%	47.56%
Shopping mall	84.52%	69.83%	66.95%	79.31%	72.13%	74.55%
Street pedestrian	35.76%	53.76%	59.83%	73.12%	61.27%	56.75%
Street traffic	85.12%	81.03%	86.21%	92.53%	83.62%	85.70%
Tram	56.25%	56.90%	68.39%	67.53%	56.03%	61.02%
Average	64.33%	66.69%	65.54%	66.35%	65.18%	65.62%

Πίνακας 4.7: Γκαουσιανό μοντέλο μίξης / stereo-pitch shift

4.2.7 Γκαουσιανό μοντέλο μίξης με stereo/white noise σύνολο δεδομένων

Σε αυτό το συνδυασμό, και πάλι έχουμε μικρά ποσοστά επιτυχίας, όπως και στο συνελικτικο νευρωνικό δίκτυο με λευκό θόρυβο. Το ποσοστό επιτυχίας είναι 57/09% με καλύτερο Fold το δεύτερο με 58.28%. Η ακουστική σκηνή με τη μεγαλύτερη επιτυχία ήταν η bus με 78.51%, ενώ η σκηνή με τη μικρότερη επιτυχία ήταν η metro_station με 40.39%.

Scene	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Airport	49.43%	75.58%	72.67%	68.90%	58.72%	65.06%
Bus	76.19%	84.48%	63.22%	72.41%	96.26%	78.51%
Metro	68.21%	63.29%	45.09%	56.65%	47.97%	56.24%
Metro station	31.55%	47.70%	41.67%	36.49%	44.54%	40.39%
Park	41.96%	49.71%	94.25%	89.08%	49.43%	64.89%
Public square	60.42%	58.91%	35.92%	31.61%	31.61%	43.69%
Shopping mall	88.69%	35.06%	48.85%	41.95%	60.63%	55.04%
Street pedestrian	27.03%	41.91%	43.93%	41.33%	51.16%	41.07%
Street traffic	75.30%	70.98%	65.23%	90.80%	75.29%	75.52%
Tram	50.00%	55.17%	59.20%	45.69%	42.24%	50.46%
Average	56.88%	58.28%	57.00%	57.49%	55.78%	57.09%

Πίνακας 4.8: Γκαουσιανό μοντέλο μίξης / stereo - white noise

4.2.8 Γκαουσιανό μοντέλο μίξης με σύνολο δεδομένων χωρισμένο σε αριστερό και δεξί κανάλι

Ο συγκεκριμένος συνδυασμός έχει σχεδόν ίδια ποσοστά με το stereo σύνολο δεδομένων. Το καλύτερο Fold ήταν το τρίτο με ποσοστό επιτυχίας 77.36%. Η ακουστική σκηνή με τη μεγαλύτερη επιτυχία ήταν η street_traffic με 89.41%, ενώ η σκηνή με τη μικρότερη επιτυχία ήταν η metro_station με 55.59%.

Scene	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Airport	74.15%	81.69%	77.62%	83.72%	74.13%	78.26%
Bus	83.63%	92.82%	79.89%	82.12%	95.40%	86.78%
Metro	75.14%	80.06%	68.50%	75.72%	87.50%	77.38%
Metro station	60.12%	58.91%	55.75%	50.86%	52.30%	55.59%
Park	80.06%	86.21%	92.24%	87.36%	94.54%	88.08%
Public square	63.99%	75.00%	71.84%	54.60%	42.24%	61.53%
Shopping mall	70.54%	61,21%	74.14%	80.46%	72.99%	71.87%
Street pedestrian	48.84%	77.17%	76.88%	80.35%	74.28%	71.50%
Street traffic	89.29%	86.78%	90.52%	93.97%	86.49%	89.41%
Tram	68.45%	70.69%	86.21%	79.02%	63.79%	73.63%
Average	71.42%	77.05%	77.36%	76.82%	74.37%	75.40%

Πίνακας 4.9: Γκαουσιανό μοντέλο μίξης / Left-Right channels

Κεφάλαιο 5

Συμπεράσματα και δυσκολίες

5.1 Συμπεράσματα

Καθώς έχουν σημειωθεί τα αποτελέσματα, σε αυτό το κεφάλαιο γίνεται η σύγκρισή τους. Δίνονται διαγράμματα σχετικά με τα ποσοστά που επιτεύχθηκαν σε κάθε μοντέλο για κάθε κατηγορία συνόλου δεδομένων. Συνολικά, έχουμε αποτελέσματα από 8 διαφορετικές υλοποιήσεις.

