ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΓΝΩΡΙΣΗ ΚΑΙ ΕΝΤΟΠΙΣΜΟΣ ΗΧΗΤΙΚΩΝ ΓΕΓΟΝΟΤΩΝ ΣΕ ΕΣΩΤΕΡΙΚΟ ΧΩΡΟ

Διπλωματική Εργασία

Στυλιανός Ροβολής

Επιβλέπων: Γεράσιμος Ποταμιάνος

ΒΟΛΟΣ 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΓΝΩΡΙΣΗ ΚΑΙ ΕΝΤΟΠΙΣΜΟΣ ΗΧΗΤΙΚΩΝ ΓΕΓΟΝΟΤΩΝ ΣΕ ΕΣΩΤΕΡΙΚΟ ΧΩΡΟ

Διπλωματική Εργασία

Στυλιανός Ροβολής

Επιβλέπων: Γεράσιμος Ποταμιάνος

ΒΟΛΟΣ 2023

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

RECOGNITION AND LOCALIZATION OF ACOUSTIC EVENTS INDOORS

Diploma Thesis

Stylianos Rovolis

Supervisor: Gerasimos Potamianos

VOLOS 2023

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων **Γεράσιμος Ποταμιάνος**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Δασκαλοπούλου Ασπασία**
Αναπληρώτρια Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Αντωνόπουλος Χρήστος**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Ποταμιάνο Γεράσιμο για την υπομονή και την πολύτιμη βοήθειά του και τους γονείς μου για την τεράστια στήριξή τους όλα αυτά τα χρόνια.

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, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει 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 Δηλών

Στυλιανός Ροβολής

Διπλωματική Εργασία

ΑΝΑΓΝΩΡΙΣΗ ΚΑΙ ΕΝΤΟΠΙΣΜΟΣ ΗΧΗΤΙΚΩΝ ΓΕΓΟΝΟΤΩΝ ΣΕ ΕΣΩΤΕΡΙΚΟ ΧΩΡΟ

Στυλιανός Ροβολής

Περίληψη

Η αναγνώριση και ο εντοπισμός ηχητικών γεγονότων είναι η διαδικασία αναγνώρισης της θέσης και της κατηγορίας των ηχητικών συμβάντων σε μια εγγραφή ήχου πολλαπλών καναλιών. Έχει πολλές πρακτικές εφαρμογές, όπως η επιτήρηση εσωτερικών χώρων, ο αυτοματισμός κτηρίων και η ρομποτική. Μερικές από τις προκλήσεις και τα προβλήματα που αντιμετωπίζουν οι ερευνητές είναι η μεγάλη ποικιλία και η μεταβλητότητα των ηχητικών συμβάντων, η ταυτόχρονη συνύπαρξη διαφορετικών ηχητικών πηγών, η παρουσία θορύβου και παρεμβολών και η έλλειψη κατάλληλα επεξεργασμένων δεδομένων.

Σε αυτή τη διπλωματική εργασία επικεντρώνουμε την προσέγγισή μας σε αυτό το πρόβλημα σύμφωνα με το συνέδριο DCASE2019 [1]. Θα αξιολογήσουμε διάφορα μοντέλα βαθιάς μάθησης χρησιμοποιώντας τις μετρικές και το σύνολο δεδομένων του παραπάνω συνεδρίου με σκοπό να βελτιώσουμε τη δοθείσα βασική μέθοδο. Τα αποτελέσματά μας δείχνουν ότι η εκπαίδευση ξεχωριστών μοντέλων για τα προβλήματα αναγνώρισης και εντοπισμού, επιδεικνύει καλύτερη απόδοση από ένα σύστημα που αναλαμβάνει ταυτόχρονα και τα δύο αυτά ζητήματα. Χρησιμοποιώντας χαρακτηριστικά log Mel και GCC-PHAT, καταφέραμε να επιτύχουμε μια μικρή βελτίωση σε σχέση με τη βασική μέθοδο, αν και δεν φτάσαμε στο επίπεδο πιο σύγχρονων συστημάτων που υπάρχουν σήμερα.

Λέξεις-κλειδιά:

αναγνώριση και εντοπισμός ήχου, ψηφιακή επεξεργασία σημάτων, βαθιά μάθηση

Diploma Thesis

RECOGNITION AND LOCALIZATION OF ACOUSTIC EVENTS INDOORS

Stylianos Rovolis

Abstract

The recognition and localization of acoustic events involves identifying the location and class of sound events in a multi-channel audio recording. This task has numerous practical applications, such as surveillance, building automation, and robotics. However, it is challenged by the wide variability of sound events, the presence of multiple active sources at once, background noise and interference, and a lack of annotated data for training and evaluation.

In this thesis, we focus on a specific approach to this problem based on the DCASE2019 workshop [1]. We evaluated various deep learning models using the provided metrics and dataset, with the goal of improving upon the baseline method. Our results show that separate training for the tasks of sound recognition and localization performs better than a single model for both tasks. By using log Mel and GCC-PHAT features, we were able to achieve a slight improvement over the baseline method, although we did not reach the level of state-of-the-art systems.

Keywords:

sound event localization, sound event detection, digital signal processing, deep learning

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

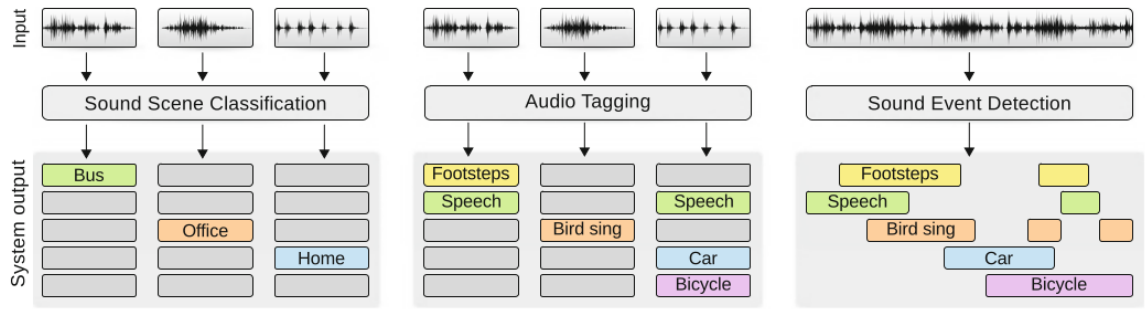
Συντομογραφίες

CNN	Convolutional Neural Network
DFT	Discrete Fourier Transform
DOA	Direction of Arrival
FC	Fully Connected
FFT	Fast Fourier Transform
GCC-PHAT	Generalized Cross Correlation Phase Transform
GRU	Gated Recurrent Unit
IR	Impulse Response
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SED	Sound Event Detection
SELD	Sound Event Localization and Detection
STFT	Short-Time Fourier Transform

Κεφάλαιο 1

Εισαγωγή

Η αναγνώριση και η κατηγοριοποίηση ηχητικών γεγονότων περιέχει ένα ευρύ φάσμα εφαρμογών. Σε ένα αστικό περιβάλλον θα μπορούσε να χρησιμοποιηθεί για την αναγνώριση διάφορων ήχων [13], σε μία αίθουσα συνεδριάσεων για τον εντοπισμό του ομιλητή [14] και για την μελέτη της βιοποικιλότητας [15]. Τα συστήματα που αντιμετωπίζουν το παραπάνω πρόβλημα χωρίζονται σε δυο κυρίως κατηγορίες. Σε εκείνα που αγνοούν τα χρονικά χαρακτηριστικά του προβλήματος και επικεντρώνονται μόνο στην σωστή ταξινόμηση των ήχων και στα μοντέλα που εξάγουν επιπλέον χρονικές πληροφορίες όπως είναι η στιγμή έναρξης και λήξης κάθε ακουστικού γεγονότος. Τέτοια χρονικά (temporal) μοντέλα στην βιβλιογραφία αναφέρονται και ως συστήματα ανίχνευσης (detection) [16]. Διαχωρίζονται με αυτόν τον τρόπο από τα μοντέλα ταξινόμησης (classification) [17] και ακουστικής επισήμανσης (audio tagging) [18] .



Σχήμα 1.1: Συστήματα ανάλυσης ηχητικών γεγονότων. Εικόνα από [2].

1.1 Αντικείμενο της Διπλωματικής

Στην παρακάτω εργασία το πρόβλημα της αναγνώρισης και εντοπισμού ηχητικών γεγονότων θα το αντιμετωπίσουμε χρησιμοποιώντας ένα σύστημα ανίχνευσης. Η περιγραφή του προβλήματος και τα δεδομένα που θα χρησιμοποιηθούν, αντλήθηκαν από το workshop DCASE 2019 challenge, task 3, Sound Event Localization and Detection [1]. Στόχος της διπλωματικής είναι η εφαρμογή και η αξιολόγηση αρχιτεκτονικών βαθιάς μάθησης με συνδυασμό πολλαπλών επιπέδων Συνελικτικών (Convolutional) και Επαναληπτικών (Recurrent) νευρωνικών δικτύων για την αντιμετώπιση του παραπάνω προβλήματος.

1.2 Συνεισφορά της Διπλωματικής

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Μελετήθηκαν διαφορετικές αρχιτεκτονικές δικτύων βαθιάς μάθησης και χαρακτηριστικών εισόδου για την αναγνώριση και τον εντοπισμό ηχητικών γεγονότων.

2. Υλοποιήθηκαν τρεις διαφορετικές αρχιτεκτονικές με συνδυασμό CNN και RNN δικτύων.

3. Αξιολογήθηκε η επίδοση των δομικών στοιχείων που τις αποτελούσαν, με διαφορετικούς συνδυασμούς επιπέδων CNN καθώς και των ίδιων αρχιτεκτονικών στο σύνολο τους.

4. Αξιοποιώντας τα παραπάνω συμπεράσματα χρησιμοποιήθηκε η πιο αποδοτική μέθοδος για την εξαγωγή των τελικών αποτελεσμάτων.

