

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ**

**ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ**

**ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ**

**Τίτλος διπλωματικής εργασίας στα Ελληνικά**

**στοιχισμένος στο κέντρο**

Διπλωματική Εργασία

Ονοματεπώνυμο φοιτητή

Επιβλέπων: Ονοματεπώνυμο επιβλέποντος

Μήνας έτος



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ**

**ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ**

**ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ**

**Τίτλος διπλωματικής εργασίας στα Ελληνικά**

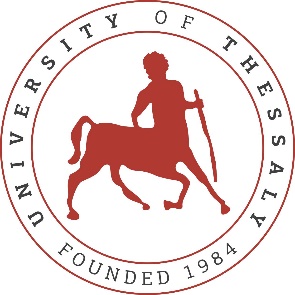
**στοιχισμένος στο κέντρο**

Διπλωματική Εργασία

Ονοματεπώνυμο φοιτητή

Επιβλέπων: Ονοματεπώνυμο επιβλέποντος

Μήνας έτος



**UNIVERSITY OF THESSALY**

**SCHOOL OF ENGINEERING**

**DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING**

**Title of the Diploma Thesis in English**

**centered**

Diploma Thesis

Name of student

Supervisor: Name of supervisor

Month year

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων/πουσα **Ονοματεπώνυμο Επιβλέποντα**

Βαθμίδα/ιδιότητα επιβλέποντα, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Ονοματεπώνυμο Μέλους 1**

Βαθμίδα/ιδιότητα μέλους 1, Τμήμα/Ίδρυμα μέλους 1

Μέλος **Ονοματεπώνυμο Μέλους 2**

Βαθμίδα/ιδιότητα μέλους 2, Τμήμα/Ίδρυμα μέλους 2

**ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ**

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελούν αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλουν οποιασδήποτε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχουν έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο/Η Δηλών/ούσα

Ονοματεπώνυμο Φοιτητή/ήτριας

**DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS**

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

The Declarant

Name of Student

# Ευχαριστίες ή Σχόλια

Θα ήθελα να ευχαριστήσω τους γονείς μου και την οικογένειά μου που με βοήθησαν και με υποστήριξαν Σελίδα με ευχαριστίες ή σχόλια του συγγραφέα (προαιρετικό)

Διπλωματική Εργασία

**Τίτλος διπλωματικής εργασίας**

**(όπως στο εξώφυλλο)**

Ονοματεπώνυμο φοιτητή

# Περίληψη

Το αντικείμενο έρευνας της διπλωματικής αφορά το πρόβλημα της αναγνώρισης ηχητικών συμβάντων και της τοποθέτησής τους στο χώρο. Ο σκοπός αυτής της διπλωματικής εργασίας είναι πάνω στην έρευνα και στην ανάπτυξη ενός συστήματος ανίχνευσης ηχητικών συμβάντων και εύρεσης της θέσης τους στο χώρο με τη χρήση τεχνητής νοημοσύνης και νευρωνικών δικτύων, το οποίο να μπορεί να ξεπεράσει σε ακρίβεια το baseline μοντέλο βάσης του διαγωνισμού DCASE 2021. Η ιδέα, το αρχικό νευρωνικό μοντέλο, η εξαγωγή των χαρακτηριστικών των αρχείων ήχου καθώς και ο τρόπος καταμέτρησης της ακρίβειας του κάθε μοντέλου βασίστηκε στο φετινό διαγωνισμό DCASE 2021 TASK 3. Όσον αφορά τη δομή της συγκεκριμένης διατριβής, αρχικά θα επεξηγηθούν ορισμοί και θεωρητικές βάσεις αναγκαίες για την αντιμετώπιση και κατανόηση του προβλήματος. Στη συνέχεια θα αναπτυχθεί το βασικό μοντέλο baseline που παρέχεται από το διαγωνισμό DCASE 2021 καθώς και η μορφή της εισόδου και της προ-επεξεγασίας των αρχείων ήχου και θα γίνει αναφορά στο τρόπο που αξιολογούνται τα αποτελέσματα και θα επεξηγηθεί πως γίνεται ο υπολογισμός της ακρίβειας ενός μοντέλου. Τέλος θα γίνει παρουσίαση των νευρωνικών δικτύων και τεχνικών που αναπτύχθηκαν στα πλαίσια της εργασίας για το σκοπό της εργασίας και θα σχολιασθούν διάφορες ακόμα τεχνικές και πειράματα για τη βελτίωση των αποτελεσμάτων. Το υπόλοιπο μέρος της διπλωματικής εργασίας είναι η παρουσίαση των αποτελεσμάτων και τα τελικά συμπεράσματα.

**Λέξεις-κλειδιά:**

Ήχος, αναγνώριση ηχητικών συμβάντων, νευρωνικά δίκτυα, τεχνητή νοημοσύνη

Diploma Thesis

**Title of Diploma thesis**

**(as in cover page)**

Name of student

# Abstract

(Usually less than a page)

**Keywords:**

Stand-alone hybrid system; HOMER Pro; simulation-optimization; techno-economic

analysis; excess electricity percentage; Net Present Cost (NPC); Levelizez Cost of Energy (LCoE)

# **Πίνακας περιεχομένων**

[Ευχαριστίες ή Σχόλια xiii](#_Toc76067406)

[Περίληψη xv](#_Toc76067407)

[Abstract xvii](#_Toc76067408)

[Πίνακας περιεχομένων xix](#_Toc76067409)

[Κατάλογος εικόνων xxi](#_Toc76067410)

[Κατάλογος σχημάτων xxiii](#_Toc76067411)

[Κατάλογος πινάκων xxv](#_Toc76067412)

[Συντομογραφίες xxvii](#_Toc76067413)

[Κεφάλαιο 1 Εισαγωγή 1](#_Toc76067414)

[1.1 Αντικείμενο της διπλωματικής 1](#_Toc76067415)

[1.1.1 Συνεισφορά 1](#_Toc76067416)

[1.2 Οργάνωση του τόμου 2](#_Toc76067417)

[Κεφάλαιο 2 Τίτλος κεφαλαίου 3](#_Toc76067418)

[2.1 Εισαγωγή 3](#_Toc76067419)

[2.2 Ορισμός Δυναμικής Τιμολόγησης 3](#_Toc76067420)

[2.3 Μεταβαλλόμενη τιμή της ενέργειας 4](#_Toc76067421)

[2.3.1 Τίτλος Υπο-ενότητας 5](#_Toc76067422)

[2.3.2 Τίτλος Υπο-ενότητας 6](#_Toc76067423)

[2.3.3 Τίτλος Υπο-ενότητας 6](#_Toc76067424)

[Κεφάλαιο 3 Τίτλος κεφαλαίου 7](#_Toc76067425)

[3.1 This us test 7](#_Toc76067426)

[3.2 Τίτλος ενότητας 7](#_Toc76067427)

[Κεφάλαιο 4 Τίτλος κεφαλαίου 9](#_Toc76067428)

[4.1 Λειτουργία του Virtual Budget 9](#_Toc76067429)

[4.1.1 Πρόβλεψη Φορτίου 10](#_Toc76067430)

[4.1.2 Πρόβλεψη της τιμής 11](#_Toc76067431)

[4.2 Αναλυτική Προσέγγιση της Μεθόδου (VB) 12](#_Toc76067432)

[4.3 Τίτλος ενότητας 13](#_Toc76067433)

[Κεφάλαιο 5 Κ.λπ. 15](#_Toc76067434)

[5.1 Τίτλος ενότητας 15](#_Toc76067435)

[Κεφάλαιο 6 - Συμπεράσματα (ή κάτι παρόμοιο) 17](#_Toc76067436)

[6.1 Σύνοψη και συμπεράσματα 17](#_Toc76067437)

[6.2 Μελλοντικές επεκτάσεις 17](#_Toc76067438)

[Βιβλιογραφία 19](#_Toc76067439)

[ΠΑΡΑΡΤΗΜΑΤΑ 21](#_Toc76067440)

[Παράρτημα Α 23](#_Toc76067441)

[Τίτλος παραρτήματος 23](#_Toc76067442)

[Α.1 Βασικές αρχές 23](#_Toc76067443)

[Α.2 Επιπλέον παρατηρήσεις 23](#_Toc76067444)

[Α.3 Παραδοχές 24](#_Toc76067445)

[Α.4 Θεωρητικό υπόβαθρο 24](#_Toc76067446)

[Παράρτημα Β 27](#_Toc76067447)

[Τίτλος παραρτήματος 27](#_Toc76067448)

# Κατάλογος εικόνων

[Εικόνα 4.1: Γραμμές μεταφοράς 9](#_Toc76067449)

# Κατάλογος σχημάτων

[Σχήμα 2.1: Διακύμανση της τιμής της ηλεκτρικής ενέργειας 5](file:///C:\Users\user\Desktop\back_32gb%20flasaki%2020201226\THMMY%20PPS\DIPLOMATIKES\DIPLOMATIKES%20THMMY\kanonismos%20diplomatikon\Diplomatiki%20Ergasia%20-%20Kanonismoi%20-%20Ypodeigmata%20-%20Templates\ΣΥΓΓΡΑΦΗ%20ΔΙΠΛΩΜΑΤΙΚΗΣ%20ΕΡΓΑΣΙΑΣ%20ΣΕ%20WORD%20(ΠΠΣ)%20-%20July%202021%20-%20ΥΠΟΔΕΙΓΜΑ.docx#_Toc76067450)

[Σχήμα 4.1:-Απεικόνιση λειτουργίας του Virtual Budget [7] 10](#_Toc76067451)

[Σχήμα 4.2: Επίπεδα της μορφοποίησης του φορτίου με τη μέθοδο VB και χρήση των παραμέτρων β,γ 13](#_Toc76067452)

# Κατάλογος πινάκων

Πίνακας 2.1: Ημερήσιες τιμές ηλεκτρικής ενέργειας σε KW της εταιρείας PJM 5

Πίνακας 4.1: Τιμές ηλεκτρικής ενέργειας σε KW 11

# Συντομογραφίες

βλπ βλέπε

κ.λπ. και λοιπά

κ.ο.κ και ούτω καθεξής

CNN Convolutional Neural Network

# Εισαγωγή

## Αντικείμενο της διπλωματικής

Ο ήχος μπορεί να περιέχει αρκετή χρήσιμη πληροφορία. Μπορούμε με αυτόν να καταλάβουμε καλύτερα το περιβάλλον (αναγνώριση σκηνών) και τι συμβαίνει σε αυτό (αναγνώριση πηγών), το οποίο δίνει δυνατότητα για την ανάπτυξη μεγάλου εύρους εφαρμογών, όπως στο τομέα αυτοκινούμενων αμαξιδίων, στο τομέα της ρομποτικής και σε έξυπνα συστήματα ασφαλείας, όπου μπορεί να αναγνωρίζει ανεπιθύμητα ηχητικά συμβάντα και να ειδοποιεί τον χρήστη. Η αναγνώριση ηχητικών συμβάντων (SED) και ο εντοπισμός τους στο χώρο (DOA) αποτελεί έναν ενεργό κλάδο έρευνας ο οποίος έχει προσελκύσει το ενδιαφέρον του επιστημονικού χώρου, ιδιαίτερα τα τελευταία χρόνια. Η αποτελεσματική έρευνα στο αντικείμενο μπορεί να αποβεί χρήσιμη σε διάφορους τομείς για την ανάπτυξη της επιστήμης, αλλά ακόμα και για τη καθημερινή ζωή.