Αρχικά, θα γίνει σύγκριση των μοντέλων που χρησιμοποιήθηκαν. Ως αποτέλεσμα, θα χρησιμοποιηθούν τα τελικά ποσοστά επιτυχίας για κάθε σύνολο δεδομένων. Αυτά, φαίνονται στον παρακάτω πίνακα.

	stereo	stereo/pitch shift	stereo/white noise	Left/Right channels
CNN	68.4%	63.0%	54.2%	65.2%
GMM	78.36%	65.62%	57.02%	75.40%

Πίνακας 5.1: Τελικά ποσοστά για τα δύο μοντέλα για κάθε σύνολο δεδομένων

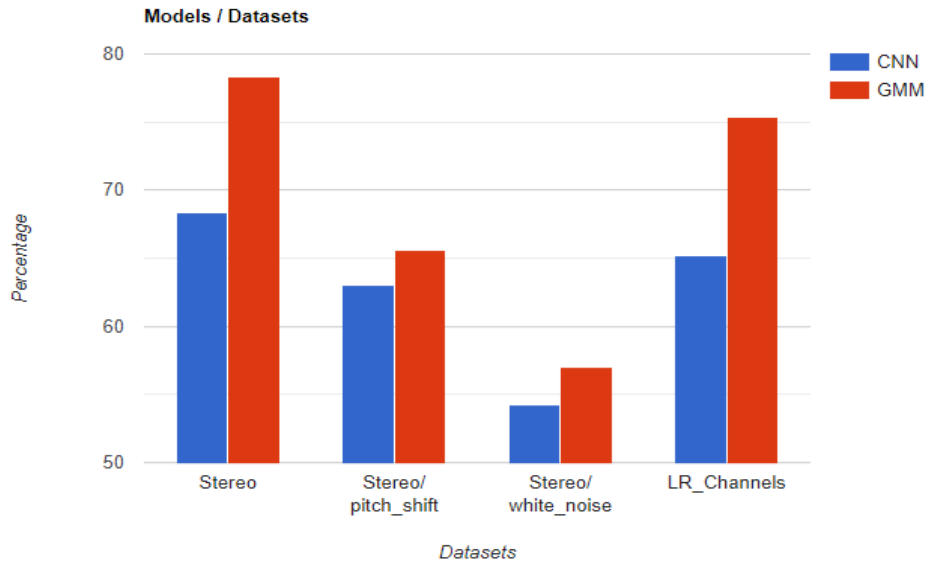
Σε αυτόν τον πίνακα, παρατηρούμε αρχικά, πως το γκαουσιανό μοντέλο μίξης λειτουργεί καλύτερα για τα συγκεκριμένα χαρακτηριστικά ήχου από το συνελικτικό νευρωνικό δίκτυο. Φαίνεται πως σε όλα τα σύνολα δεδομένων, παρουσιάζει μεγαλύτερα ποσοστά. Ωστόσο, καταναλώνει περισσότερο χρόνο από το συνελικτικό νευρωνικό δίκτυο.

Έπειτα, παρατηρούμε πως στην κατηγορία του συνόλου με τον λευκό θόρυβο, και τα δύο μοντέλα πέφτουν αρκετά χαμηλά και πιο συγκεκριμένα, χαμηλότερα από το πρωτότυπο σύστημα (59.7%). Αυτό σημαίνει πως, παρότι η πρόσθεση λευκού θορύβου λειτουργεί σε κάποια πειράματα, στο συγκεκριμένο υπολειτουργεί.