1.3 Οργάνωση του Τόμου

Το υπόλοιπο της διπλωματικής εργασίας οργανώνεται ως εξής:

- Στο Κεφάλαιο 2 γίνεται μια θεωρητική αναφορά στα τεχνητά νευρωνικά δίκτυα και την βαθιά μάθηση. Από τον πρώτο τεχνητό νευρώνα φτάνουμε στα περίπλοκα Συνελικτικά και Επαναληπτικά Νευρωνικά δίκτυα.

- Στο Κεφάλαιο 3 αναφέρονται βασικές τεχνικές επεξεργασίας των ηχητικών σημάτων.

- Στο Κεφάλαιο 4 αναλύεται η βάση δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση και την αξιολόγηση των μοντέλων μας.

- Στο Κεφάλαιο 5 εμβαθύνουμε στα προβλήματα του εντοπισμού και της ανίχνευσης της ηχητικής πηγής όπως αυτά περιγράφονται στο DCASE2019. Περιγράφονται οι μετρικές που συνδέονται με το πρόβλημα καθώς και οι μέθοδοι επεξεργασίας των δεδομένων εισόδου και εξόδου.

- Στο Κεφάλαιο 6 εξηγούμε την μέθοδο εκπαίδευσης για τα μοντέλα μας από την εξαγωγή των χαρακτηριστικών μέχρι τις διαφορετικές αρχιτεκτονικές που θα μελετηθούν στο πειραματικό μέρος.

- Το Κεφάλαιο 7 αποτελεί το πειραματικό μέρος της εργασίας και παρουσιάζονται τα τελικά αποτελέσματα χρησιμοποιώντας την πιο αποδοτική μέθοδο.

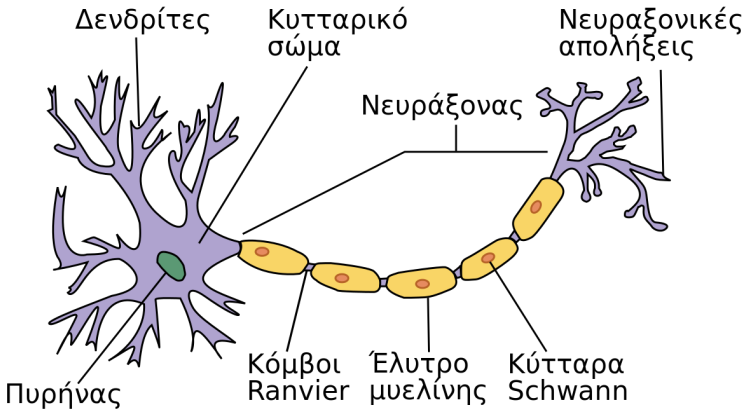
- Τέλος στο Κεφάλαιο 8 συνοψίζονται τα συμπεράσματα της διπλωματικής εργασίας και προτείνονται ιδέες για την βελτίωση και εξέλιξη των μεθόδων που χρησιμοποιήθηκαν.

Κεφάλαιο 2

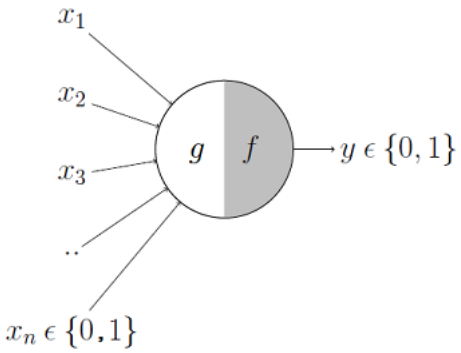
Τεχνητά Νευρωνικά Δίκτυα και Βαθιά Μάθηση

2.1 Νευρώνες και ο Αλγόριθμος Perceptron

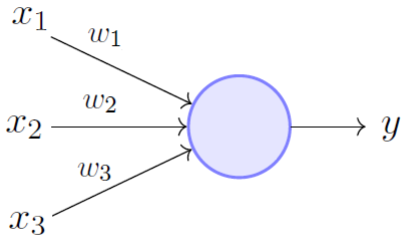
Ο ανθρώπινος εγκέφαλος αποτέλεσε πηγή έμπνευσης για την δημιουργία και την μελέτη των τεχνητών νευρωνικών δικτύων. Η έρευνα των βιολογικών νευρώνων και η προσπάθεια μοντελοποίησής τους με μαθηματικά εργαλεία οδήγησε τελικά στην δημιουργία του πρώτου τεχνητού νευρώνα από τους McCulloch και Pitts [19], τον οποίο σύντομα διαδέχθηκε το μοντέλο Perceptron.



Σχήμα 2.1: Ο βιολογικός νευρώνας. Εικόνα από [3].



Σχήμα 2.2: Ο τεχνητός νευρώνας των McCulloch-Pitts. Εικόνα από [4].



Σχήμα 2.3: Ο νευρώνας Perceptron. Εικόνα από [5].

Στην πιο απλή του λειτουργία ένας νευρώνας ενεργοποιείται όταν το άθροισμα της εισόδου ξεπερνάει κάποιο κατώφλι θ. Η συνάρτηση μεταφοράς του Perceptron είναι:

$$y = f(\sum_{i=1}^{n} w_i x_i - \theta) \tag{2.1}$$

Η συνάρτηση f ονομάζεται συνάρτηση ενεργοποίησης και οι παράμετροι w_i είναι τα συναπτικά βάρη του νευρώνα. Οι πιο διαδεδομένες μορφές της συνάρτησης $f()$ είναι οι παρακάτω:

- Βηματική -1/1 (step function):

$$f(u) = \begin{cases} -1 & \text{, αν } u \leq 0, \\ 1 & \text{, διαφορετικά.} \end{cases}$$

- Σιγμοειδής (sigmoid):

$$f(u) = \frac{1}{1 + e^{-u}}$$

- Υπερβολική εφαπτομένη (hyperbolic tangent):

$$f(u) = \tanh u = \frac{e^u - e^{-u}}{e^u + e^{-u}}$$

- Συνάρτηση κατωφλίου (threshold function):

$$f(u) = \begin{cases} 0 & , \text{αν } u \leq 0, \\ u & , \text{αν } 0 < u < 1, \\ 1 & , \text{αν } u \geq 1. \end{cases}$$

- Συνάρτηση διορθωμένης γραμμικής μονάδας (Rectified Linear Unit - ReLU):

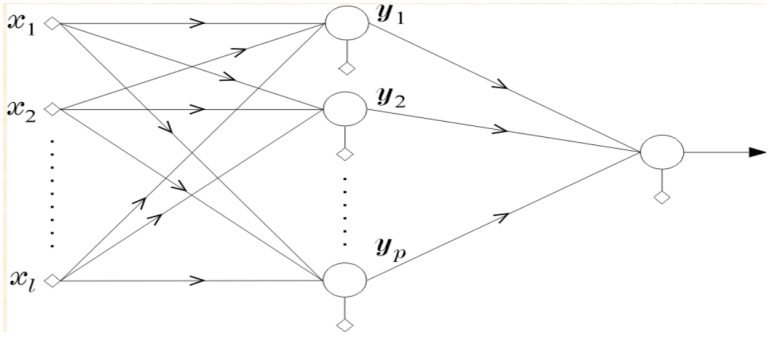
$$f(u) = \begin{cases} 0 & , \text{αν } u \leq 0, \\ u & , \text{αν } u > 0. \end{cases}$$

- Γραμμική (linear):

$$f(u) = u$$

2.2 Τα Πρώτα Δίκτυα Νευρώνων

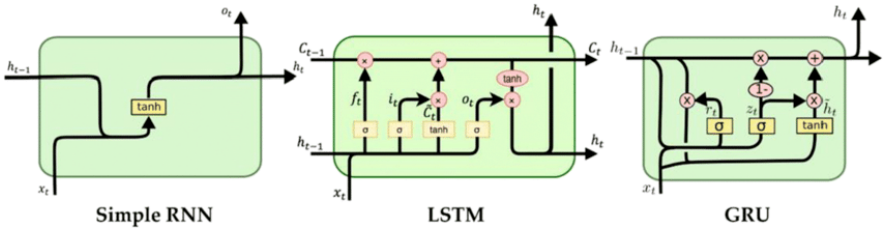
Ο απλός αλγόριθμος Perceptron με ένα νευρώνα είναι αποτελεσματικός σε απλά προβλήματα ταξινόμησης όπου τα δεδομένα είναι γραμμικά διαχωρίσιμα. Με την σύνδεση πολλαπλών νευρώνων σε επίπεδα και την επινόηση του αλγορίθμου οπισθο-διάδοσης (Back-Propagation) για την ανανέωση των βαρών και την εκπαίδευση του μοντέλου, καλύφθηκε επιπλέον ένα τεράστιο πλήθος εφαρμογών που περιέχει και μη γραμμικά προβλήματα. Αυτό είχε ως αποτέλεσμα τα νευρωνικά δίκτυα να γνωρίσουν μεγάλη άνθηση και να αρχίσουν έτσι να εφαρμόζονται σε πλήθος επιστημονικών πεδίων από την πληροφορική και την οικονομία, μέχρι την ιατρική και την βιολογία.



Σχήμα 2.4: Δίκτυο Perceptron δύο επιπέδων. Εικόνα από [6].

2.3 Επαναληπτικά Νευρωνικά Δίκτυα

Επόμενο βήμα στην εξέλιξη των νευρωνικών δικτύων αποτέλεσε η εισαγωγή της έννοιας του χρόνου. Η επινόηση των επαναληπτικών δικτύων έδωσε την δυνατότητα στους νευρώνες να επικοινωνούν με νευρώνες προηγούμενων στρωμάτων ή ακόμη και με τον εαυτό τους. Αυτή η διαδικασία ανατροφοδότησης προσφέρει δυναμική συμπεριφορά στο δίκτυο και ταυτόχρονα δίνει χαρακτηριστικά μνήμης στο σύστημα. Στην Σχήμα 2.5 παρουσιάζονται παραδείγματα νευρώνων τέτοιων δικτύων.