Για την προσέγγιση και ανάπτυξη ενός συστήματος ανίχνευσης ηχητικών συμβάντων, απαιτείται βαθιά γνώση σε τομείς σημάτων και επεξεργασίας ήχου. Η λύση του προβλήματος αυτού μέχρι τώρα γινόταν με διάφορα μοντέλα ταξινόμησης βασισμένα στη μηχανική μάθηση (SVM, Random Forest, KNN, Gaussian Mixture Model κ.λπ.) [1], όμως, τα τελευταία χρόνια, ο τρόπος προσέγγισης του θέματος κυρίως για προβλήματα πολλαπλών κλάσεων (multi-label classification) πραγματεύεται κυρίως με λύσεις που έχουν βάση στη Βαθιά Μάθηση, η οποία αποτελεί υποτομέα της Μηχανικής Μάθησης και πιο συγκεκριμένα στη τεχνική των νευρωνικών δικτύων. Η χρήση αυτών των υπολογιστικών συστημάτων για την επίλυση του προβλήματος αυτού έχει δείξει ότι είναι αρκετά επιτυχής από τεχνικές μηχανικής μαθήσεως, και ειδικά με τη ραγδαία ανάπτυξη της τεχνολογίας μια τέτοια προσέγγιση πλέον είναι προτιμότερη. Αρκετές μελέτες σε αυτό το κλάδο έχουν αναπτύξει συστήματα νευρωνικών μοντέλων, τα οποία πολλές φορές είναι πιο ακριβή και από τον άνθρωπο. Τα πιο επιτυχημένα μοντέλα σε αυτό το τομέα είναι τα Βαθιά Αναδρομικά Συνελικτικά δίκτυα ( Deep CRNN) τα οποία παρά τη χρήση τους για προβλήματα επεξεργασίας εικόνων έχουν αποβεί αποτελεσματικά και στο πρόβλημα αναγνώρισης και εντοπισμού πηγής ηχητικών συμβάντων (SELD). Τα μοντέλα αυτά πλεονεκτούν των απλών αλγορίθμων ταξινόμησης, καθώς εμφανίζονται στο προσκήνιο νέα προβλήματα όπως SELD σε περιβάλλον με επικαλυπτόμενα συμβάντα, όπου απαιτείται πιο περίπλοκος χειρισμός και έξυπνη ταξινόμηση,.

Ωστόσο, υπάρχει ακόμα αρκετή ανάγκη για περαιτέρω μελέτη και έρευνα, καθώς σε πραγματικά περιβάλλοντα τα συστήματα αυτά δε λειτουργούν με πλήρη ακρίβεια και δε δίνουν ικανοποιητικά αποτελέσματα. Αυτό συμβαίνει, καθώς βασικό πρόβλημα της αναγνώρισης ηχητικών συμβάντων είναι κυρίως η απομόνωσή του κάθε γεγονότος από άλλα που τυχόν μπορεί να προκύπτουν και να το επικαλύπτουν.

### Συνεισφορά

Η συγκεκριμένη διπλωματική εργασία αποτέλεσε έναυσμα για τη δημιουργία ενός συστήματος ταξινόμησης ηχητικών συμβάντων και τοποθέτησής τους στο χώρο με χρήση νευρωνικών δικτύων. Στόχος της αποτελεί την εύρεση ενός τέτοιου συστήματος που να ξεπερνάει σε ακρίβεια το baseline μοντέλο του διαγωνισμού DCASE 2021 task 3.

Παρουσιάζεται η έρευνα πάνω σε διάφορες παραμέτρους και τεχνικές που επηρεάζουν την ακρίβεια στα αποτελέσματα ενός μοντέλου, όπως τεχνικές αύξησης των δεδομένων εισόδου (data augmentation), αλλαγή των αποκωδικοποιητών του δικτύου (decoders), καθώς και τα τελικά μοντέλα που αναπτύχθηκαν για τον αναφερόμενο.

Πιο συγκεκριμένα, αναπτύχθηκαν τρία (3) μοντέλα, η αρχιτεκτονική των οποίων βασίστηκε στη Βαθιά Μάθηση και στα συνελικτικά νευρωνικά δίκτυα (CNN). Το πρώτο μοντέλο χρησιμοποιεί την αρχιτεκτονική του μοντέλου ResNet, το οποίο χρησιμοποιείται συχνά σε προβλήματα κατηγοριοποίησης εικόνων. Θα αναφερθεί η σημασία της αρχιτεκτονικής τους και στο πρόβλημα κατηγοριοποίησης ηχητικών συμβάντων καθώς και η σημασία του μεγέθους και του αριθμού στρωμάτων ενός τέτοιου μοντέλου και πως επηρεάζει τα αποτελέσματα. Ως δεύτερο μοντέλο, γίνεται η χρήση μοντέλου Conformer, το οποίο αποτελεί τη τελευταία σε προβλήματα ακολουθιών κειμένου και σε εφαρμογές μετάφρασης αλλά έχουν αποδειχθεί αποτελεσματικά και σε κατηγοριοποίηση multi-label ηχητικών συμβάντων. Τέλος, ερευνήθηκε η μέθοδος Ensembling τριών μοντέλων CRNN. Και στα τρία (3) μοντέλα επιλέχθηκε η ταυτόχρονη πρόβλεψη του SED και DOA με χρήση του ACCDOA output format [2, 3].

## Οργάνωση του τόμου

Στο **Κεφάλαιο 1** γίνεται μια εισαγωγή στο πρόβλημα της αναγνώρισης ηχητικών γεγονότων και παρουσιάζονται οι λόγοι που οδηγούν στην ανάπτυξή νευρωνικών δικτύων για την επίλυσή του.

Στο **Κεφάλαιο 2** γίνεται αναφορά στο θεωρητικό υπόβαθρό αναγκαίο για την αντιμετώπιση του προβλήματος.

Στο **Κεφάλαιο 3** περιγράφεται τα αρχεία εισόδου, η δομή της αρχιτεκτονικής του βασικού μοντέλου και των κανόνων του DCASE 2021 καθώς και τεχνικές προ-επεξεργασίας, μετα-επεξεργασίας και παραμέτρων που χρησιμοποιεί

Στο **Κεφάλαιο 4** παρουσιάζονται τα μοντέλα αρχιτεκτονικής που χρησιμοποιήθηκαν για την εύρεση ενός καλύτερου συστήματος από του baseline, καθώς και τεχνικές για επιπλέον αποτελεσματικότητα.

Στο **Κεφάλαιο 5** παρουσιάζονται συγκεντρωτικά τα κυριότερα αποτελέσματα πειραμάτων της παρούσας εργασίας, οι λόγοι πρωτοτυπίας της έρευνας και της προσέγγισης, η συμβολή της στη διερεύνηση του θέματος και οι δυσκολίες που υπήρξαν στην ανάπτυξή του

Στο **Κεφάλαιο 6** αναφέρονται τα συμπεράσματα της διατριβής και πιθανές προτάσεις για περαιτέρω διερεύνηση του θέματος στο μέλλον.

# Θεωρητικό Υπόβαθρο

## Επεξεργασία ήχου

Ο ήχος αποτελεί σήμα το οποίο κρύβει μέσα του αρκετή χρήσιμη πληροφορία στο πεδίο της συχνότητας αλλά και του χρόνου. Ωστόσο έχει παρατηρηθεί ότι το πρόβλημα ταξινόμησης ήχου είναι αρκετά όμοιο με αυτό της ταξινόμησης εικόνων, συνεπώς είναι πολύ πιο εύκολο να εξάγουμε κάποιο συμπέρασμα ή κάποια πληροφορία από τα χαρακτηριστικά του ήχου αν τον οπτικοποιήσουμε και εφαρμόσουμε σε αυτό μεθόδους και τεχνικές επεξεργασίας εικόνας. Θα αναφέρουμε παρακάτω πως γίνεται η μετατροπή ενός ηχητικού σήματος σε mel-Spectogram για την επεξεργασία του και πως επίσης γίνεται η εξαγωγή των χαρακτηριστικών του ήχου.

Για την εξαγωγή των χαρακτηριστικών που παίρνουμε από ηχητικά σήματα εισόδου και για την ψηφιακή αναπαράσταση τους, γίνεται δειγματοληψία τους σε συγκεκριμένη συχνότητα δειγμάτων ανά χρονική μονάδα. Έτσι μετατρέπεται το σήμα σε διακριτό, αναπαρίσταται δηλαδή από μια ακολουθία δειγμάτων, το οποίο το καθιστά εύκολο για ανάλυση. Όμως, ένα ηχητικό σήμα κρύβει αρκετή πληροφορία και στο πεδίο της συχνότητας. Για να μπορεί να γίνει αποτελεσματική ανάλυση και εξαγωγή της πληροφορίας ενός σήματος, χρειάζεται να πληρεί τρεις προϋποθέσεις:

1. Να γνωρίζουμε την χρονο-συχνοτική του αναπαράσταση.
2. Να έχουμε λογαριθμικά αντιληπτό τρόπο αναπαράστασης του ύψους της συχνότητας.
3. Να έχουμε λογαριθμικά αντιληπτό τρόπο αναπαράστασης των διαστημάτων των συχνοτήτων.

Για να ικανοποιηθεί η πρώτη προϋπόθεση, χρησιμοποιείται ο **μετασχηματισμός Φουριέ**, ο οποίος μετασχηματίζει μέσω μαθηματικών υπολογισμών μια εξίσωση που περιγράφεται από το πεδίο του χρόνου στο πεδίου της συχνότητας. [4, 5, 6].

Η αναπαράσταση μιας σειράς Φουριέ πεπερασμένου μήκους καλείται Διακριτός Μετασχηματισμός Φουριέ (DFT). Ο DFT είναι ένας αλγόριθμος επεξεργασίας σημάτων, ο οποίος διακριτοποιεί ένα περιοδικά μη χρονικά μεταβαλλόμενο σήμα εισόδου πεπερασμένης διαρκείας. Ουσιαστικά είναι αναπαράσταση μιας ακολουθίας πεπερασμένης διάρκειας από μια περιοδική ακολουθία εισόδου, όπου μια περίοδος της είναι η πεπερασμένη ακολουθία. Μία παραλλαγή του DFT είναι ο Fast Fourier Transform (FFT), ο οποίος δίνει πιο γρήγορα αποτελέσματα και με λιγότερους υπολογισμούς.

Diagram

Description automatically generated

Μετασχηματισός Φουριέ 1

Εικόνα 1: Οπτικοποίηση μετατροπής ενός σήματος από το πεδίου του χρόνου στο πεδίο της συχνότητας.