Επίσης, παρατηρούμε οτι και στα δύο μοντέλα, τα καλύτερα ποσοστά

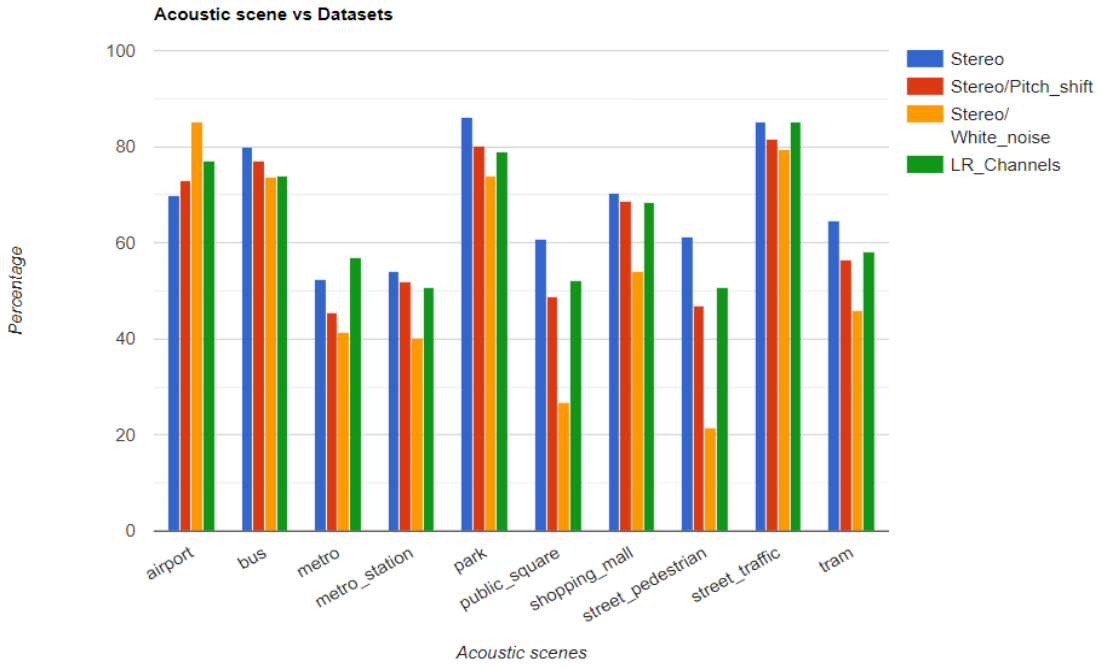
παρουσιάζονται με το κανονικό σύνολο δεδομένων, που δίνεται απο το DCASE. Παρακάτω, αναπαριστώνται τα ποσοστά αυτά σε διαγράμματα, για την καλύτερη κατανόηση της σύγκρισής τους.

Στο πρώτο διάγραμμα, γίνεται σύγκριση των μοντέλων που χρησιμοποιήθηκαν σε σχέση με τα σύνολα δεδομένων.



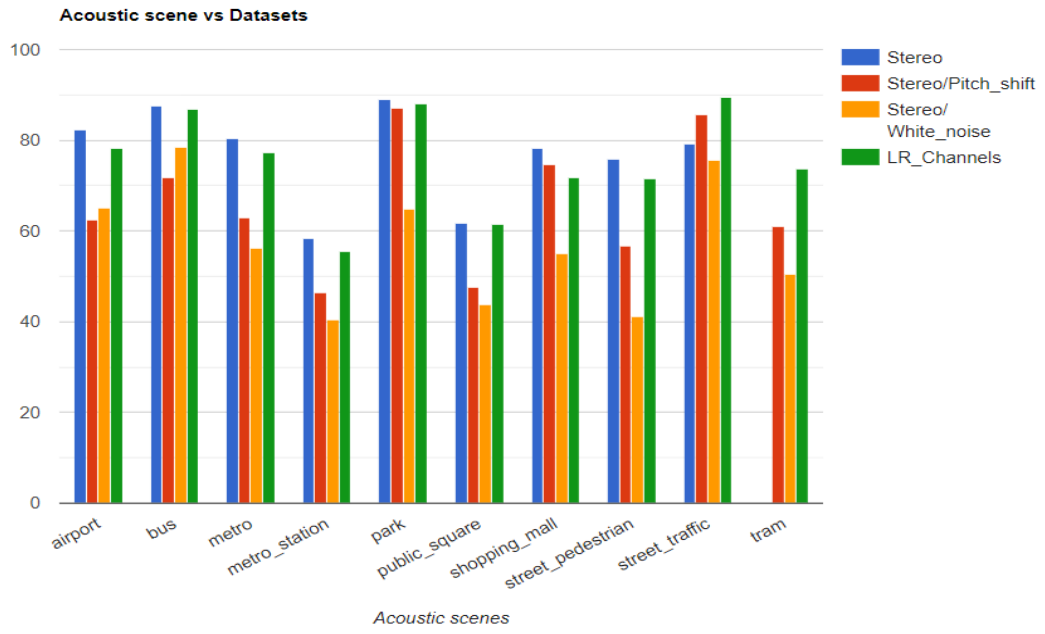
Σχήμα 5.1: Διάγραμμα αποτελεσμάτων μοντέλων / συνόλων δεδομένων. Δημιουργήθηκε από το [15]

Στο επόμενο διάγραμμα, γίνεται σύγκριση των αποτελεσμάτων με βάση τις ακουστικές σκηνές για κάθε σύνολο δεδομένων στο συνελικτικό νευρωνικό δίκτυο.