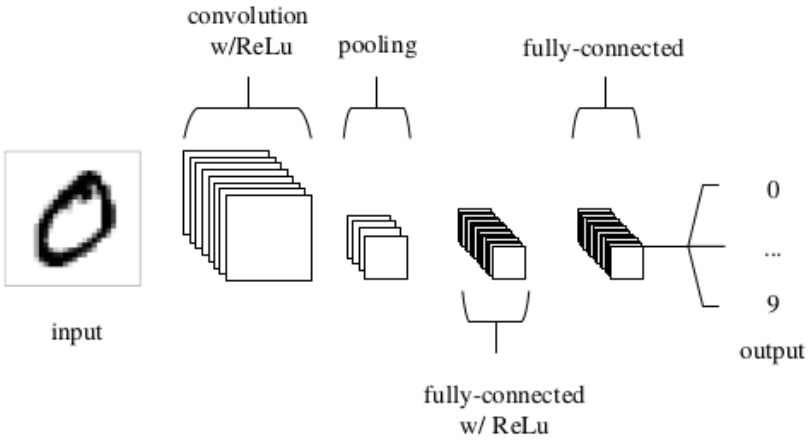


Σχήμα 2.5: Νευρώνες RNN. Εικόνα από [7].

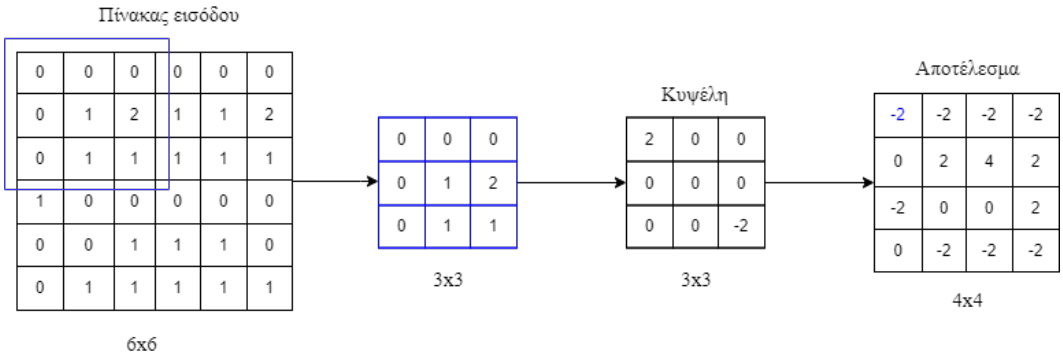
2.4 Συνελικτικά Νευρωνικά Δίκτυα

Όπως αναφέραμε και προηγουμένως τα δίκτυα Perceptron χρησιμοποιήθηκαν ευρέως στην ταξινόμηση προτύπων. Βασικό τους μειονέκτημα ωστόσο, ήταν ότι τα διάφορα χαρακτηριστικά έπρεπε να υπολογιστούν και να εισαχθούν χειροκίνητα ως είσοδοι. Η ανάγκη για αυτόματη εξαγωγή χαρακτηριστικών οδήγησε στην επινόηση των Συνελικτικών Νευρωνικών Δικτύων.

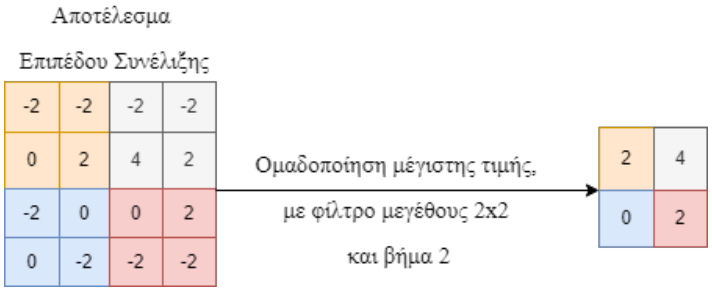
Ένα συνελικτικό νευρωνικό δίκτυο (CNN) αποτελείται από τρία βασικά επίπεδα (layers): Συνελικτικό επίπεδο (Convolution layer), επίπεδο Ομαδοποίησης (Pooling Layer), επίπεδο πλήρως συνδεδεμένων (Fully Connected) νευρώνων.



Σχήμα 2.6: Παράδειγμα αρχιτεκτονικής CNN. Εικόνα από [8].



Σχήμα 2.7: Λειτουργία του επιπέδου συνέλιξης. Σχεδιάστηκε στο [9].



Σχήμα 2.8: Μείωση διαστάσεων με ομαδοποίηση μέγιστης τιμής. Σχεδιάστηκε στο [9].

Τα Συνελικτικά Νευρωνικά Δίκτυα εφαρμόζονται κυρίως σε δεδομένα εικόνων και λύνουν προβλήματα που σχετίζονται με την ταξινόμηση τους ή την αναγνώριση αντικειμένων εντός τους. Με την ίδια λογική χρησιμοποιούνται σε βίντεο με ικανοποιητικά αποτελέσματα. Τα τελευταία χρόνια άρχισαν να ενσωματώνονται και σε εφαρμογές σχετικές με ήχο. Αντιμετωπίζοντας μια κυματομορφή ή ένα φασματογράφημα ενός μετασχηματισμού Fourier ως

εικόνα μπορούμε να αφήσουμε ένα δίκτυο CNN να εξάγει τα απαραίτητα χαρακτηριστικά χωρίς να χρειαστεί εξωτερική παρέμβαση.

Κεφάλαιο 3

Επεξεργασία Ηχητικών Σημάτων για Μηχανική Μάθηση

3.1 Εισαγωγή

Ο ήχος παράγεται από δόνηση, προκαλώντας στα κοντινά σωματίδια του αέρα μια παλινδρομική κίνηση, η οποία διαδίδεται ως ένα μηχανικό κύμα. Το ηχητικό κύμα φτάνει στο αυτί μας και η παραμόρφωση που δημιουργείται στο τύμπανο μεταφράζεται ως ήχος μέσω της ακοής. Μέσα σε αυτό το μηχανικό κύμα ταξιδεύει πλήθος πληροφοριών και έτσι μπορούμε να αναγνωρίζουμε τους διαφορετικούς ήχους και να εντοπίζουμε την πηγή τους.

3.2 Κυματομορφές

Η βασική και πιο συνήθης μορφή αναπαράστασης ενός ηχητικού κύματος είναι η κυματομορφή του. Μέσα από αυτήν διακρίνονται τα χρονικά χαρακτηριστικά του ήχου (αρχή, τέλος, διάρκεια), η έντασή του, η συχνότητα του ηχητικού κύματος κ.α. Ο ήχος σε αυτήν την μορφή αποτελεί ένα αναλογικό ή χρονοσυνεχές σήμα, το οποίο είναι αδύνατο να χρησιμοποιηθεί αυτούσιο από ένα ψηφιακό σύστημα όπως ο ηλεκτρονικός υπολογιστής. Η ψηφιακή επεξεργασία σήματος καλύπτει λοιπόν αυτό το πρόβλημα, αναλύοντας πλήθος διαφορετικών τεχνικών για την επίτευξη της παραπάνω διαδικασίας.

3.3 Ο Διακριτός Μετασχηματισμός Fourier

Ο διακριτός μετασχηματισμός Fourier (Discrete Fourier Transform - DFT) διαδραματίζει σημαντικό ρόλο στην ανάλυση, τον σχεδιασμό και την υλοποίηση αλγόριθμων και συστημάτων επεξεργασίας σημάτων διακριτού χρόνου, διότι οι βασικές ιδιότητές του καθιστούν ιδιαίτερα βολική την ανάλυση και τον σχεδιασμό συστημάτων στο πεδίο Fourier. Εξίσου σημαντικό είναι το γεγονός ότι υπάρχουν αποδοτικοί αλγόριθμοι για τον ακριβή υπολογισμό του DFT. Συνεπώς ο DFT αποτελεί μία σημαντική συνιστώσα σε πολλές πρακτικές εφαρμογές συστημάτων διακριτού χρόνου. Μια ιδιαίτερα αποδοτική κατηγορία αλγορίθμων για τον ψηφιακό υπολογισμό του ονομάζονται αλγόριθμοι ταχέως μετασχηματισμού Fourier (Fast Fourier Transform Algorithms - FFT). Σε σχέση με τον αριθμό των δειγμάτων N η απόδοση στον χρόνο υπολογισμού ενός FFT σε σχέση με τον DFT είναι τάξης μεγέθους καλύτερη.

- Discrete Fourier Transform: $O(N^2)$
- Fast Fourier Transform: $O(N \log N)$

Ένα μειονέκτημα όμως τέτοιων μετασχηματισμών στο πεδίο της συχνότητας είναι ότι εφαρμόζονται στο σύνολο των δειγμάτων του σήματος και έτσι χάνονται οι πληροφορίες στο πεδίο του χρόνου.

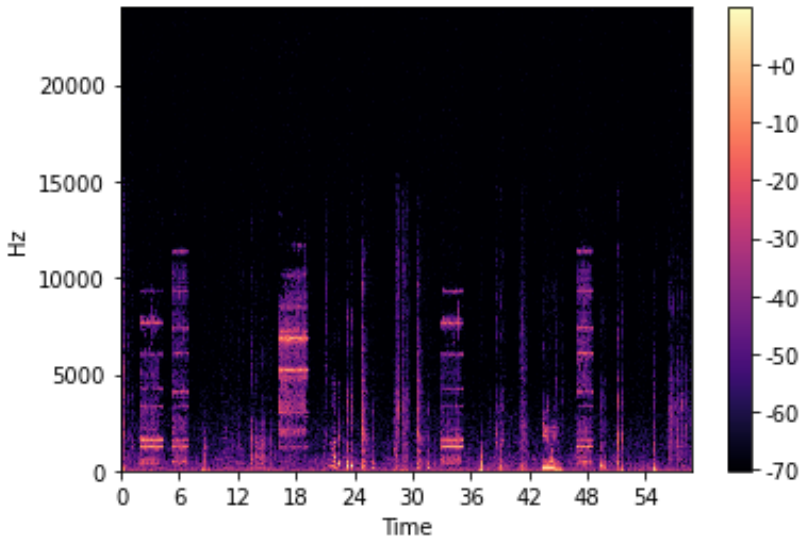
3.4 Ο Μετασχηματισμός Fourier Βραχέως Χρόνου

Ο μετασχηματισμός Fourier βραχέως χρόνου (Short-time Fourier Transform - STFT) είναι μια υποκατηγορία του διακριτού μετασχηματισμού Fourier που σκοπό έχει την διατήρηση μέρους των χρονικών χαρακτηριστικών του σήματος. Συνοπτικά είναι η εφαρμογή FFT σε μικρότερα κομμάτια του συνολικού σήματος (χρονικά πλαίσια). Αναλυτικά η διαδικασία για τον υπολογισμό του STFT είναι η εξής:

1. Χωρίζουμε τον ήχο σε χρονικά πλαίσια, ορισμένης διάρκειας.
2. Εφαρμόζουμε μία συνάρτηση παραθύρωσης (windowing function) σε κάθε πλαίσιο.
3. Υπολογίζουμε τον FFT στο αποτέλεσμα.