Όμως τα περισσότερα σήματα δεν είναι περιοδικά, όπως και τα αρχεία εισόδου που χρησιμοποιεί η συγκεκριμένη εργασία. Για το σκοπό ανάλυσης τέτοιων χρονικά μεταβαλλόμενων σημάτων χρησιμοποιείται ο αλγόριθμος Short Time Fourier Transform (STFT). Ο αλγόριθμος αυτός ουσιαστικά διαχωρίζει το σήμα εισόδου σε μικρότερα επικαλυπτόμενα ή μη τμήματα, τα οποία πολλαπλασιάζονται με κάποια εξίσωση «παραθύρου», συνήθως παραθύρου Hann, και στη συνέχεια εφαρμόζει τον FFT επαναλαμβανόμενα σε αυτά. Μια οπτική αναπαράσταση της διαδικασίας φαίνεται στην Εικόνα 1.

Diagram

Description automatically generated

Εικόνα 2: Ο αλγόριθμος STFT.

## Mel-Spectogram

Με τη χρήση του STFT στο σήμα εισόδου, εξάγουμε το spectrogram του. Το spectrogram είναι μια οπτική αναπαράσταση ενός σήματος, στην οποία φαίνεται η σχέση της πληροφορίας του στο πεδίο του χρόνου και της συχνότητας. Ένα παράδειγμα spectrogram φαίνεται στη παρακάτω Εικόνα , όπου ο άξονας x είναι οι τιμές στο πεδίο χρόνου, ο άξονας y στο πεδίο της συχνότητας και η αλλαγή στην ένταση του χρώματος μαρτυρά το μέγεθος του σήματος που βρίσκεται στη συγκεκριμένη συχνότητα και χρονική στιγμή.

Ωστόσο, η κλίμακα συχνοτήτων δεν αλλάζει γραμμικά προς την ανθρώπινη αντίληψη. Οι άνθρωποι έχουμε λογαριθμική αντίληψη του ήχου, δηλαδή καταλαβαίνουμε μεγαλύτερη διαφορά σε διαστήματα χαμηλότερων συχνοτήτων παρά σε υψηλότερων. Για να έχουμε μια πιο γραμμική κλίμακα συχνοτήτων, χρησιμοποιείται η κλίμακα Mel, εφεύρεση των Stevens, Volkmann και Newmann το 1937, στην οποία το διάστημα μεταξύ γειτονικών συχνοτήτων είναι το ίδιο αντιληπτό στο ανθρώπινο αυτί. [7, 8, 9]

A screenshot of a computer

Description automatically generated with medium confidenceΕφαρμόζοντας την κλίμακα mel στο spectrogram ενός σήματος, έχουμε το αποκαλούμενο mel-spectogram, με το οποίο μπορεί να γίνει και η ανάλυση του σήματος, ολοκληρώνοντας και την τρίτη προϋπόθεση.

Spectogram. 1

Εικόνα 3: Spectogram.

## Βαθιά μάθηση και νευρωνικά δίκτυα

Τα τελευταία χρόνια η επιστημονική κοινότητα ασχολείται με τον τομέα της Βαθιάς Μαθήσεως, η οποία αποτελεί υποκατηγορία ενός πιο ευρύ επιστημονικού πεδίου, αυτό της Τεχνητής Νοημοσύνης. Η Τεχνητή Νοημοσύνη προσπαθεί να βρει λύση στο ερώτημα που απασχολεί την ανθρωπότητα από την αρχαιότητα, το οποίο είναι αν μπορεί να υπάρξει κάποια έξυπνη ανεξάρτητη μηχανή που να μπορεί να μιμηθεί τον ανθρώπινο λογισμό. Ο τομέας της Βαθιάς Μάθησης, ιδιαίτερα ανεπτυγμένος τα τελευταία χρόνια, ασχολείται με την ανάπτυξη αλγορίθμων και μαθηματικών μοντέλων με σκοπό ένας υπολογιστής μπορεί να εκτελεί ανθρωπίνως δυνατές λειτουργίες που απαιτούν κριτική σκέψη, όπως η κατηγοριοποίησή αντικειμένων ή η εκμάθηση κάποιου στόχου. Βασικές εφαρμογές της είναι η πρόβλεψη και δημιουργία νέων δεδομένων μέσω των αλγορίθμων αυτών από δοσμένα γνωστά δεδομένα. Μερικές από τις τεχνικές που χρησιμοποιεί είναι τα νευρωνικά δίκτυα. Σε αυτό το κεφάλαιο θα αναφερθούμε σε δομές νευρωνικών δικτύων που προτιμούνται και χρησιμοποιούνται ευρέως σε προβλήματα ταξινόμησης και κατηγοριοποίησης ήχων καθώς και στα πλαίσια της διατριβής, όπως τα CNN, ResNet, Dense στρώματα καθώς και Transformers.

### Νευρωνικά δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (ANN) είναι υπολογιστικά συστήματα εμπνευσμένα από τη Βιολογία και λειτουργούν όπως και οι βιολογικοί νευρώνες σε έναν ανθρώπινο εγκέφαλο. Αποτελούνται από σειρές από ανεπτυγμένους αλγορίθμους οι οποίοι μπορούν και εκτελούν υψηλού επιπέδου λειτουργίες, όπως η κατηγοριοποίησή και εύρεση σχέσεων μεταξύ δοσμένων.

Diagram

Description automatically generatedΌπως ο βιολογικός νευρώνας δέχεται ως είσοδο έναν ερεθισμό, το οποίοι στη συνέχεια μετατρέπεται σε ένα μικρό ηλεκτρικό σήμα και τέλος μετατρέπεται σε κίνηση, έτσι και ο τεχνητός δέχεται ως είσοδο κάποια δεδομένα και μέσω κάποιων «βαρών» και εξισώσεων ενεργοποίησης τα μετατρέπει σε καινούρια επεξεργασμένα δεδομένα. Η σύνδεση πολλών νευρώνων δημιουργούν ένα νευρωνικό δίκτυο και όπως και ένας ανθρώπινος εγκέφαλος αποτελείται από ένα δίκτυο βιολογικών νευρώνων που επικοινωνούν μεταξύ τους, έτσι ένα τεχνητό νευρωνικό δίκτυο αποτελείται από πολλούς «νευρώνες» οι οποίοι είναι οργανωμένοι σε στρώματα τα οποία συνδέονται μεταξύ τους ανταλλάσσοντας και μετατρέποντας μια δοσμένη πληροφορία.

Εικόνα 4: Οπτικοποίηση της κοινής δομής ενός βιολογικού νευρώνα (αριστερά) και ενός τεχνητού (δεξιά).

Πιο συγκεκριμένα, ένας τεχνητός νευρώνας πολλαπλασιάζει την είσοδο του, η οποία έιναι συνήθως κάποιο διάνυσμα που αντιπροσωπεύει κάποια πληροφορία, με κάποιες τιμές που λέγονται βάρη. Έπειτα το άθροισμα των γινομένων των βαρών με την είσοδο περνάει από μία εξίσωση ενεργοποίησης ώστε να δώσει ένα τελικό αποτέλεσμα (output). Το αποτέλεσμα αυτό μπορεί να περαστεί ως είσοδος στο επόμενο συνδεδεμένο στρώμα του δικτύου και η διαδικασία συνεχίζεται . Η γενική εξίσωση ενός νευρώνα είναι:

Η εξίσωση ενεργοποίησης επιλέγεται με βάση τη λειτουργία το δικτύου και κάποιες από αυτές που χρησιμοποιήθηκαν και στα πλαίσια της διπλωματικής εργασίας είναι:

* ReLu:

Ιδιαίτερα χρήσιμη για Βαθιά νευρωνικά δίκτυα, καθώς αντιμετωπίζει το πρόβλημα των εξαφανιζόμενων κλίσεων (vanishing gradient, [10] [11]).

* Tanh:

Χρησιμοποιείται στο Πυκνό στρώμα του σταδίου αποκωδικοποίησης της εισόδου (decoder). Δίνει τιμές στο διάστημα -1 ως 1 και καθώς το κέντρο συμμετρίας της είναι στο 0, μπορεί να κατηγοριοποιεί τις τιμές εισόδου σε πιο έντονα αρνητικές και έντονα θετικές.

* Softmax:

Χρησιμοποιείται σε προβλήματα κατηγοριοποίησης πολλαπλών κλάσεων, καθώς ομαλοποιεί τις τιμές των δεδομένων και τις δίνει στο διάστημα 0 ως Έτσι δίνει την πιθανοτική κατανομή κάθε εισόδου για μια συγκεκριμένη κλάση.

* Sigmoid:

Η sigmoid χρησιμοποιείται για επίλυση μη γραμμικών προβλημάτων. Το πλεονέκτημά της είναι πως είναι παραγωγίσιμη και έτσι δίνει πιο ομαλές κλίσεις και αλλαγές στις τιμές εξόδου.

* Swish:

Η swish είναι συνεχής ομαλή συνάρτηση και επιτρέπει και την έξοδο αρνητικών τιμών. Δίνει συχνά πιο ακριβή αποτελέσματα από τη χρήση της ReLu και εφαρμόζεται στο μοντέλο του Conformer.

Diagram

Description automatically generatedΈνα στρώμα νευρώνων είναι πολλοί νευρώνες μαζί και πάνω από ένα στρώματα που συνδέονται μεταξύ τους οδηγούν σε αρχιτεκτονικές πολλαπλών στρωμάτων (Multi-layer ANN). Τα νευρωνικά δίκτυα πολλαπλών στρωμάτων προσφέρουν τη δυνατότητα επίλυσης πιο περίπλοκων πράξεων και προβλημάτων, καθώς συνδυάζονται οι εξισώσεις ενεργοποίησής τους. Ένα απλό δίκτυο πολλαπλών στρωμάτων Feed Forward φαίνεται στην Εικόνα 5 , όπου το δίκτυο συνδέει όλες τις εξόδους του με όλες τις εισόδους του. Όταν συμβαίνει αυτό, δηλαδή όταν όλες οι είσοδοι ενός στρώματος συνδέονται με τις εξόδους ενός διπλανού στρώματος, τότε έχουμε Πυκνά Νευρωνικά Δίκτυα (Dense). Τα Πυκνά στρώματα προτιμούνται σε πολλές εφαρμογές, καθώς λόγω της πλήρης σύνδεσής όλων των εισόδων και εξόδων τους το καθιστούν σχεδόν αδύνατο να χαθεί σημαντική πληροφορία από τα δεδομένα, έτσι μαθαίνει και παράγει νέα χαρακτηριστικά χρησιμοποιώντας τα χαρακτηριστικά όλων των προηγούμενων στρωμάτων. Δομές όπως αυτή των Πυκνών Στρωμάτων συνήθως επιλέγονται αρκετές φορές ακόμα και από CNN, καθώς απαιτούνε τη χρήση λιγότερων παραμέτρων στο δίκτυο και έτσι μειώνεται και ο χρόνος που απαιτείται για να εκπαιδευτούν. Ακόμα και ένα τόσο απλό μοντέλο όπως στην Εικόνα 5 μπορεί να κάνει περίπλοκους υπολογισμούς και να δώσει ικανοποιητικά αποτελέσματα σε προβλήματα κατηγοριοποίησης.