Σχήμα 5.2: Διάγραμμα αποτελεσμάτων ακουστικών σκηνών / συνόλων δεδομένων στο συνελικτικό νευρωνικό δίκτυο. Δημιουργήθηκε από το [15]

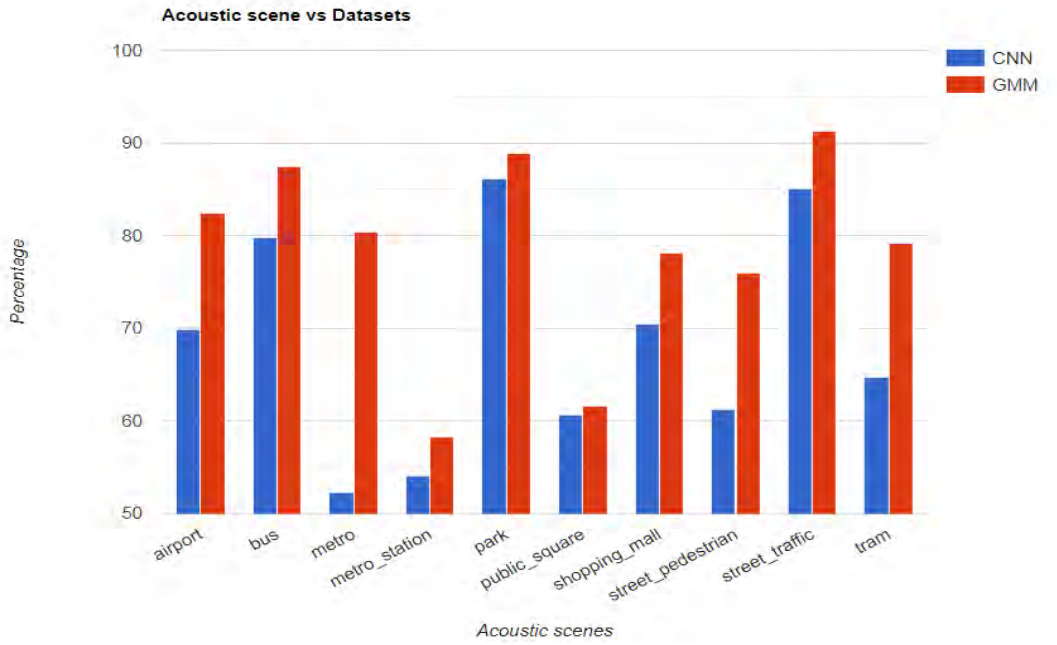
Παρακάτω, φαίνονται τα ποσοστά για κάθε ακουστική σκηνή με τη χρήση των συνόλων δεδομένων στο γκαουσιανό μοντέλο μίξης.



Σχήμα 5.3: Διάγραμμα αποτελεσμάτων ακουστικών σκηνών / συνόλων δεδομένων στο γκαουσιανό μοντέλο μίξης. Δημιουργήθηκε από το [15]

Τέλος, γίνεται σύγκριση των μοντέλων με βάση τις ακουστικές σκηνές,

στο πιο επιτυχημένο σύνολο δεδομένων, δηλαδή το stereo.



Σχήμα 5.4: Διάγραμμα αποτελεσμάτων μοντέλου / ακουστικών σκηνών στο σύνολο δεδομένων stereo. Δημιουργήθηκε από το [15]

5.2 Δυσκολίες

Αρχικά, η πρώτη δυσκολία που αντιμετώπισα σχετικά με τις υλοποιήσεις μου, ήταν η περίπλοκη διαχείρηση του πρωτότυπου κώδικα του DCASE 2018. Ο κώδικας αυτός χρησιμοποιεί την βιβλιοθήκη dcase_util, από την οποία διαχειρίζεται το μεγαλύτερο μέρος των ενεργειών του (εξαγωγή χαρακτηριστικών, κανονικοποίηση, εκπαίδευση μοντέλου κ.α.). Ήταν, λοιπόν, περίπλοκη και χρονοβόρα η εξοικείωση με τη συγκεκριμένη βιβλιοθήκη.