Η έξοδος του DFT είναι μία ακολουθία των συντελεστών Fourier για κάθε συχνότητα. Αντίθετα η έξοδος του STFT είναι ένας δισδιάστατος πίνακας από συντελεστές Fourier που

υπολογίζονται σε κάθε χρονικό πλαίσιο και για κάθε συχνότητα. Όπως αναφέραμε ο πιο απλός τύπος αναπαράστασης ενός ηχητικού σήματος είναι η κυματομορφή του. Με την μεταφορά στο πεδίο της συχνότητας και την εφαρμογή του STFT εξάγεται το φασματογράφημά του (spectrogram).



Σχήμα 3.1: Φασματογράφημα ενός αρχείου ήχου από DCASE 2019.

3.5 Φασματογράφημα σε Κλίμακα Mel

Το φασματογράφημα σε κλίμακα Mel (Mel Spectrogram) αποτελεί εξέλιξη του φασματογραφήματος που παράγεται από έναν STFT. Κύριος κορμός τους είναι η αντικατάσταση των γραμμικά κατανεμημένων συχνοτήτων όπου υπολογίζεται ο STFT, από ζώνες οι οποίες προσομοιώνουν τον λογαριθμικό τρόπο με τον οποίο αντιλαμβάνεται ο ανθρώπινος εγκέφαλος τις συχνότητες [20]. Η αντιστοίχηση των συχνοτήτων f στην κλίμακα Mel γίνεται με τον παρακάτω τύπο:

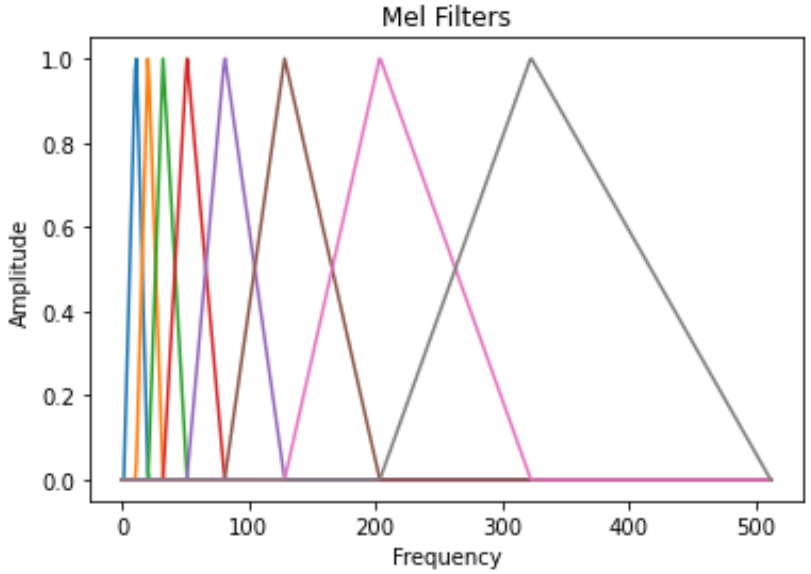
$$M(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.1}$$

και αντίστροφα:

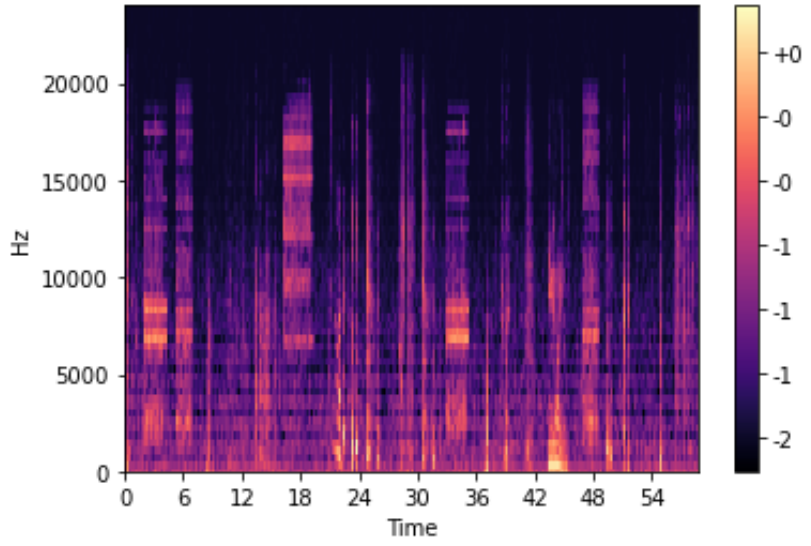
$$F(m) = 700 \cdot (10^{m/2595} - 1) \tag{3.2}$$

Επόμενο βήμα στην μετατροπή ενός STFT είναι η επιλογή του αριθμού των ζωνών Mel (Mel Bands) και η δημιουργία των τριγωνικών φίλτρων που θα εφαρμοστούν στον STFT. Το αποτέλεσμα είναι ένα λογαριθμικό φασματογράφημα Mel. Αποτελεί και αυτό πίνακα με

διαστάσεις: (Αριθμός ζωνών Mel) x (αριθμός χρονικών πλαισίων). Το αντίστοιχο φασματογράφημα Mel του φασματογραφήματος του Σχήματος 3.1 φαίνεται στο Σχήμα 3.3.



Σχήμα 3.2: Παράδειγμα 8 Τριγωνικών φίλτρων Mel.



Σχήμα 3.3: Φασματογράφημα σε κλίμακα Mel (με 64 φίλτρα) του ίδιου αρχείου ήχου με το Σχήμα 3.1.

Κεφάλαιο 4

Ανάλυση της Βάσης Δεδομένων

Η δική μας εργασία θα βασιστεί στο σύνολο δεδομένων “TAU Spatial Sound Events 2019 - Microphone Array” [21]. Για την συλλογή των κρουστικών αποκρίσεων (impulse responses - IRs) χρησιμοποιήθηκε μια διάταξη σφαιρικών μικροφώνων Eigenmike και ένα ηχείο το οποίο τοποθετείται σε διάφορες θέσεις γύρω από αυτή.

Σχήμα 4.1: Διάταξη μικροφώνων Eigenmike. Εικόνα από [10].

Οι κρουστικές αποκρίσεις συλλέχθηκαν στις παρακάτω κατευθύνσεις:

- 36 IRs για κάθε 10°αζιμούθιου, για 9 ανυψώσεις από -40°έως 40°σε απόσταση 1 μέτρου απο το μικρόφωνο. Σύνολο 324 IRs.
- 36 IRs για κάθε 10°αζιμούθιου, για 5 ανυψώσεις από -20°έως 20°, σε απόσταση 2 μέτρων από το μικρόφωνο. Σύνολο 180 IRs.

Η ηχογράφησή τους πραγματοποιήθηκε σε 5 διαφορετικούς εσωτερικούς χώρους στο Πανεπιστήμιο του Τάμπερε:

1. Language Center – Μεγάλη ανοιχτή αίθουσα με πολλά τραπέζια, καθίσματα και μοκέτα. Άνθρωποι που δουλεύουν και συνομιλούν.

2. Reaktori Building – Μεγάλος χώρος καφετέριας με καθίσματα, τραπέζια και μοκέτα. Άνθρωποι που συζητούν και τρώνε.

3. Festia Building – Ψηλοτάβανος διάδρομος με σκληρό πάτωμα. Άνθρωποι που περπατούν και συνομιλούν.

4. Tietotalo Building – Διάδρομος με αίθουσες και σκληρό πάτωμα. Άνθρωποι που περπατούν και συνομιλούν.

5. Sähkötalo Building – Μεγάλος διάδρομος με πολυθρόνες και τραπέζια. Στο πάτωμα υπάρχει μοκέτα σε διάφορα σε σημεία. Άνθρωποι που συνομιλούν και περπατούν.

Το σύνολο των μεμονωμένων ηχητικών γεγονότων προέρχεται από το DCASE 2016 task 2 [22] και αποτελείται από 11 κλάσεις με 20 δείγματα για την κάθε μία.

1. χτύπημα (knock)

2. άνοιγμα συρταριού (drawer)

3. καθάρισμα λαιμού (clearthroat)

4. τηλέφωνο (phone)

5. ήχος κλειδιών που πέφτουν (keysdrop)

6. ομιλία (speech)

7. πληκτρολόγιο (keyboard)

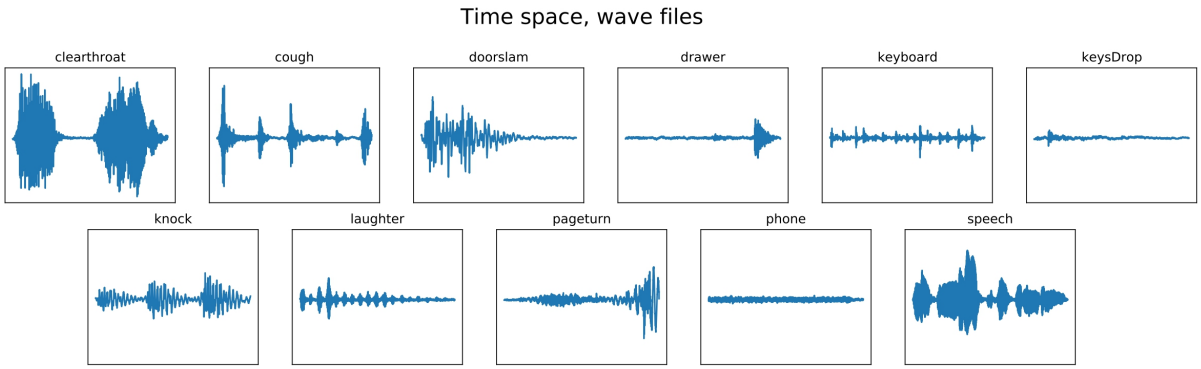
8. γύρισμα σελίδας (pageturn)

9. βήχας (cough)

10. δυνατό κλείσιμο πόρτας (doorslam)

11. γέλιο (laughter)

Κάθε ηχογράφηση δημιουργείται επιλέγοντας τυχαία ένα ηχητικό γεγονός και ορίζοντάς του χρόνο έναρξης και λήξης. Ύστερα γίνεται συνέλιξη του παραπάνω ήχου με την αντίστοιχη κρουστική απόκριση ώστε να "τοποθετηθούν" σε μία συγκεκριμένη απόσταση από το μικρόφωνο και στις κατάλληλες γωνίες αζιμούθιου και ανύψωσης. Επιπλέον στις μισές ηχογραφήσεις προστίθενται ήχοι έτσι ώστε να συνυπάρχουν έως δύο επικαλυπτόμενα ηχητικά γεγονότα.