Εικόνα 5: Απεικόνιση ενός πλήρως συνδεδεμένου Feed Forward δικτύου με ένα κρυφό στρώμα.(Εικόνα από [37])

### Συνελικτικά Δίκτυα (CNN)

Diagram

Description automatically generatedΈνα από τα ευρέως χρησιμοποιούμενα δίκτυα στους τομείς κατηγοριοποίησης είναι τα Συνελικτικά Νευρωνικά Δίκτυα (CNN). Χρησιμοποιούνται ευρέως σε εφαρμογές Όρασης Υπολογιστών(Computer Vision), επεξεργασίας εικόνων (Image Processing) και αναγνώρισης ομιλίας (ASR), λόγω του ότι είναι ικανά να αντιμετωπίζουν την ολική πληροφορία των δεδομένων εισόδου σε μορφή εικόνας και να αναγνωρίζουν τα μοτίβα σε αυτή. Τα Συνελικτικά δίκτυα βασίζονται στην τεχνική των κοινών βαρών, όπου χρησιμοποιούν ίδιες τιμές βαρών πάνω στα τμήματα της εισόδου για να μειωθούν οι παράμετροι που είναι αναγκαίοι για την εκπαίδευση του δικτύου. Τα βάρη αυτά αποτελούν και τα φίλτρα του κάθε CNN στρώματος και με βάση αυτά εξάγονται και τα μοτίβα της εισόδου μέσω συνελίξεων με αυτά. Πιο αναλυτικά, ένα CNN στρώμα έχει ένα τρισδιάστατο πίνακα που ονομάζεται φίλτρο όπου κάθε θέση του περιέχει κάποιες τιμές, τα βάρη. Συνήθως οι διαστάσεις των φίλτρων είναι τετραγωνικές, πολύ μικρότερες από τον πίνακα εισόδου που εφαρμόζεται η συνέλιξη. Αφού εφαρμοστεί πλήρως πάνω στον πίνακα εισόδου, υπολογίζει το εσωτερικό γινόμενο του επικαλυπτόμενου τμήματος της εικόνας με το φίλτρο. Η διαδικασία της συνέλιξης συνεχίζει μέχρι το φίλτρο να «περάσει» πάνω από όλο τον πίνακα εισόδου. Έστω για παράδειγμα ότι έχουμε για είσοδο μια εικόνα με διαστάσεις 16x16x1 και χρειάζεται να τη συνελίξουμε. Έστω ότι έχουμε ένα Συνελικτικό στρώμα με διαστάσεις φίλτρου 4x4x1. Τότε, υπολογίζεται το εσωτερικό γινόμενο του φίτρου με τα 4x4x1 πρώτα πάνω αριστερά πίξελ της εικόνας, μετά θα ολισθήσει κατά ένα πίξελ δεξιά στα επόνεα 4x4x1 πίξελ κ.ο.κ Οι διαστάσεις της συνελιγμενης εικόνας είναι (16-4 + 1)x(16-4 +1)x1 = 13x13x1. Ο τρόπος που λειτουργούν φαίνεται και στην Εικόνα 6.

Εικόνα 6: Απεικόνιση συνελίξεως εικόνας εισόδου 7x7x1 με CNN φίλτρου μεγέθους 3x3x1( Εικόνα από [33]).

Το μέγεθος του πίνακα εξόδου και επιλογή των πίξελ εισόδου που γίνεται η συνέλιξη μπορεί να επηρεαστεί και από δύο παραμέτρους: το stride και το padding. Το stride καθορίζει κατά πόσο θα «γλιστρήσει» το φίλτρο σε κάθε βήμα συνέλιξης, ενώ το padding γεμίζει νοητά τον πίνακα εισόδου με μηδενικές τιμές γύρω από τις πλευρές του για να αυξ΄σει το τελικό μέγεθος της εξόδου. Στο πάνω παράδειγμα το stride θεωρήθηκε ίσο με τη μονάδα και δε χρησιμοποιήθηκε padding.

Diagram, engineering drawing

Description automatically generated

Εικόνα 7: Συνολική αρχιτεκτονική ενός Συνελικτικού δικτύου (Εικόνα από [12]).

Ένα πυκνό νευρωνικό συνελικτικό δίκτυο περιέχει και άλλα στρώματα πέρα του συνελικτικού, όπως Dense (τα οποία αναφέρθηκαν παραπάνω) και Pooling (Εικόνα 7). Τα Pooling στρώματα χρησιμοποιούνται για τη μείωση των διαστάσεων εξόδου ενός CNN, έτσι ώστε να μειωθεί η περιπλοκότητα των πράξεων. Λειτουργούν χωρίζοντας νοητά σε τμήματα τους νευρώνες ενός στρώματος και συνδυάζοντας τους σε ένα μόνο νευρώνα εξόδου. Συχνοί τύποι Pooling στρωμάτων είναι τα Max Pooling και τα Average Pooling στρώματα. Στα πρώτα, επιλέγεται για έξοδος από τους νευρώνες εισόδου, ο νευρώνας αυτός που έχει την μέγιστη τιμή. Στη δεύτερη κατηγορία Pooling, η έξοδος καθορίζεται από το μέσο όρο τιμών όλων των νευρώνων εισόδου. Τα Pooling στρώματα είναι επίσης χρήσιμα για την καταπολέμιση του προβλήματος overfitting.

### Diagram Description automatically generated Recurent Networks

Εικόνα 8: Μηχανισμός ενός απλού Αναδρομικού Δικτύου (RNN)(Εικόνα από [35]).

Τα Αναδρομικά Δίκτυα (RNN) χρησιμοποιούνται ευρέως σε προβλήματα που πραγματεύονται ακολουθιακά δεδομένα. Ο βασικός μηχανισμός τους είναι η αναδρομική ανατροφοδότηση της εξόδου του δικτύου στην είσοδο (forward propagation) έχοντας ως αποτέλεσμα σε μια καλύτερη τιμή πρόβλεψης στην έξοδο του στρώματος (Εικόνα 7). Γνωστοί τύποι τέτοιων δικτύων είναι τα LSTM και τα GRU. Σε προβλήματα ταξινόμησης και αναγνώρισης ήχων συνηθίζεται να αναπτύσσονται μοντέλα που συνδυάζουν Συνελικτικά μοντέλα και αναδρομικά, τα οποία ονομάζονται CRNN και μέχρι τα τελευταία χρόνια αποτελούσαν την πιο ακριβέστερη μέθοδο Βαθιάς Μαθήσεως για την αντιμετώπιση του SELD προβλήματος. Ειδικά πριν την εμφάνιση των Transformers αποτελούσαν την πιο ευρέως αποδεκτή επιλογή για προβλήματα αυτόματης αναγνώρισης ομιλίας (ASR)( [13], [14], [15]) λόγω της ικανότητας τους να αναγνωρίζουν με επιτυχία της χρονικές εξαρτήσεις σε οποιαδήποτε ακολουθία δεδομένων εισόδου.

### Residual Networks

Diagram

Description automatically generatedΗ χρήση CNN ή CRNN δικτύων χρησιμοποιούνται για τη πλειοψηφία μοντέλων αναγνώρισης ήχου, καθώς έχουν τη δυνατότητα να αναγνωρίζουν τις τοπικές σχέσεις των δεδομένων εισόδων μεταξύ τους, και έτσι να γίνεται καλύτερη κατηγοριοποίηση. Καθώς οι αρχιτεκτονικές εξελίσσονται, τόσο περισσότερα στρώματα χρησιμοποιούνται σε παραπάνω αρχιτεκτονικές για να αυξηθεί η ακρίβεια της κατηγοριοποίησης, ειδικά σε επικαλυπτόμενα ηχητικά συμβάντα όπου η είσοδος χρειάζεται παραπάνω επεξεργασία. Έτσι οδηγούμαστε αρκετές φορές στο βασικότερο πρόβλημα των Βαθιών δικτύων, το οποίο είναι αυτό του εξαφανισμένου κλάσματος (vanishing gradient). Μια λύση για αυτό το πρόβλημα βρέθηκε με την εισαγωγή skip συνδέσεων στο δίκτυο, δηλαδή τον συγκαταλογισμό της αρχικής εισόδου του στρώματος κατά τον υπολογισμό της εξόδου. Χρήση αυτής της προσέγγισης κάνουν τα υπολειμματικά δίκτυα (Residual Networks ή ResNets). Η πρωτοτυπία τους είναι στην εισαγωγή των υπολειμματικά μπλοκ (residual blocks) που εμπνέονται από την ιδέα των skip συνδέσεων. Πιο συγκεκριμένα, αποτελούνται από ομάδες CNN μπλοκ, του ταυτοτικού μπλοκ και του συνελικτικού μπλοκ. Κάθε ομάδα αποτελείται αρχικά από ένα συνελικτικό μπλοκ, όπου η skip είσοδος συνελίσσεται, και από κάποιον αριθμό ταυτοτικών μπλοκ, όπου δεν υπάρχει συνέλιξη του skip. Ο αριθμός ομάδων και ταυτοτικών μπλοκ εξαρτάται από το βάθος του δικτύου, το οποίο μαρτυράται από τον αριθμό στο όνομα του ResNet. Για παράδειγμα, ένα ResNet-34 αποτελείται από 34 στρώματα και πιο συγκεκριμένα από ένα συνελικτικό στρώμα 7x7, 4 ομάδες υπολειμματικών μπλοκ (6, 8, 12 και 6 αντίστοιχα σε κάθε ομάδα) και τέλος Diagram

Description automatically generatedένα average pooling στρώμα.

Εικόνα 9: Απεικόνιση ενός ταυτοτικού μπλοκ σε ένα υπολειμματικό δίκτυο. Με μαύρο βελάκι αναπαρίσταται και η skip σύνδεση (Εικόνα από [18]).

Εικόνα 10: Ένα υπολειμματικό μπλοκ (Εικόνα από [36]).

Ας υποθέσουμε ότι έστω f(x) η έξοδος ενός πολυστρωματικού CRNN και με x συμβολίζουμε την είσοδο του δικτύου αυτού. Το δίκτυο προσπαθεί να προσεγγίσει μια βέλτιστη τιμή εξόδου g(x), η οποία είναι αρκετά περίπλοκη να λυθεί. Τώρα όσον αφορά το residual block δεν είναι τίποτα παρά ένας προσθέτης που απλοποιεί την εξίσωση που πρέπει να υπολογισθεί, εδώ την g(x), σε μια απλούστερη f(x) = g(x) − x ( [16], [17], [18]). Οπότε πλέον αντί να χρειάζεται να λυθεί μια περίπλοκη εξίσωση, αρκεί να λυθεί μόνο η αθροιστική εξίσωση f(x) + x = g(x), δηλαδή προστίθεται το skip x μαζί με την επεξεργασμένη συνελιγμένη έξοδο f(x) και δίνει ως αποτέλεσμα το μέσο όρο των τιμών τους (Εικόνα 9). Αυτή η μέθοδος έχει παρατηρηθεί ότι ξεπερνά σε επιδόσεις αρχιτεκτονικής πλήρους CRNN και η βασική ιδέα των skip συνδέσεων έχει εμπνεύσει αρκετές επιτυχημένες αρχιτεκτονικές όπως το DenseNet ( [19]).