Πριν την μεταχείριση του πρωτότυπου κώδικα του DCASE 2018, σκοπός μου ήταν να φτιάξω από την αρχή ένα δικό μου μοντέλο, με όσα ήδη γνώριζα. Αυτό, φυσικά, ήταν ως ένα σημείο πιο εύκολο, διότι διαχειριζόμουν βιβλιοθήκες που ήταν ήδη γνωστές. Ωστόσο, δυσκολεύτηκα αρκετά να εισάγω την βιβλιοθήκη sed_eval, την οποία έπρεπε να χρησιμοποιήσω για την εκτίμηση της εκπαίδευσης των μοντέλων μου. Έτσι, θα είχα μία σύγκριση με τα αποτελέσματα των ομάδων που δούλεψαν πάνω στη συγκεκριμένη εργασία.

Το σημείο που με δυσκόλεψε περισσότερο ήταν η αρχιτεκτονική των

μοντέλων μου. Καθώς άλλαζα συνεχώς τις παραμέτρους για να εξάγω καλύτερα αποτελέσματα, έπρεπε να τρέχω τον κώδικά μου αρκετές φορές. Κάθε τρέξιμο απαιτούσε αρκετό χρόνο και πολλές φορές δεν υπήρχε κάποια βελτίωση.

Κεφάλαιο 6

Μελλοντικές κατευθύνσεις έρευνας

6.1 Ανασκόπηση της διπλωματικής

Σε αυτή τη διπλωματική διερευνήθηκε το πρόβλημα της κατάταξης αρχείων ήχου σε δοσμένες ακουστικές σκηνές. Για τον σκοπό αυτό χρησιμοποιήθηκαν τα αρχεία απο το DCASE 2018 - Task1A. Πάνω σε αυτά τα αρχεία έγιναν διάφορες αλλαγές ως προς την αρχιτεκτονική του μοντέλου, την εξαγωγή των χαρακτηριστικών των αρχείων ήχου κ.α.

Η εργασία ξεκίνησε με την περιγραφή του προβλήματος, καθώς και με την ανασκόπηση της λειτουργίας του DCASE. Στη συνέχεια, δόθηκε το απαραίτητο θεωρητικό υπόβαθρο όλων των λειτουργιών και χαρακτηριστικών που χρησιμοποιήθηκαν. Έγινε μία εκτενής αναφορά στα νευρωνικά δίκτυα και στους φασματικούς συντελεστές συχνότητας Mel.

Στο πειραματικό σκέλος της εργασίας, έγιναν αλλαγές με τη σειρά στα εξής:

1. Εξαγωγή χαρακτηριστικών των αρχείων ήχου
2. Διασταυρωμένη επικύρωση αποτελεσμάτων με τη μέθοδο K-fold
3. Μοντέλα που χρησιμοποιήθηκαν
4. Αρχιτεκτονική του κάθε μοντέλου

Έπειτα, παρουσιάστηκαν σε πίνακες τα αποτελέσματα όλων των συνδυασμών των μοντέλων με τα διάφορα σύνολα δεδομένων. Έγινε σχολιασμός για κάθε πίνακα, δηλαδή σε ποια ακουστική σκηνή βρίσκονται τα καλύτερα και τα χειρότερα ποσοστά, καθώς και σύγκριση των αποτελε-

σμάτων με τους υπόλοιπους πίνακες. Επιπροσθέτως, παρουσιάστηκαν διαγράμματα για την καλύτερη κατανόηση των ποσοστών. Έγινε, δηλαδή, μία οπτική αναπαράσταση των συγκρίσεων των μοντέλων και των ακουστικών σκηνών.

Τέλος, βγήκαν διάφορα συμπεράσματα για την πειραματική αυτή εργασία και αναφέρθηκαν όσες δυσκολίες αντιμετώπισα στην παρούσα διπλωματική.