Σχήμα 4.2: Παραδείγματα κυματομορφών των 11 διαφορετικών ηχητικών γεγονότων.

Κεφάλαιο 5

Τα Προβλήματα Ανίχνευσης και Εντοπισμού Ηχητικών Πηγών

5.1 Ανίχνευση Ηχητικών Γεγονότων

Για το πρόβλημα της ανίχνευσης των ηχητικών γεγονότων (SED) σκοπός είναι η πρόβλεψη των ενεργών κλάσεων σε κάθε χρονική στιγμή. Όπως αναφέρθηκε στο Κεφάλαιο 4 στο DCASE 2019 task 3 υπάρχουν 11 διαφορετικές κλάσεις και σε κάθε χρονική στιγμή μπορούν να συνυπάρχουν μέχρι δύο από αυτές. Το πρόβλημα εμπίπτει στην κατηγορία της ταξινόμησης πολλαπλών ετικετών (multi label classification). Οι μετρικές που χρησιμοποιούνται για την αξιολόγηση του SED είναι δύο, συγκεκριμένα το λάθος ταξινόμησης (Error Rate) και το F-Score [11] (βλέπε επίσης Σχήμα 5.1). Για τον υπολογισμό του F-Score χρησιμοποιούνται οι μετρικές Precision (P) και Recall (R), δηλαδή τα:

$$P = \frac{TP}{TP + FP},\ R = \frac{TP}{TP + FN} \tag{5.1}$$

όπου:

- TP: συμβολίζει τα ορθώς ανιχνευμένα γεγονότα (True Positives).
- FP: συμβολίζει γεγονότα που ανιχνεύθηκαν αλλά δεν είναι σωστά σύμφωνα με τον ορισμό (False Positives).
- FN: συμβολίζει γεγονότα που δεν ανιχνεύθηκαν σωστά σύμφωνα με τον ορισμό (False Negatives).

και στην συνέχεια υπολογίζεται το F-Score, ως

$$F = \frac{2P \cdot R}{P + R} \tag{5.2}$$

Το λάθος ταξινόμησης (Error Rate - ER) υπολογίζει το σύνολο το λαθών βάση των εισαγωγών (insertions 'I'), διαγραφών (deletions 'D'), και αντικαταστάσεων (substitutions 'S'):

$$ER = \frac{S + D + I}{N} \tag{5.3}$$

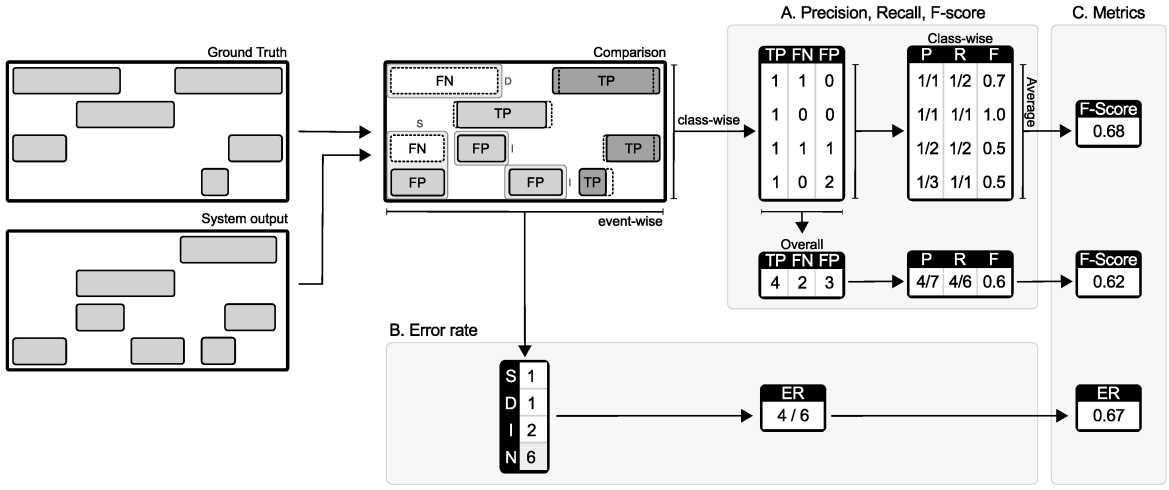
όπου:

- Αντικαταστάσεις S: Τα ηχητικά γεγονότα στην έξοδο του συστήματος που έχουν σωστή χρονική θέση αλλά λανθασμένη ετικέτα.

- Εισαγωγές I: Τα ηχητικά γεγονότα στην έξοδο του συστήματος που δεν είναι ούτε σωστά, ούτε αντικαταστάσεις.

- Διαγραφές D: Τα ηχητικά γεγονότα που υπάρχουν στα πραγματικά δεδομένα (ground truth) και δεν είναι ούτε σωστά ούτε αντικαταστάθηκαν.

- Γεγονότα αναφοράς N: Ο αριθμός των ηχητικών γεγονότων στα πραγματικά δεδομένα.



Σχήμα 5.1: Διάγραμμα υπολογισμού μετρικών. Εικόνα από [11].

5.2 Εντοπισμός Ηχητικής Πηγής

Το πρόβλημα του εντοπισμού της ηχητικής πηγής απαιτεί την εκτίμηση των σφαιρικών συντεταγμένων, αζιμούθιου και ανύψωσης, των ενεργών ηχητικών γεγονότων. Οι γωνίες του αζιμούθιου βρίσκονται σε ένα εύρος από -180° έως 170° και της ανύψωσης από -40° έως 40°. Είναι διακριτές τιμές και υπολογίζονται για κάθε 10°. Το πρόβλημα μπορεί να αντιμετωπιστεί είτε ως ταξινόμηση (classification) είτε ως γραμμική παλινδρόμηση (linear regression). Για την αξιολόγηση της κατεύθυνσης άφιξης (Direction of Arrival - DOA) χρησιμοποιούνται δυο μετρικές στο διαγωνισμό DCASE2019 Task 3, συγκεκριμένα το σφάλμα DOA (DOA error) και η "ανάκληση πλαισίου" (frame recall), όπως επεξηγείται στο [23]. Ένα ιδανικό μοντέλο για την εκτίμηση του DOA θα είχε σφάλμα DOA ίσο με 0° και frame recall ίσο με 100%.

5.3 Επεξεργασία των Δεδομένων

Η βάση δεδομένων είναι χωρισμένη σε 4 προκαθορισμένα τμήματα (folds). Για κάθε τμήμα ορίζονται το σύνολο εκπαίδευσης (training set), το σύνολο επιβεβαίωσης (validation set), και το σύνολο ελέγχου (test set). Ο παραπάνω διαχωρισμός γίνεται για να εφαρμοστεί η μέθοδος γενίκευσης διασταυρωμένης επικύρωσης (Cross-Validation) [24] για τα μοντέλα. Η τελική επίδοση θα είναι ο μέσος όρος των επιδόσεων σε κάθε τμήμα.

	Training Split	Validation Split	Test Split
Fold 1	3,4	2	1
Fold 2	4,1	3	2
Fold 3	1,2	4	3
Fold 4	2,3	1	4

Πίνακας 5.1: Διαχωρισμός δεδομένων βάσης για διασταυρωμένη επικύρωση.

Κάθε αρχείο ηχογράφησης της βάσης δεδομένων έχει την μορφή:
"split_[αριθμός]_ir_[αριθμός τοποθεσίας]_ov[αριθμός επικαλυπτόμενων ηχητικών γεγονότων]_[αύξων αριθμός].wav". Ακολουθείται από ένα αντίστοιχο αρχείο .csv όπου καταγράφονται οι ετικέτες για την συγκεκριμένη ηχογράφηση, του οποίου τα περιεχόμενα έχουν την εξής τυποποίηση: "ηχητικό γεγονός, χρόνος έναρξης, χρόνος λήξης, αζιμούθιο, ανύψωση",

για παράδειγμα “cough, 0.2, 0.5, 30, 25”

Στην συνέχεια γίνεται επεξεργασία της παραπάνω πληροφορίας για να είναι κατάλληλη για εισαγωγή σε ένα μοντέλο εκπαίδευσης. Δεν είναι τόσο εύκολο όσο ένα απλό πρόβλημα ταξινόμησης ήχου όπου κάθε αρχείο ηχογράφησης αντιστοιχεί σε μία κλάση. Κωδικοποιούμε λοιπόν τα παραπάνω σε δύο πίνακες εφάπαξ κωδικοποίησης (one-hot encoding), συγκεκριμένα:

1. Για τις ετικέτες SED: Πίνακας διαστάσεων (αριθμός χρονικών πλαισίων) $\times$ (αριθμός κλάσεων)

2. Για τις ετικέτες DOA: Πίνακας διαστάσεων (αριθμός χρονικών πλαισίων) $\times$ (2$\cdot$ αριθμός κλάσεων)

Ο συνολικός αριθμός χρονικών πλαισίων για κάθε ηχογράφηση, διάρκειας 60s, εξαρτάται από την συχνότητα δειγματοληψίας και το μέγεθος του άλματος (hop size) που εφαρμόζεται στον STFT. Στο πρόβλημα μας τα συνολικά πλαίσια είναι 3000 με μέγεθος άλματος 20ms και ρυθμό δειγματοληψίας στα 48kHz. Παραδείγματα των τελικών πινάκων που παράγονται με τις ετικέτες για κάθε υποεργασία, φαίνονται στους Πίνακες 5.2 και 5.3.

	classes	knock	drawer	clearthroat	phone	keysdrop	speech	keyboard	pageturn	cough	doorslam	laughter
		0	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
frames	***1***	0	0	0	0	0	0	0	0	1	0	0
	2	0	0	0	0	0	0	0	0	1	0	0
	3	1	0	0	1	0	0	0	0	0	0	0
	4	1	0	0	1	0	0	0	0	0	0	0
	5	0	0	1	0	0	0	0	0	0	0	0
	6	0	0	1	0	0	0	0	0	0	0	0
	...											
	3000	0	0	0	0	0	0	0	0	0	0	1

Πίνακας 5.2: Παράδειγμα one-hot κωδικοποίησης ετικετών SED.