### Transformers

Η ιδέα των Transformers ανακοινώθηκε πρώτη φορά στο άρθρο «All You Need Is Attention» [20] όπου περιγράφει ένα μοντέλο για μετάφραση οποιαδήποτε ακολουθίας εισόδου σε μια επιθυμητή ακολουθία εξόδου μέσω μηχανισμών προσοχής (attention) και υπολειμματικών συνδέσεων (residuals). Ο μηχανισμός προσοχής είναι η διαδικασία αντιστοίχισης διανυσμάτων ζήτησης (query) και ενός ζεύγους κλειδιού-τιμής (key-value) προς μια επιθυμητή έξοδο, η οποία υπολογίζεται από το εσωτερικό γινόμενο κάποιων βαρών με τις τιμές εισόδου. Τα βάρη υπολογίζονται μέσω κάποιας συνάρτησης συσχέτισης του query με το αντίστοιχο κλειδί του και με αυτά. Οι τιμές τις εξόδου μαρτυρούν τη συσχέτιση των queries και keys μεταξύ τους και καθορίζουν ανάλογα πόσο να επικεντρωθεί το δίκτυο σε αυτές τις σχέσεις, ή αλλιώς πόση προσοχή να δώσει, για αυτό και ο μηχανισμός ονομάζεται έτσι.

Εικόνα 11: Η αρχιτεκτονική ενός Transformer. Το αριστερό σκέλος αποτελεί τον κωδικοποιητή, ενώ το δεξί τον αποκωδικοποιητή (Εικόνα από [20]).

Diagram

Description automatically generated

Η εισαγωγή των Transformers ( [20] [21] [22] [23]) και των attention μηχανισμών έχει φέρει επανάσταση στον τομέα της επεξεργασίας και αναγνώρισης ήχου και σε πληθώρα εφαρμογών, όπως σε μοντέλα μηχανικής μετάφρασης γλώσσας, chatbots κ.λ.π. Τα τελευταία χρόνια έχουν εξελιχθεί τόσο ώστε να έχουν αντικαταστήσει τα Recurrent δίκτυα σε πολλά σύγχρονα μοντέλα ASR και language models λόγω του ότι είναι ικανά να κατανοούν και να βρίσκουν τις σχέσεις εξαρτήσεων των δεδομένων εισόδου μεγαλύτερης εμβέλειας και μήκους. Η επιτυχία του αυτή βασίζεται στο μηχανισμό αυτό-προσοχής (self-attention) όπου έχει απεριόριστο σε μήκος παράθυρο αναφοράς για να βρει τις σχέσεις στα δεδομένα εισόδου, σε αντίθεση με τα RNN, με αποτέλεσμα να μπορεί να βρει το συνολικό πλαίσιο συνάφειας των δεδομένων εισόδου και εξόδου και να συγκλίνει πιο γρήγορα σε κάποια πρόβλεψη. Μερικά από τα σύγχρονα μοντέλα που κάνουν χρήση Transformers είναι το BERT, GPT, 1D-DETR [24], SepFormer [25] και πολλά ακόμα. Ένα ακόμα σηματνικό πλεονέκτημα των Transformers είναι η μικρότερή τους επιβάρυνσης σε σχέση με τα CRNN κατά τη διάρκεια της εκπαίδευσης, καθώς αποτελούνται από λιγότερες παραμέτρους λόγω της έλλειψής πράξεων συνέλιξης.

Η δομή τους περιγράφεται στο άρθρο [20] και πιο αναλυτικά φαίνεται στην Εικόνα 10. Γενικά δομές νευρωνικών δικτύων για προβλήματα μετατροπής ακολουθιακών δεδομένων χρησιμοποιούν τη τεχνική κωδικοποίησης-αποκωδικοποίησης (encoder-decoder), στην οποία τα δεδομένα εισόδου περνάνε σε ένα στρώμα κωδικοποιητή, στο οποίο γίνεται αντιστοίχιση των τιμών τους σε μια συνεχή ακολουθιακή αναπαράσταση που να μπορεί να επεξεργασθεί από το δίκτυο. Αφού γίνει η επεξεργασία, τα δεδομένα περνάνε από τον αποκωδικοποιητή όπου και παράγεται μια νέα ακολουθία με κάθε σύμβολο να δημιουργείται με βάση τα προηγουμένως παραγόμενα σύμβολα στην ακολουθία (autoregression). Ο Transformer στο στάδιο του encoder χρησιμοποιεί μηχανισμούς ενσωμάτωσης (Embedding), κωδικοποίησης θέσης (positional-encoding) , Πολυκεφαλικής Αυτό-Προσοχής (Multi-Head Self Attention ή MHSA) και Feed Forward δικτύων.

Τα στρώματα Ενσωμάτωσης απλά μετατρέπουν τις ακολουθίες εισόδου σε τιμές που μπορούν να αναγνωριστούν από το δίκτυο. Η τεχνική του positional-encoding χρησιμοποιείται, εφόσον δεν υπάρχει μοντέλο συνέλιξης ή αναδρομής στο δίκτυο για να παραχθεί κάποια πληροφορία για τη σχέση σειράς των δεδομένων εισόδου μεταξύ τους. Η σχέση αυτή υπολογίζεται με βάση τις εξισώσεις:

,

Diagram

Description automatically generated, όπου οι διαστάσεις της ακολουθίας εισόδου, η θέση και η διάσταση κάθε ενσωματωμένου διανύσματος. Τα διανύσματα που παράγονται προστίθενται με τα αρχικά ενσωματωμένα διανύσματα και παράγουν την είσοδο για το στρώμα MHSA.

Εικόνα 12: Αναπαράσταση στρώματος MHSA (Εικόνα από [20]).

Η Πολύ-Κεφαλική Αυτό-Προσοχή απλά εκτελεί τον μηχανισμό προσοχής πολλαπλές φορές παράλληλα και είναι Αυτό-Προσοχής γιατί τα διανύσματα query και key ταυτίζονται. Επίσης χρησιμοποιεί Προσοχή Κλιμακωμένου Εσωτερικού Γινομένου (Scaled Dot-Product Attention) για την ανάκτηση της επιβαρυμένης εισόδου, η εξίσωση της οποίας είναι:

, όπου τα *query, key, value* σε μορφή πινάκων, οι διαστάσεις των keys και η εξίσωση ενεργοποίησης. Οι έξοδοι κάθε Κεφαλής συμπτύσσονται μεταξύ τους σε μία όπου μέσω skip σύνδεσης προστίθεται σε αυτή τα δεδομένα πριν την επεξεργασία από το στρώμα MHSA. Το άθροισμα αυτό έπειτα περνάει ως είσοδος σε ένα Γραμμικό στρώμα (Linear) όπου αποτελεί και το τελευταίο στρώμα στον κωδικοποιητή και παράγει την τελική έξοδο κωδικοποίησης. Η έξοδος αυτή πλέον με τη πληροφορία της Αυτό-Προσοχής βοηθάει τον αποκωδικοποιητή να δώσει έμφαση σε συγκεκριμένα σημεία των δεδομένων. Ο κωδικοποιητής χρησιμοποιεί όμοια στρώματα και μηχανισμούς με τον κωδικοποιητή ενώ έχει ένα πρόσθετο στρώμα MHSA.

### Conformer

Diagram

Description automatically generatedΣε αντίθεση με τους Transformers, τα CNN και RNN είναι καλύτερα στην εύρεση εξαρτήσεων ακολουθίας εισόδου σε τοπικό επίπεδο. Μάλιστα έρευνες και πειραματισμοί με διάφορα μοντέλα, όπως στην πρόταση Wu [26] ο οποίος προτείνει τον διαχωρισμό των δεδομένων εισόδου σε δύο ξεχωριστούς κλάδους που κάνουν χρήση Transformers και Συνελίξεως ανεξάρτητα, έχουν αποδείξει ότι για καλύτερα αποτελέσματα σε προβλήματα ASR είναι χρήσιμες οι τοπικές αλλά και ολικές εξαρτήσεις των δεδομένων. Για αυτό το λόγο δημιουργήθηκε μια αρχιτεκτονική που να συνδυάζει τα οφέλη των CNN και των Transformers, το Conformer. Το μοντέλο του Conformer πρωτοπαρουσιάστηκε στην αναφορά με τίτλο *«Conformer: Convolution-augmented Transformer for Speech Recognition»* [27]και είναι κυρίως εμπνευσμένο από την πρόταση του Wu στο πολυκλαδικό του μοντέλο. Αλλά αντίθετα με το μοντέλο του Wu όπου η τελική έξοδος των δεδομένων συμπτύσσεται αφού έχει διαχωριστεί και εφαρμοστεί ξεχωριστά σε αυτά Αυτό-Προσοχή και Συνέλιξη, στο Conformer συνδυάζονται οι μηχανισμοί αυτοί μαζί με Feed Forward δίκτυα για την εξαγωγή πληροφορίας ταυτόχρονα, σε ένα κλάδο.

Εικόνα 13: Απεικόνιση ενός μπλοκ κωδικοποιητή Conformer (αριστερά) και ενός Conformer μπλοκ (δεξιά) (Εικόνα από [27]).

Diagram

Description automatically generatedDiagram

Description automatically generatedΗ αρχιτεκτονική ενός Conformer χωρίζεται στον κωδικοποιητή και αποκωδικοποι­ητή. Το μπλοκ του κωδικοποιητή αποτελείται από ένα αρχικό στρώμα Συνέλιξης και από πολλαπλά στοιβαγμένα Conformer blocks (Εικόνα 13). Ένα Conformer block αποτελείται από τέσσερις (4) διαφορετικές μονάδες αρχιτεκτονικών, οι οποίες είναι παραλλαγές των Feed Forward δικτύου, MHSA, Συνελικτικού Δικτύου και στρώματος κανονικοποίησης (Layer Normalization), το οποίο κανονικοποιεί τις τιμές ενεργοποί­ησης των προηγούμενων στρωμάτων. Σε κάθε μονάδα αρχιτεκτονικής εφαρ­μόζεται και η τεχνική της υπολειπομένης σύνδεσης, η οποία προστίθεται στην έξοδο κάθε μονάδας. Η μονάδα Feed Forward (Εικόνα 14) ακολουθεί τη δομή παρόμοιων Feed Forward αρχιτεκτονικών σε μοντέλα Transformers, μόνο που στο Conformer block χρησιμοποιεί επιπλέον Swish συνάρτηση ενεργοποίησης για την κανονικοποίησης του δικτύου.

Εικόνα 14: (Πάνω) Feed Forward μονάδα. (Κάτω) Συνελικτική μονάδα (Εικόνες από [27]).

Η μονάδα MHSA είναι όμοια με αυτή που εφαρμόζεται σε έναν Transformer μόνο που στο κομμάτι της Αυτό-Προσοχής χρησιμοποιεί την εξίσωση της σχεσιακής κωδικοποίησης σχέσης (relative positional encoding) [28], η οποία έχει καλύτερα αποτελέσματα στη γενίκευση του δικτύου ανεξάρτητα από το μήκος εισόδου.