6.2 Μελλοντικές κατευθύνσεις έρευνας

Στη συγκεκριμένη διπλωματική εργασία, τα αποτελέσματα των πειραμάτων που παρουσιάστηκαν ήταν κυρίως θετικά. Ωστόσο, υπάρχουν αρκετά περιθώρια βελτίωσης των ποσοστών. Ορισμένες μελλοντικές κατευθύνσεις αναγράφονται παρακάτω:

- Υλοποίηση διαφορετικών ταξινομητών, πέρα από το συνελικτικό νευρωνικό δίκτυο και το γκαουσιανό μοντέλο μίξης.
- Εξαγωγή διαφορετικών χαρακτηριστικών, καθώς και εφαρμογή διαφορετικών αρχιτεκτονικών των μοντέλων
- Εφαρμογή της μεθόδου συνόλου, η οποία δέχεται περισσότερα από ένα μοντέλα, τα συνδυάζει και στο τέλος εξάγει αποτελέσματα με βάση την τακτική απόφασης. Η συγκεκριμένη υλοποίηση χρησιμοποιήθηκε από τις περισσότερες ομάδες που συμμετείχαν στο διαγωνισμό του DCASE. Παρόλο που είναι περίπλοκη μέθοδος, βοηθάει πολύ στην εξαγωγή των αποτελεσμάτων.
- Εφαρμογή των μοντέλων που υλοποιήθηκαν σε διαφορετικά σύνολα δεδομένων, με διαφορετικά χαρακτηριστικά μηχανημάτων εγγραφής.

Bibliography

[1] "Classification of general audio data for content-based retrieval." [Online]. Available: https://www.researchgate.net/figure/The-organization-of-tools-for-audio-feature-extraction_fig1_222666693

[2] "Python, pitch shifting, and the pianoputer." [Online]. Available: http://zulko.github.io/blog/2014/03/29/soundstretching-and-pitch-shifting-in-python/

[3] "The dummy's guide to MFCC." [Online]. Available: https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

[4] "A-guide-mel-frequency-cepstral-coefficients-mfccs." [Online]. Available: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[5] "From fiction to reality: A beginner's guide to artificial neural networks." [Online]. Available: https://towardsdatascience.com/from-fiction-to-reality-a-beginners-guide-to-artificial-neural-networks-d041

[6] "Enhanced environmental sound classification with a CNN." [Online]. Available: https://medium.com/ai%C2%B3-theory-practice-business/enhanced-environmental-sound-classification-with-a-cnn-1ca388748bc9

[7] "Fit Gaussian mixture model to data." [Online]. Available: https://www.mathworks.com/help/stats/fitgmdist.html

[8] "A beginner's guide to convolutional neural networks (CNNs)." [Online]. Available: https://skymind.ai/wiki/convolutional-network

[9] "A comprehensive guide to convolutional neural networks—the eli5 way." [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd

[10] "Dropout: A simple way to prevent neural networks from overfitting." [Online]. Available: http://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf

[11] "Convolutional neural networks (CNN): Step 3 - flattening." [Online]. Available: https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-step-3-flattening

[12] "Convolutional neural networks (CNN): Step 1(b) - ReLU layer." [Online]. Available: https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-step-1b-relu-layer/

[13] "K-fold cross validation." [Online]. Available: https://medium.com/datadriveninvestor/k-fold-cross-validation-6b8518070833

[14] "A multi-device dataset for urban acoustic scene classification." [Online]. Available: https://arxiv.org/pdf/1807.09840.pdf

[15] "Bar graph maker." [Online]. Available: https://www.rapidtables.com/tools/bar-graph.html

[16] "Acoustic scene classification: An overview of DCASE 2017 challenge entries." [Online]. Available: http://www.cs.tut.fi/~mesaros/pubs/mesaros-iwaenc2018-asc-in-dcase2017.pdf

[17] "Investigation of acoustic and visual features for acoustic scene classification." [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419300661

[18] "Acoustic scene classification." [Online]. Available: http://dcase.community/challenge2019/task-acoustic-scene-classification

[19] "Acoustic scene classification with fully convolutional neural networks and i-vectors." [Online]. Available: http://dcase.community/documents/challenge2018/technical_reports/DCASE2018_Dorfer_97.pdf

[20] "Artificial neural network." [Online]. Available: https://en.wikipedia.org/wiki/Artificial_neural_network

[21] "Finite-state machine." [Online]. Available: https://en.wikipedia.org/wiki/Finite-state_machine

[22] “Convolutional neural network.” [Online]. Available: https://en.wikipedia.org/wiki/Convolutional_neural_network

[23] “Gaussian mixture model.” [Online]. Available: https://www.geeksforgeeks.org/gaussian-mixture-model/

[24] “Gaussian mixture model.” [Online]. Available: https://brilliant.org/wiki/gaussian-mixture-model/