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	class index	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
frames	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	40	0	0
	2	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	40	0	0
	3	10	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0
	4	10	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0
	5	0	20	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0
	6	0	20	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0
	...																						
	3000	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	60

Πίνακας 5.3: Παράδειγμα one-hot κωδικοποίησης ετικετών DOA.

Κεφάλαιο 6

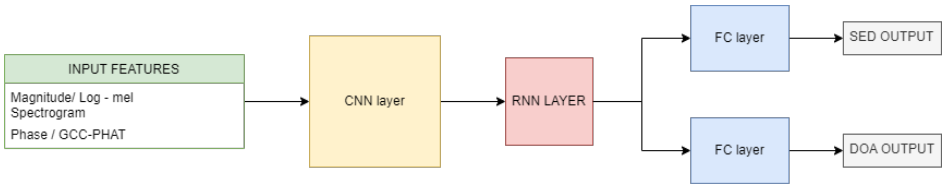
Μέθοδος Εκπαίδευσης

6.1 Εξαγωγή Χαρακτηριστικών

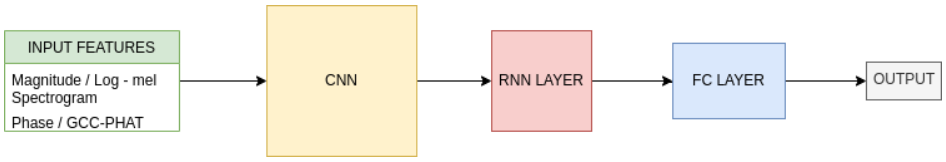
Η επιλογή των κατάλληλων χαρακτηριστικών αποτελεί πολύ σημαντικό κεφάλαιο στα προβλήματα που σχετίζονται με ήχο και επιλύονται με μηχανική μάθηση. Υπάρχει πληθώρα διαφορετικών λύσεων είτε στο πεδίο του χρόνου π.χ. κυματομορφές, είτε στο πεδίο της συχνότητας όπως είναι οι μετασχηματισμοί Fourier. Εμείς επιλέξαμε να δράσουμε στο πεδίο της συχνότητας, και πειραματιστήκαμε τόσο με κλασσικά φασματογραφήματα, όσο και με φασματογραφήματα Mel. Η τελική επιλογή ήταν τα δεύτερα με 64 φίλτρα Mel, τα οποία συμπληρώθηκαν από χαρακτηριστικά GCC-PHAT (Generalized Cross Correlation Phase Transform) [25] που θα βοηθήσουν πολύ στον εντοπισμό της ηχητικής πηγής. Η συχνότητα δειγματοληψίας του ηχητικού σήματος είναι 48kHz. Για τον αρχικό STFT χρησιμοποιείται παράθυρο Hann 1024 σημείων. Το μέγεθος του άλματος ορίζεται σε 20ms, χωρίς επικαλυπτόμενα χρονικά πλαίσια. Επομένως, για κάθε ηχητικό δείγμα 60 δευτερολέπτων λαμβάνουμε 3000 πλαίσια. Επιλέγουμε επιπλέον να διαιρέσουμε το σύνολο αυτό σε ακολουθίες των 128 πλαισίων. Για τον μετασχηματισμό GCC-PHAT ο αριθμός δειγμάτων καθυστέρησης ορίζεται και αυτός ίσος με 64, όσος και τα φίλτρα Mel. Για ένα ηχητικό σήμα 4 καναλιών υπολογίζονται 6 ζευγάρια GCC-PHAT. Συνεπώς η τελική μορφή των δεδομένων της εισόδου είναι: $128 \times 64 \times 10$.

6.2 Αρχιτεκτονικές

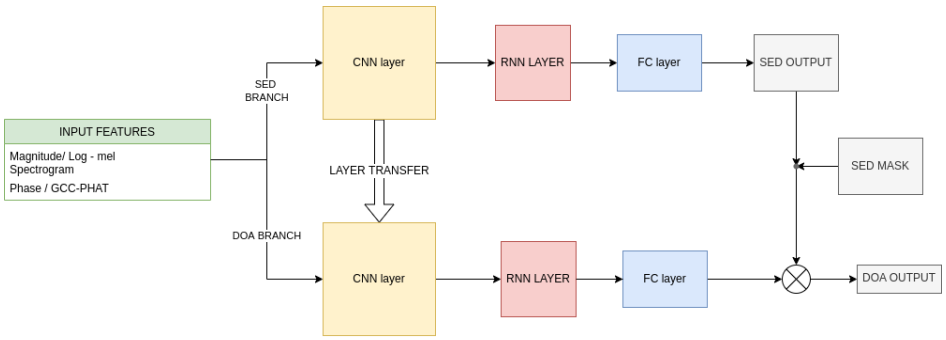
Για το πρόβλημα της αναγνώρισης και του εντοπισμού ήχου κατά την διάρκεια του DCASE 2019 προτάθηκαν διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων για την αντιμετώπιση του. Στην μέθοδο Baseline όπως αυτή περιγράφεται στο [26], ένα σύστημα είναι υπεύθυνο και για τις δύο υποεργασίες. Από την άλλη πλευρά υπάρχουν συστήματα που χρησιμοποιούν ξεχωριστούς κλάδους για την αναγνώριση και τον εντοπισμό, οι οποίοι συνεργάζονται μεταξύ τους [12]. Στα Σχήματα 6.1, 6.2, 6.3 παρουσιάζονται τα διαγράμματα των διαφορετικών αρχιτεκτονικών που θα εξετάσουμε.



Σχήμα 6.1: Αρχιτεκτονική #1, βασισμένη στην μέθοδο Baseline του DCASE 2019. Σχεδιάστηκε στο [9].



Σχήμα 6.2: Αρχιτεκτονική #2, ανεξάρτητος κλάδος SED / DOA. Σχεδιάστηκε στο [9].



Σχήμα 6.3: Αρχιτεκτονική #3, μέθοδος δύο σταδίων (“Two Stage”) από [12]. Σχεδιάστηκε στο [9].

6.3 Τα Επίπεδα του Νευρωνικού Δικτύου

Κάθε αρχιτεκτονική αποτελείται από τρία κύρια μπλοκ, τα οποία συνδυάζονται μεταξύ τους για τον υπολογισμό της εξόδου. Το πρώτο μπλοκ απαρτίζεται από συνδυασμό πολλαπλών επιπέδων CNN, ακολουθείται από το RNN επίπεδο που αποτελείται από περιφραγμένες επαναλαμβανόμενες μονάδες (Gated Recurrent Units - GRU), με τελευταίο επίπεδο σε ιεραρχία μια συστοιχία πυκνά συνδεδεμένων (Fully Connected - FC) νευρώνων, από τους οποίους θα παραχθεί και η τελική έξοδος.

6.3.1 Μπλοκ Συνελικτικών Νευρωνικών Δικτύων

Στο πρώτο επίπεδο θα χρησιμοποιήσουμε τα CNNs για να εξάγουμε χαρακτηριστικά από την είσοδο. Η ικανότητά τους να λειτουργούν σε όλα τα κανάλια της εισόδου μας δίνει την δυνατότητα να εξάγουμε διακαναλικά χαρακτηριστικά, απαραίτητα τόσο για την αναγνώριση όσο και για τον εντοπισμό του ήχου. Η κατασκευή τους γίνεται με τρεις διαφορετικούς τρόπους:

1. **Baseline**: Όπως αναφέρεται στην βασική μέθοδο του DCASE 2019 [26], είναι συνδυασμός τριών επιπέδων 2D-CNN με το καθένα να αποτελείται από 64 φίλτρα, με μέγεθος πυρήνα 3x3. Κάθε CNN ακολουθείται από κανονικοποίηση παρτίδας (Batch Normalization), την συνάρτηση κανονικοποίησης ReLU και το στρώμα μέγιστης συγκέντρωσης (Max Pooling).

Είσοδος (128, 64 , 10)
Conv2D: 64 filters, 3x3 kernel size Batch Normalization ReLU 1x4 Max Pooling
Conv2D: 64 filters, 3x3 kernel size Batch Normalization ReLU 1x4 Max Pooling
Conv2D: 64 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling

Πίνακας 6.1: Baseline CNN layers

2. **“Groups”**: Όπως αναφέρεται στην μέθοδο δύο επιπέδων [12], αποτελείται από 4 ζευγάρια 2D-CNN. Κάθε ζευγάρι έχει 64, 128, 256, 512 φίλτρα αντίστοιχα. Κάθε CNN ακολουθείται από κανονικοποίηση παρτίδας (Batch Normalization), την συνάρτηση κανονικοποίησης ReLU και το στρώμα μέγιστης συγκέντρωσης (Max Pooling).