Η μονάδα Συνέλιξης, εμπνευσμένη από την έρευνα του Wu [26], περιέχει Pointwise συνέλιξη και GLU μονάδας. Η Pointwise συνέλιξη χρησιμοποιείται σε προβλήματα αναγνώρισης τρισδιάστατων αντικειμένων με δεδομένα σημεία εισόδου (point cloud) και λειτουργεί εφαρμόζοντας τη πράξη της συνέλιξης με φίλτρο 1x1 προς κάθε σημείο της εισόδου. Ακολουθεί μονοδιάστατη συνέλιξη κατά βάθος (1D Depthwise Convolution) με Swish συνάρτηση ενεργοποίησης. Η Συνέλιξη κατά βάθος σπάει την είσοδο σε κανάλια και εφαρμόζει την πράξη της συνέλιξης σε κάθε ένα ξεχωριστά. Αυτού του είδος η συνέλιξη προτιμάται σε βαθιά νευρωνικά δίκτυα, καθώς αντιμετωπίζει το πρόβλημα της υπερμοντελοποίησης (overfitting) [29]. Υποθέτοντας ότι με συμβολίζουμε την είσοδο και με την έξοδο του Conformer, η εξίσωση της εξόδου μπορεί να εκφραστεί:

Όπου FFN(.), MHSA(.), Conv(.) και Layernorm(.) οι μονάδες Feed Forward, Multi-Head Self-Attention, Convolution και Layer Normalization.

# Επεξήγηση του διαγωνισμού DCASE 2021 task 3

## Ο διαγωνισμός DCASE 2021

Το θέμα της αναγνώρισης και ταξινόμησης ηχητικών συμβάντων έχει προσελκύσει τόσο πολύ το ενδιαφέρον διάφορων ερευνητών με αποτέλεσμα τη δημιουργία του παγκόσμιου διαγωνισμού DCASE (Detection and Classification of Acoustic Sound Events [30]. Ο διαγωνισμός καλεί ερευνητές φοιτητές ή και επαγγελματίες από όλο το κόσμο που ασχολούνται με πεδία επεξεργασίας σημάτων και Μηχανικής Μάθησης με σκοπό την ανταλλαγή ιδεών και γνώσεων και την έρευνα πάνω στην αντιμετώπιση προβλημάτων ταξινόμησης και αναγνώρισης ήχου σε διάφορα περιβάλλοντα. Ο διαγωνισμός του DCASE αποτελεί ευκαιρία για την εξέλιξη του κλάδου και συνεχώς αυξάνεται η δημοτικότητά του, προσελκύοντας κάθε χρόνο όλο και περισσότερους ερευνητές. Ο πρώτος διαγωνισμός άρχισε το 2013 και από το 2016 και μετά εγκαθιδρύθηκε να γίνεται ετησίως, προσφέροντας και σχετικά εργαστήρια (workshops) πέρα από προκλήσεις. Αυτή τη χρονιά, ο διαγωνισμός DCASE 2021 αποτελούνταν από 6 κατηγορίες προκλήσεων (tasks), πιο αναλυτικά αυτές οι προκλήσεις αφορούσαν:

* Task 1: Ηχητική αναγνώριση περιβάλλοντος.
* Task 2: Αναγνώριση ηχητικών ανωμαλιών που προέρχονται από μηχανικές πηγές με χρήση μοντέλου ανίχνευσης ανωμαλιών.
* Task 3: Αναγνώριση ηχητικών συμβάντων και εύρεση θέσης τους στο χώρο με παρεμβατικές πηγές.
* Task 4: Αναγνώριση και διαχωρισμός ηχητικών γεγονότων σε εσωτερικούς χώρους.
* Task 5: Βιοακουστική αναγνώριση ήχων.
* Task 6: Αυτόματη παραγωγή επεξηγηματικού κειμένου περιγραφής ακουστικών γεγονότων.

Το θέμα της συγκεκριμένης διπλωματικής εργασίας αφορά το task 3 (SELD) [31].

## Το πρόβλημα SELD

Ένα Sound event Detection and Localization (SELD) σύστημα για το task 3 θα πρέπει να αναγνωρίζει ποιες και πόσες κλάσεις είναι ενεργές σε κάθε υπολογισμένο χρονικό πλαίσιο της εισόδου (SED) καθώς και την θέση της κάθε κλάσης που εντοπίζεται στο χώρο σε σφαιρικές συντεταγμένες, δηλαδή μέσω της Αζιμούθια γωνίας τους και του ύψους τους από το μικρόφωνο εγγραφής (DOA). Για να καθορίσει τον εντοπισμό ενός συμβάντος, το σύστημα πρέπει να βγάζει ως έξοδο μια γραμμή που υποδουλώνει την τοπική του δραστηριότητα στο συγκεκριμένο χρονικό πλαίσιο μαζί με δύο ακόμα χωρικές τροχιές που προσδιορίζουν τη γωνία Αζιμούθια και την τιμή ανύψωσης του. Οι κλάσεις στο γράφημα προσδιορίζονται από το διαφορετικό τους χρώμα. Το τελικό γράφημα πρέπει να περιέχει όλα τα ενεργά συμβάντα σε κάθε χρονικό πλαίσιο μαζί με τη θέση τους για ένα δοσμένο αρχείο. Στο φετινό διαγωνισμό υπάρχουν 12 κλάσεις αναγνώρισης:

* Σειρήνα (0)
* Μωρό που κλαίει (1)
* Ήχος πρόσκρουσης (2)
* Σκυλί που γαβγίζει (3)
* Γυναικεία κραυγή (4)
* Γυναικεία ομιλία (5)
* Ήχος βημάτων (6)
* Χτύπημα πόρτας (7)
* Ανδρική κραυγή (8)
* Ανδρική ομιλία (9)
* Τηλέφωνο (10)
* Πιάνο (11)

Η διαφορά του με το διαγωνισμό DCASE 2020 είναι πως φέτος προστίθενται, εκτός από τη περιβαλλοντικό θόρυβο και αντήχησης στα καταγεγραμμένα αρχεία ήχου, και ηχητικές παρεμβολές, οι οποίες δεν ανήκουν σε ηχητικό συμβάν των δοσμένων κλάσεων. Για κάθε καταγεγραμμένη ηχητική εγγραφή, τα παρεμβαλλόμενα ηχητικά συμβάντα είναι χωροθετημένα με τον ίδιο τρόπο που θα ήταν και ένα οποιοδήποτε συμβάν στόχο. Συγκεκριμένα, ο αριθμός ηχητικών δειγμάτων που χρησιμοποιήθηκαν για τον διαμοιρασμό τους σε δοσμένες κλάσεις ήταν 500, ενώ τα δείγματα που χρησιμοποιήθηκαν για παρεμβολές ήταν 400. Έτσι ο φετινός διαγωνισμός γίνεται πιο προκλητικός για τους ερευνητές, καθώς αποτελεί και μια πιο κοντά σε πραγματικές συνθήκες προσομοίωση.

Επίσης ο μέγιστος αριθμός των επικαλυπτόμενων ηχητικών συμβάντων αυξήθηκε σε τρία (3), με τη δυνατότητα επικάλυψης συμβάντων ίδιας κλάσης.

Ωστόσο η πιο σημαντική αλλαγή στο φετινό διαγωνισμό είναι η σύμπτυξη των κλάδων για το έργο του SED και DOA, χρησιμοποιώντας μια νέα μορφή στόχου-εξόδου των προβλέψεων, την ACCDOA [32]. Το δίκτυο πλέον βγάζει ως έξοδο ένα διάνυσμα ACCDOA για κάθε ηχητικό συμβάν σε κάθε χρονικό πλαίσιο της εγγραφής εισόδου. Στη μορφή ACCDOA, ομαδοποιείται η απόκλιση της ταξινόμησης και

Εικόνα 15: ACCDOA αναπαράσταση για 2 κλάσεις (γάβγισμα και βήματα) για ένα χρονικό πλαίσιο Τ, σκιαγραφημένο με γκρι πλαίσιο (αριστερά). Δεξιά φαίνεται πως για κάθε κλάση αναπαρίσταται με ένα διάνυσμα στο χώρο. Η κλάση «βήματα» δεν είναι ενεργή στο Τ οπότε το μέτρο του διανύσματος είναι μηδέν. Αντιθέτων, το «γάβγισμα» είναι ενεργό, και η κατεύθυνσή του φαίνεται με την κατεύθυνση του διανύσματος με βάση τις Καρτεσιανές DOA συντεταγμένες (Εικόνα από [3]).

Diagram

Description automatically generated

τοποθέτησης κάθε πρόβλεψης σε ένα διάνυσμα. Η κατάσταση μίας κλάσης, αν είναι δηλαδή ενεργή ή όχι σε ένα χρονικό πλαίσιο, καθορίζεται από το μήκος του διανύσματος αυτού. Για κάθε ενεργό συμβάν επίσης καθορίζεται το DOA του με Καρτεσιανές συντεταγμένες με βάση την κατεύθυνση του διανύσματος ACCDOA. Ένα παράδειγμα για 2 κλάσεις με ACCDOA διάνυσμα παρουσιάζεται στην Εικόνα 15.

## Δεδομένα εισόδου

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση και τον έλεγχο του μοντέλου έγιναν με τη βάση δεδομένων **TAU-NIGENS Spatial Sound Events 2021** [32], η οποία περιέχει πολλαπλές χωρικές ηχογραφήσεις σκηνών και συμβάντων διαφορετικών κατευθύνσεων και αποστάσεων από το σημείο εγγραφής. Πιθανές τιμές της Αζιμούθιας γωνίας είναι στο διάστημα και ανύψωσης στο διάστημα . Η χωροθέτηση κάθε ηχητικού συμβάντος σε κάθε ηχητική σκηνή έγινε με χρήση των χωρητικών αποκρίσεων τους (Impulse response ή IR). Κάθε IR ηχογραφήθηκε με τη βοήθεια ενός σφαιρικού μικροφώνου Eigenmike σε 13 διαφορετικές τοποθεσίες εσωτερικού χώρου στο Πανεπιστήμιο Ταμπερέ της Φιλανδίας (Tampere University). Κάθε ηχητικό κλιπ αντιστοιχεί σε μια από τις 13 τοποθεσίες και αποτελεί σύνθεση πολλαπλών IR (SRIR). Η διαδικασία συλλογής και σύνθεσης των SRIR περιγράφεται στο [2]. Συνολικά για κάθε δωμάτιο μπορεί να ανατεθούν από 1184 μέχρι 6480 πιθανές SRIR θέσεις στο χώρο. Παράλληλα με τα IR, συλλέχθηκαν και ηχογραφήσεις περιβαλλοντικού θορύβου από τις ίδιες τοποθεσίες, συνολικής διάρκειας 30 λεπτών της ώρας και με το μέγεθος του θορύβου να έχει ομοιόμορφη αναλογία σήματος προς ήχο (SNR) από 30-60db, αθόρυβο περιβάλλον προς περιβάλλον με αρκετό θόρυβο.