Είσοδος (128, 64 , 10)
2xConv2D: 64 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling
2xConv2D: 128 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling
2xConv2D: 256 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling
Conv2D: 512 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling

Πίνακας 6.2: Επίπεδα CNN στη μέθοδο “Groups”.

3. **VGG16**: Είναι αντίγραφο των επιπέδων CNN της διάσημης αρχιτεκτονικής που χρησιμοποιείται για την ταξινόμηση εικόνων. Τα βάρη των επιπέδων δεν θα μεταφερθούν από το ήδη εκπαιδευμένο μοντέλο αλλά θα υπολογιστούν από την αρχή.

Είσοδος (128, 64 , 10)
2xConv2D: 64 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling
2xConv2D: 128 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling
3xConv2D: 256 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling
3xConv2D: 512 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling
3xConv2D: 512 filters, 3x3 kernel size Batch Normalization ReLU 1x2 Max Pooling

Πίνακας 6.3: Επίπεδα CNN του μοντέλου VGG16.

Σε κάθε περίπτωση η συνάρτηση Max-Pooling εφαρμόζεται κατά μήκος του άξονα της συχνότητας έτσι ώστε να διατηρείται ακέραιο το μήκος της ακολουθίας των χρονικών πλαισίων. Το μέγεθος των δεδομένων σε κάθε επίπεδο είναι:

- input: (L , B , C)
- 1st-Layer Conv2D: (L , B , F_1)
- 2nd-Layer Conv2D: (L , B / P_1 , F_2)
- …

- N_{th}-Layer Conv2D: (L , B / P_{n-1} , F_n)

όπου:

- L = Μέγεθος ακολουθίας
- B = Φίλτρα Mel
- C = Αριθμός καναλιών
- F = Μέγεθος φίλτρου
- P = Μέγεθος στρώματος συγκέντρωσης

Για παράδειγμα στην περίπτωση των "Groups" έχουμε:

- L = 128
- B = 64
- C = 10
- F = [64 , 128 , 256, 512]
- P = [2 , 2 , 2 , 2]. Άρα μετά το τελευταίο επίπεδο η έξοδος θα έχει διαστάσεις: $(128 \times 4 \times 512)$.

Πριν τροφοδοτηθεί το επόμενο μπλοκ θα χρειαστεί να αλλάξουμε τις διαστάσεις της εξόδου ώστε να γίνουν κατάλληλες για το επίπεδο RNN. Χρησιμοποιώντας την συνάρτηση Flatten του keras [27], η νέα έξοδος θα έχει μέγεθος: $128 \times (4 \cdot 512) = 128 \times 2048$. Σημαντικό είναι το γεγονός πως η διάσταση του χρόνου δεν έχει αλλάξει και παραμένει ίση με 128 πλαίσια.

6.3.2 Μπλοκ Επαναληπτικών Νευρωνικών Δικτύων

Για το RNN επίπεδο έγινε η επιλογή των GRUs [28] μεγέθους 256 και της συνάρτησης ενεργοποίησης $\tanh$. Το επίπεδο RNN χρησιμοποιείται κυρίως για την εξόρυξη χρονικών χαρακτηριστικών από την έξοδο των CNN.

6.3.3 Πλήρως Διασυνδεδεμένο Επίπεδο

Το τελευταίο επίπεδο είναι ένα σύνολο πλήρως συνδεδεμένων νευρώνων που θα μας δώσει την τελική έξοδο. Η συνάρτηση ενεργοποίησης εξαρτάται σε κάθε περίπτωση από την φύση του προβλήματος. Για την αναγνώριση του ήχου που απαιτεί την ταξινόμηση πολλαπλών ετικετών (multi-label classification), ταιριάζει η σιγμοειδής συνάρτηση ενεργοποίησης (sigmoid), ενώ για τον εντοπισμό της ηχητικής πηγής, που είναι ένα πρόβλημα γραμμικής παλινδρόμησης, χρησιμοποιείται η γραμμική (linear) συνάρτηση.

Κεφάλαιο 7

Πειράματα

7.1 Περιγραφή

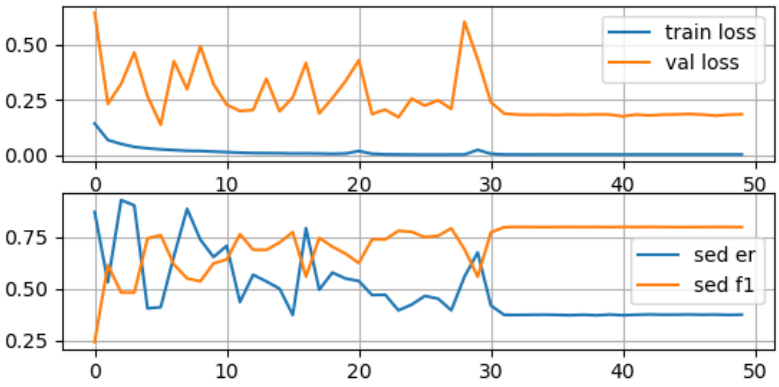
Θα πειραματιστούμε με όλες τις αρχιτεκτονικές και τα διαφορετικά επίπεδα CNN που αναφέρονται στο Κεφάλαιο 6. Για να έχουμε συγκρίσιμα αποτελέσματα και ο χρόνος εκπαίδευσης να είναι διαχειρίσιμος, θα εκπαιδεύσουμε κάθε μοντέλο για 50 εποχές. Ο ρυθμός εκπαίδευσης είναι 0.001 για τις 30 πρώτες εποχές, και θα μειώνεται κατά 10% για κάθε εποχή που απομένει. Επιλέγεται συντελεστής dropout ίσος με 0. Στην συνέχεια χρησιμοποιώντας την πιο αποδοτική αρχιτεκτονική θα υπολογίσουμε τα τελικά αποτελέσματα στα 4 διαφορετικά τμήματα (folds) που είναι χωρισμένη η βάση δεδομένων και θα εφαρμόσουμε την διασταυρωμένη επικύρωση (Cross-Validation). Τέλος θα απεικονίσουμε τις προβλέψεις του μοντέλου με τα κατάλληλα γραφήματα και θα σχολιάσουμε τις επιδόσεις του στην ανίχνευση των ηχητικών γεγονότων και στον εντοπισμό της ηχητικής πηγής.

7.2 Δοκιμή CNN στην Αναγνώριση Ηχητικών Γεγονότων

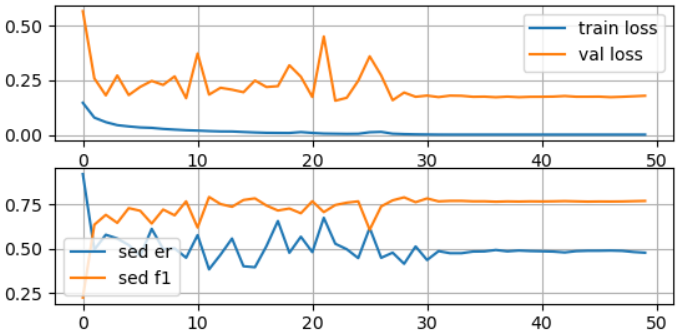
Χρησιμοποιώντας την αρχιτεκτονική#2 (Σχήμα 6.2) θα εκπαιδεύσουμε μοντέλα για τα τρία διαφορετικά CNN (Baseline, Groups, VGG16). Τα αποτελέσματα για κάθε μέθοδο παρουσιάζονται στον παρακάτω πίνακα:

CNN Architecture	Error Rate	F-score
Baseline	0.36%	0.79
Groups	0.55%	0.73
VGG16	0.42%	0.78

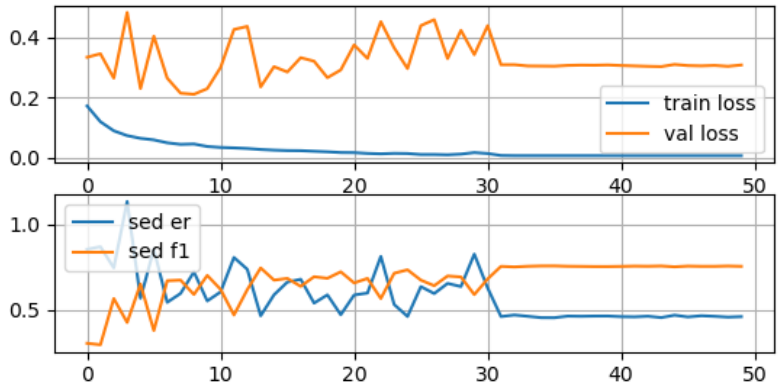
Πίνακας 7.1: Αποτελέσματα CNN στην ανίχνευση ηχητικών γεγονότων (SED), χρησιμοποιώντας την αρχιτεκτονική #2.



Σχήμα 7.1: Πορεία εκπαίδευσης για Baseline (SED).



Σχήμα 7.2: Πορεία εκπαίδευσης για Groups (SED).



Σχήμα 7.3: Πορεία εκπαίδευσης για VGG16 (SED).

Γίνεται εύκολα αντιληπτό πως υπάρχει πρόβλημα υπερπροσαρμογής (overfitting) του συστήματος στα δεδομένα ακόμα και στις πρώτες εποχές (<20), κυρίως λόγω του σχετικά μικρού μεγέθους δεδομένων για την φύση του προβλήματος. Με την μείωση του ρυθμού εκπαίδευσης μετά την 30η εποχή επιτυγχάνουμε μία μικρή βελτίωση, αν και πάλι η απόκλιση μεταξύ σετ εκπαίδευσης και σετ ελέγχου είναι σημαντική.

7.3 Δοκιμή CNN στον Εντοπισμό της Ηχητικής Πηγής

Με παρόμοιο τρόπο όπως στην Παράγραφο 7.2 θα συγκρίνουμε τις επιδόσεις κάθε συνόλου CNN στον εντοπισμό της ηχητικής πηγής. Να σημειωθεί πως κατά την διάρκεια της εκπαίδευσης χρησιμοποιούμε τις ετικέτες των ενεργών κλάσεων ως μάσκα έτσι ώστε το σφάλμα ταξινόμησης να υπολογίζεται μόνο στα χρονικά πλαίσια όπου υπάρχουν ενεργές κλάσεις. Αυτό σημαίνει πως η αποτελεσματικότητα της μεθόδου εξαρτάται σημαντικά από την επίδοση του μοντέλου αναγνώρισης (SED). Επομένως στα παρακάτω αποτελέσματα απουσιάζει η μετρική frame recall και αξιολογούμε τα μοντέλα μόνο βάση του σφάλματος εντοπισμού σε μοίρες.