Κάθε ηχογράφηση έχει στατικά και μη ηχητικά συμβάντα με την ίδια πιθανότητα να ανήκει κάποιο σε μία από αυτές τις κλάσεις. Για τα μη στατικά συμβάντα, συντίθενται με γωνιακή ταχύτητα 10deg/sec, 20deg/sec ή 40deg/sec ανάλογα με το αν η πηγή κινείται αργά, με μέτρια ταχύτητα ή γρήγορα.

Η βάση δεδομένων αποτελείται από 800 ηχογραφημένα κλιπ του ενός λεπτού το καθένα, από τα οποία 600 αποτελούν την βάση ανάπτυξης (development dataset) και τα υπόλοιπα 200 τη βάση αξιολόγησης (evaluation dataset). Οι ηχογραφήσεις του development dataset είναι οι καταγραφές που έγιναν στα 11 από τα 13 συνολικά δωμάτια και διαμοιράζονται σε 6 σετ δεδομένων (splits) των 100 ηχογραφήσεων το κάθε ένα. Οι ηχογραφήσεις των υπόλοιπων 2 δωματίων (200 συνολικά) δίνονται στο evaluation dataset. Πιο συγκεκριμένα για το development dataset, το split 1-4 είναι για εκπαίδευση (training), το 5 για επαλήθευση (validation) και το 6 για έλεγχο (testing). Κάθε ηχογράφηση σκηνής δίνεται σε δύο μορφές καταγραφής χώρου 4 καναλιών η κάθε μία, σε τετράεδρο πίνακα μικροφώνου (microphone array ή **MIC)** και Ambisonic πρώτης τάξης (first-order Ambisonics ή **FOA)**. Στα πλαίσια της διπλωματικής, επιλέχθηκε να ασχοληθούμε μόνο με τη μορφή MIC. Για τη μορφή FOA προστίθονται 3 ακόμα κανάλια για τα διανύσματα ακουστικής εντάσεως (acoustic intensity vectors), ενώ για τη MIC προστίθενται ακόμα 6 για τις γενικοποιημένες ακολουθίες διασυσχέτισης (GCC-PHAT) Σε κάθε κλιπ έγινε δειγματοληψία A picture containing chart

Description automatically generatedστα 24kHz [2].

Εικόνα 16: Γραφική απεικόνιση μιας συντιθέμενης ηχογράφησης. Τα χρωματισμένα αντικείμενα υποδεικνύουν τις κλάσεις-στόχους, τα γκρι τον θόρυβο και τα παρεμβαλόμενα συμβάντα και με βέλη τα μη στατικά συμβάντα (Εικόνα από [31]).

## Το baseline μοντέλο

Table

Description automatically generated with low confidence

Εικόνα 17: Δομή του CRNN baseline μοντέλου βασισμένη στο SELDnet (Εικόνα από [2]).

Η δομή και αρχιτεκτονική του baseline μοντέλου για το φετινό διαγωνισμό (Εικόνα 17) είναι παρόμοια με της προηγούμενης χρονιάς. Υιοθετεί ένα CRNN νευρωνικό δίκτυο που βασίζεται στην αρχιτεκτονική SELDnet [34]. Το μοντέλο αυτό συνδυάζει το SED έργο με την τοποθέτηση στο χώρο συνδυαστικά για την ταυτόχρονη αξιολόγησή τους και πρόβλεψης κάθε ηχητικού γεγονότος.

Κατά την εξαγωγή των χαρακτηριστικών, το baseline μοντέλο για κάθε αρχείο ηχογράφησης εισόδου εξάγει πολυκαναλικά log-mel spectrograms σε 64 mel-δέσμες (mel-bands) από FFT 1024 σημείων, χρησιμοποιώντας παράθυρο Hamming 40 ms με 20 ms μήκος άλματος (50% επικάλυψη) στα 24kHz συχνότητα δειγματοληψίας. Σε πειράματα που δοκιμάστηκαν αργότερα, αποδείχθηκε ότι η αλλαγή της συχνότητας σε 16kHz και των mel-bands σε 128 δε βελτίωσε τα αποτελέσματα, οπότε υιοθετήσαμε τις ίδιες παραμέτρους εξαγωγής των ακουστικών χαρακτηριστικών με το baseline μοντέλο στα δικά μας πειράματα. Αυτή η διαδικασία δίνει μια ακολουθία από T αριθμό STFT πλαισίων. Αφού γίνει η εξαγωγή των χαρακτηριστικών, η οποία συμπτύσσει το log-mel spectrogram με το διάνυσμα intensity vector ή GCC-PHAT ανάλογα με την επιλεγμένη μορφή δεδομένων, γίνεται η συνέλιξή τους από ένα δίκτυο CNN τριών στρωμάτων για τον εντοπισμό των τοπικών αμετάβλητων μετατοπίσεων (shift-invariant features) με συνάρτηση ενεργοποίησης ReLu και υποδειγματοληψίας (downsampling) μέσω Max-Pooling. Ο εντοπισμός των shift-invariant χαρακτηριστικών βοηθάει το δίκτυο στην καλύτερη προσαρμογή και αναγνώριση του σε ίδια δεδομένα εισόδου με μικρές μετατοπίσεις. Μετά το στάδιο αυτό, προκύπτει ακολουθία με T/5 πλαίσια. Στη συνέχεια ακολουθεί RNN δίκτυο bi-GRU, το οποίο χρησιμεύει στην εκμάθηση των χρονικών συσχετίσεων της εισόδου. Τέλος, τα δεδομένα περνάνε από δύο Fully Connected στρώματα και δίνει τελική έξοδο ένα ACCDOA διάνυσμα διαστάσεων T/5 x 3\*12, όπου κάθε μία από τις 12 κλάσεις-στόχους αντιπροσωπεύεται από 3 κόμβους που αντιστοιχούν στις συντεταγμένες τους x, y, z στο χώρο. Η τελική έξοδος αναγνωρίζει ένα γεγονός ως ενεργό και επιλέγεται η υπολογισμένη του DOA, εφόσον το μήκος του διανύσματος δραστηριότητας για τη συγκεκριμένη κλάση ξεπερνάει το όριο τιμής 0.5.

Ωστόσο, το δοσμένο μοντέλο δεν είναι σε θέση να ταξινομεί επικαλυπτόμενα ηχητικά συμβάντα της ίδιας κλάσης, αν και αυτή η περίπτωση μπορεί να προκύψει καθώς περιλαμβάνεται στο σετ των δεδομένων. Μια από τις αρχιτεκτονικές που παρουσιάζεται παρακάτω στη διπλωματική εργασία προσπαθεί να καταπολεμήσει και αυτό το πρόβλημα.

## Καταμέτρηση ακρίβειας- Metrics

Για την αξιολόγηση της ακρίβειας κάθε μοντέλου, χρησιμοποιούνται 4 μετρικές, από τις οποίες 2 συγκεντρώνονται στο κομμάτι ακρίβειας του SED και οι υπόλοιπες 2 στην ακρίβεια χωρικής τοποθέτησης. Συγκεκριμένα, οι μετρικές που στοχεύουν στην αξιολόγηση του SED, αλλιώς αναφερόμενες και ως μετρικές με επίγνωση στη τοποθεσία (location-aware), πλέον εξαρτώνται από την τοποθεσία των συμβάντων, Με τις προβλέψεις που λαμβάνονται υπόψιν να είναι κάτω από ένα όριο απόστασης . Οι μετρικές αυτές είναι το σφάλμα απόκλισης (error rate) και στο F-σκορ (F-score).

Οι υπόλοιπες 2 μετρικές, αλλιώς αναφερόμενες και ως ταξικά συνειδητοποιημένες, εξαρτώνται από τις εντοπισμένες κλάσεις, δηλαδή κάθε μετρική υπολογίζεται για κάθε κλάση συνολικά και όχι για όλη την έξοδο συνολικά. Αυτές οι μετρικές είναι το σφάλμα τοποθέτησης , το οποίο εκφράζει τη μέση γωνιακή απόσταση μεταξύ των προβλέψεων και αναφορών στην ίδια κλάση, και η δεύτερη μετρική είναι μια ανάκληση τοποθέτησης , η οποία εκφράζει το πραγματικό ποσό των υπολογισμένων τοποθεσιών που αναγνωρίσθηκαν σε μία κλάση σε σχέση με όλες τις εντοπισμένες κλάσεις. Περισσότερα για τον τρόπο υπολογισμό τους, υπάρχει στον σύνδεσμο [35].

Επίσης σημειώνεται πως κάθε μετρική υπολογίζεται σε μη-επικαλυπτόμενα χρονικά πλαίσια διαρκείας ενός δευτερολέπτου.

Γενικά ένα επιτυχημένο σύστημα προσπαθεί να μειώσει όσο περισσότερο γίνεται τα και και να αυξήσει τα και .

# Προσέγγιση του DCASE 2021 task 3

## Πρώτη προσέγγιση – ResNet

# Πειράματα και αποτελέσματα

## Τίτλος ενότητας

# - Συμπεράσματα

## Σύνοψη και συμπεράσματα

## Μελλοντικές επεκτάσεις

# Βιβλιογραφία

|  |  |
| --- | --- |
| [1] | «Types of classification algorithms,» [Ηλεκτρονικό]. Available: https://analyticsindiamag.com/7-types-classification-algorithms/. |
| [2] | A. a. A. S. a. K. D. a. D. A. a. S. P. a. V. T. Politis, «A Dataset of Dynamic Reverberant Sound Scenes with Directional Interferers for Sound Event Localization and Detection,» November 2021. [Ηλεκτρονικό]. Available: https://dcase.community/workshop2021/proceedings. |
| [3] | K. a. K. Y. a. T. N. a. T. S. a. M. Y. Shimada, «ACCDOA: Activity-Coupled Cartesian Direction of Arrival Representation for Sound Event Localization and Detection,» αρ. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2021. |
| [4] | «But what is the Fourier Transform? A visual introduction.,» [Ηλεκτρονικό]. Available: https://www.youtube.com/watch?v=spUNpyF58BY. |
| [5] | «Fast Fouier Transform,» [Ηλεκτρονικό]. Available: https://www.nti-audio.com/en/support/know-how/fast-fourier-transform-fft. |
| [6] | «Fourier Transform,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Fourier\_transform. |
| [7] | «Understanding the Mel Spectrogram,» 6 May 2020. [Ηλεκτρονικό]. Available: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53. |
| [8] | «Mel Spectograms explained,» [Ηλεκτρονικό]. Available: https://www.youtube.com/watch?v=9GHCiiDLHQ4. |
| [9] | «Mel scale example,» [Ηλεκτρονικό]. Available: https://www.sfu.ca/sonic-studio-webdav/handbook/Mel.html. |
| [10] | «Addressing the vanishing gradients problem for beginners,» [Ηλεκτρονικό]. Available: https://analyticsindiamag.com/addressing-the-vanishing-gradient-problem-a-guide-for-beginners/. |
| [11] | «How to fix vanishing gradients problem using the rectified lnear activation function,» [Ηλεκτρονικό]. Available: https://machinelearningmastery.com/how-to-fix-vanishing-gradients-using-the-rectified-linear-activation-function/. |
| [12] | «A Comprehensive Guide to Convolutional Neural Networks,» [Ηλεκτρονικό]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. |
| [13] | T. N. S. Y. W. R. P. P. N. C.-C. Chiu, ΙΕΕΕ, 2018. |
| [14] | H. S. a. R. P. K. Rao, «Exploring architectures,data and units for streaming end-to-end speech recognition with rnn-transducer,» ΙΕΕΕ, 2017. |
| [15] | A. Graves, 2012. |
| [16] | X. Z. S. R. J. S. Kaiming He, «Deep Residual Learning for Image Recognition». |
| [17] | S. Bhattacharyya, «Understand and Implement ResNet-50 with TensorFlow 2.0,» [Ηλεκτρονικό]. Available: https://towardsdatascience.com/understand-and-implement-resnet-50-with-tensorflow-2-0-1190b9b52691. |
| [18] | «How to code your ResNet from scratch in Tensorflow?,» [Ηλεκτρονικό]. Available: https://www.analyticsvidhya.com/blog/2021/08/how-to-code-your-resnet-from-scratch-in-tensorflow/. |
| [19] | Z. L. K. Q. W. L. v. d. M. Gao Huang, «Densely Connected Convolutional Networks,» 2016. |
| [20] | «All You Need Is Attention,» Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, 2017. |
| [21] | «Transformers explanation and implementation». |
| [22] | «Illustrated Guide to Transformers Neural Network: A step by step explanation». |
| [23] | «Transformer Neural Networks - EXPLAINED! (Attention is all you need)». |
| [24] | «Sound Event Detection Transformer:,» Zhirong Ye, Xiangdong Wang, Hong Liu, Yueliang Qian, Rui Tao, Long Yan, Kazushige Ouchi. |
| [25] | «Attention is All You Need in Speech Separation,» Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, Jianyuan Zhong, 2020. |
| [26] | Z. L. J. L. Y. L. S. H. Zhanghao Wu, «https://arxiv.org/abs/2004.11886,» 2020. |
| [27] | «Conformer: Convolution-augmented Transformer for Speech Recognition,» Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang, 2020. |
| [28] | «Transformer-xl: Attentive language models beyond a fixed-length context,» 2019. |
| [29] | «Υπερμοντελοποίηση στην Μηχανική Μάθηση (overfitting),» [Ηλεκτρονικό]. Available: https://contia.gr/%CE%BC%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE-%CE%BC%CE%AC%CE%B8%CE%B7%CF%83%CE%B7/. |
| [30] | «DCASE community,» [Ηλεκτρονικό]. Available: https://dcase.community/. |
| [31] | «DCASE 2021 TASK 3,» [Ηλεκτρονικό]. Available: https://dcase.community/challenge2021/task-sound-event-localization-and-detection. |
| [32] | «TAU dataset,» [Ηλεκτρονικό]. Available: https://zenodo.org/record/4844825#.YnAYW-1BxD8. |
| [33] | «An example of convolutional operation in 2D,» [Ηλεκτρονικό]. Available: https://www.researchgate.net/figure/An-example-of-convolution-operation-in-2D-2\_fig3\_324165524. |
| [34] | «Building a ResNet in Keras,» [Ηλεκτρονικό]. Available: https://towardsdatascience.com/building-a-resnet-in-keras-e8f1322a49ba. |
| [35] | «Neural Networks,» [Ηλεκτρονικό]. Available: https://medium.com/nerd-for-tech/neural-networks-68531432fb5. |
| [36] | «An Overview of ResNets and its variants,» [Ηλεκτρονικό]. Available: https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035. |
| [37] | «What is a neural network,» [Ηλεκτρονικό]. Available: https://www.psychz.net/client/kb/en/what-is-a-neural-network.html. |

# ΠΑΡΑΡΤΗΜΑΤΑ

(Εάν υπάρχει μόνο ένα παράρτημα αυτή η σελίδα παραλείπεται και εμφανίζεται μόνο η επόμενη σελίδα με το παράρτημα χωρίς αλφαβητική αρίθμηση. Δηλ. απλά ως ΠΑΡΑΡΤΗΜΑ)

# Παράρτημα Α

# Τίτλος παραρτήματος

Τα παραρτήματα περιλαμβάνουν συνοδευτικό, υποστηρικτικό υλικό (πίνακες, φωτογραφίες, ερωτηματολόγια, στατιστικά στοιχεία, αποδείξεις, περιγραφές λογισμικών προγραμμάτων, παραδείγματα, περιγραφές πολύπλοκων διαδικασιών, λίστα με πρωτογενή στοιχεία, λεπτομερή περιγραφή και προδιαγραφές εξοπλισμού, οδηγίες εγκατάστασης λογισμικού, κ.λπ.), ή αλλιώς ό,τι θεωρείται χρήσιμο να περιγραφεί, αλλά δεν συνηθίζεται να εντάσσεται μέσα στο κυρίως κείμενο της Εργασίας. Στο κυρίως κείμενο της Εργασίας πρέπει να γίνονται οι κατάλληλες παραπομπές προς τα παραρτήματα, όπου το κείμενο σχετίζεται με υλικό που περιλαμβάνεται σε αυτά. Ένα παράρτημα, αναλόγως με το περιεχόμενό του, μπορεί να είναι ενιαίο, ή να χωρίζεται σε ενότητες.

## Βασικές αρχές

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed consectetur pellentesque quam nec ultricies. Pellentesque sollicitudin ultrices vehicula. Vivamus luctus nunc nec nibh laoreet fermentum. Vestibulum sit amet faucibus nulla, eu pulvinar magna. Donec vel metus ex. Praesent tempus nisl id rhoncus vulputate. Curabitur arcu lorem, volutpat nec placerat eu, elementum a felis. Cras non dapibus arcu. Duis vel quam lacinia, tempor nisl vel, volutpat purus. Duis id risus vitae erat venenatis lobortis sed a tellus. Mauris ex neque, euismod et cursus nec, porttitor id metus. Aliquam malesuada tortor id ipsum placerat accumsan.

## Επιπλέον παρατηρήσεις

Sed eget lacus augue. Maecenas placerat sodales lectus, et aliquet turpis. Sed eget enim mollis, auctor risus ac, dignissim orci. Curabitur vitae viverra nibh, ut tristique nibh. Phasellus rhoncus dignissim eleifend. Aenean rhoncus consequat ipsum, at tristique tellus. Integer tortor est, faucibus eget consequat in, facilisis et nisi. Phasellus pellentesque fringilla turpis sed ultrices. Nullam porttitor, nisl sit amet consequat vulputate, purus dui laoreet justo, id ultrices arcu lacus a nisi. Fusce maximus orci ut velit eleifend, sed vestibulum nibh viverra. Phasellus id turpis aliquet, fermentum neque nec, rutrum massa. Integer a quam nulla. Sed porta, orci ullamcorper rhoncus pellentesque, ipsum nulla pellentesque tortor, eget mattis magna nisl sed dui. Curabitur et gravida sem. Nulla nisl turpis, vehicula in sapien id, tincidunt sodales libero. Donec nec dignissim eros, rutrum facilisis nibh.

## Παραδοχές

Aenean sodales vehicula nisl eu finibus. Quisque id neque a odio tempor dignissim ut sed odio. Maecenas at nunc vestibulum, laoreet tortor quis, tincidunt ex. In in velit ut nibh luctus egestas. Cras aliquet sem felis, nec aliquet purus convallis sit amet. Ut non elit eget metus porttitor elementum ut vel nunc. Pellentesque leo neque, sollicitudin at libero ac, scelerisque finibus nibh. Aenean mollis scelerisque dui eget tincidunt. Vestibulum aliquam et ante nec consequat. Quisque mauris urna, ultricies eget erat eu, porttitor venenatis est. Sed mollis diam quam, nec maximus purus varius at. In condimentum tellus sed hendrerit iaculis. Duis in dui sit amet nulla aliquam dignissim. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Suspendisse quis egestas massa, quis vulputate ipsum. Maecenas in pellentesque justo, vitae vehicula enim. Sed eget dolor nec nisi euismod ultrices ut nec velit. Aliquam consectetur non est vitae eleifend. Nam vulputate odio sed erat tempor ultricies. Curabitur sodales fermentum dui, a tincidunt odio egestas vitae. Aenean ac ultrices nibh. Nulla sit amet finibus purus. Phasellus at sodales erat. Suspendisse nulla mauris, feugiat posuere convallis in, auctor vitae lacus. Donec dignissim arcu quis fermentum feugiat. Praesent varius semper est a vehicula. Phasellus dignissim ante ut quam fermentum ornare. Fusce nec neque nec tellus vestibulum blandit at in tellus. Ut fermentum tincidunt orci nec consectetur.

## Θεωρητικό υπόβαθρο

Quisque ut condimentum leo. Etiam pharetra nisl purus, eget porta ex consectetur aliquet. Vestibulum consequat viverra faucibus. Integer fermentum, eros vel tincidunt aliquet, enim sem bibendum urna, a ullamcorper mauris eros vitae neque. Suspendisse sed neque ac augue efficitur congue. Integer commodo varius pretium. Quisque non nisl vitae libero vulputate commodo aliquet vel justo. Duis iaculis, purus vulputate ultricies pharetra, lacus nisl cursus tortor, eget faucibus eros nunc sed massa. Integer luctus malesuada iaculis. Mauris et imperdiet quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Quisque ut condimentum leo. Etiam pharetra nisl purus, eget porta ex consectetur aliquet. Vestibulum consequat viverra faucibus. Integer fermentum, eros vel tincidunt aliquet, enim sem bibendum urna, a ullamcorper mauris eros vitae neque. Suspendisse sed neque ac augue efficitur congue. Integer commodo varius pretium. Quisque non nisl vitae libero vulputate commodo aliquet vel justo. Duis iaculis, purus vulputate ultricies pharetra, lacus nisl cursus tortor, eget faucibus eros nunc sed massa. Integer luctus malesuada iaculis. Mauris et imperdiet quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos.

# Παράρτημα Β

# Τίτλος παραρτήματος

Τα παραρτήματα περιλαμβάνουν συνοδευτικό, υποστηρικτικό υλικό (πίνακες, φωτογραφίες, ερωτηματολόγια, στατιστικά στοιχεία, αποδείξεις, περιγραφές λογισμικών προγραμμάτων, παραδείγματα, περιγραφές πολύπλοκων διαδικασιών, λίστα με πρωτογενή στοιχεία, λεπτομερή περιγραφή και προδιαγραφές εξοπλισμού, οδηγίες εγκατάστασης λογισμικού, κ.λπ.), ή αλλιώς ό,τι θεωρείται χρήσιμο να περιγραφεί, αλλά δεν συνηθίζεται να εντάσσεται μέσα στο κυρίως κείμενο της Εργασίας. Στο κυρίως κείμενο της Εργασίας πρέπει να γίνονται οι κατάλληλες παραπομπές προς τα παραρτήματα, όπου το κείμενο σχετίζεται με υλικό που περιλαμβάνεται σε αυτά. Ένα παράρτημα, αναλόγως με το περιεχόμενό του, μπορεί να είναι ενιαίο, ή να χωρίζεται σε ενότητες.