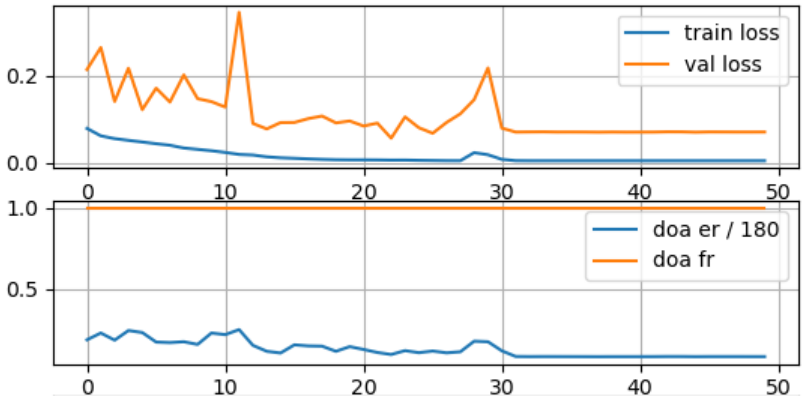
CNN Architecture	DOA Error
Baseline	14.7°
Groups	18.86°
VGG16	24.62°

Πίνακας 7.2: Αποτελέσματα CNN στον εντοπισμό της ηχητικής πηγής (DOA), χρησιμοποιώντας την αρχιτεκτονική #2.

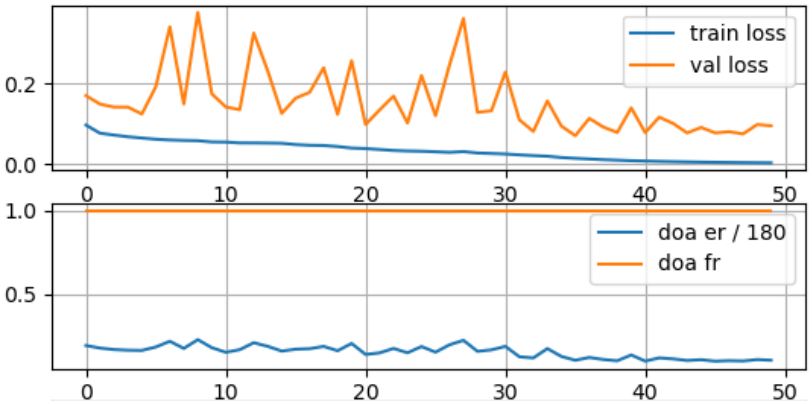
Είσοδος (128, 64 , 10)
Convolution Layer (baseline, groups, VGG16)
Flatten
GRU: 256, tanh
Dense: 256 units
Dense: 2N units
Multiply × Mask [SED labels]
Output (128, 2N)

Πίνακας 7.3: Μοντέλο εντοπισμού (DOA), N:κλάσεις

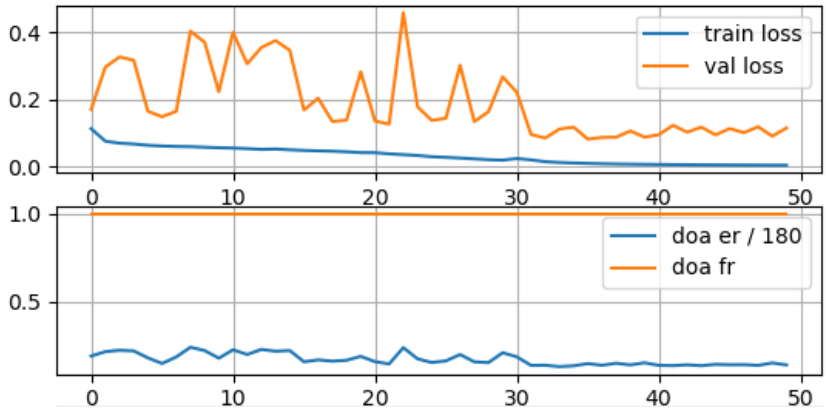
Στις παρακάτω εικόνες παρουσιάζονται και οι πορείες εκπαίδευσης για κάθε μοντέλο.



Σχήμα 7.4: Πορεία εκπαίδευσης για Baseline (DOA).



Σχήμα 7.5: Πορεία εκπαίδευσης για Groups (DOA).



Σχήμα 7.6: Πορεία εκπαίδευσης για VGG16 (DOA).

7.4 Μεταφορά των επιπέδων CNN

Για την υλοποίηση της αρχιτεκτονικής#3 της παραγράφου 6.2 θα μεταφερθούν τα ήδη εκπαιδευμένα βάρη από τα επίπεδα CNN του κλάδου SED στον κλάδο DOA. Τα βάρη των RNN και FC επιπέδων θα ανανεώνονται κανονικά. Στα αποτελέσματα που παρουσιάζονται στον Πίνακα 7.4 παρατηρούμε μικρή βελτίωση τόσο στο Baseline όσο και στα Groups. Αυτό σημαίνει πως η πληροφορία που μεταφέρθηκε από τον κλάδο της ανίχνευσης (SED) είχε θετική επιρροή στα αποτελέσματα του εντοπισμού (DOA).

CNN Architecture	DOA Error
Baseline	14.57°
Groups	19.27°
VGG16	22.64°

Πίνακας 7.4: Αποτελέσματα μεταφοράς επιπέδων CNN για τον εντοπισμό της ηχητικής πηγής.

7.5 Συνδυασμός των κλάδων SED και DOA

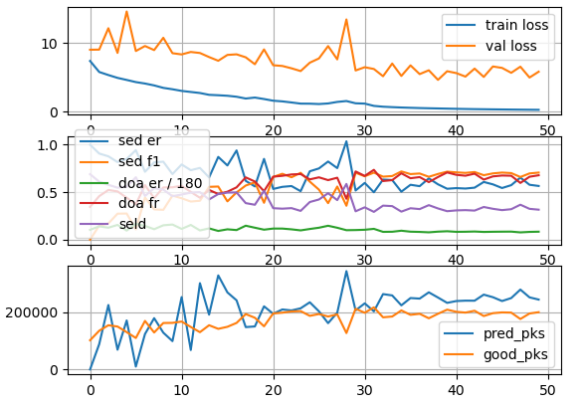
Τελικό βήμα για την πλήρη εφαρμογή της αρχιτεκτονικής#3 είναι ο συνδυασμός των δύο κλάδων για να παραχθούν οι τελικές εκτιμήσεις εντοπισμού και αναγνώρισης. Στο κλάδο DOA δεν χρησιμοποιούνται πια οι ετικέτες αναφοράς από τον κλάδο SED, αλλά οι δύο κλάδοι λειτουργούν συνδυαστικά. Στον Πίνακα 7.5 παρουσιάζονται τα συνολικά αποτελέσματα.

CNN Architecture	SED Error	F-score	DOA Error	Frame Recall	SELD Score
Baseline	0.36	0.79	11.63°	0.79	0.21
Groups	0.55	0.72	12.93°	0.66	0.31
VGG16	0.42	0.78	18.56°	0.72	0.25
Baseline (layer transfer)	0.36	0.79	11.54°	0.79	0.21
Groups (layer transfer)	0.55	0.73	13.38°	0.66	0.31
VGG16 (layer transfer)	0.42	0.78	16.9°	0.72	0.25

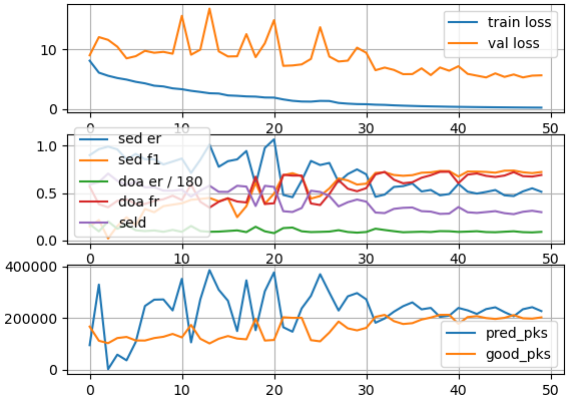
Πίνακας 7.5: Συγκεντρωτικά αποτελέσματα CNN στην ανίχνευση και τον εντοπισμό των ηχητικών γεγονότων με ξεχωριστούς κλάδους SED και DOA.

7.6 Κοινά Επίπεδα CNN - RNN για Ανίχνευση και Εντοπισμό

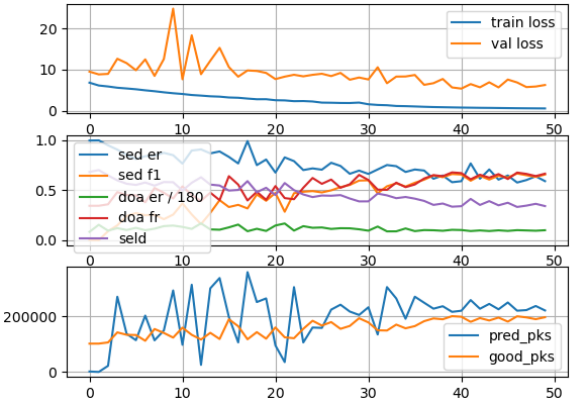
Τέλος θα γίνει αξιολόγηση των διαφορετικών επιπέδων CNN στην βασική αρχιτεκτονική#1, που αναφέρεται ως Baseline στο [26]. Σε αυτήν την μέθοδο δεν έχουμε ξεχωριστά παρακλάδια για κάθε υποπρόβλημα, αλλά τα επίπεδα CNN και RNN είναι κοινά και για τα δύο. Παρακάτω παρουσιάζονται οι πορείες εκπαίδευσης και τα τελικά αποτελέσματα. Η μόνη διαφορά στα γραφήματα εκπαίδευσης είναι ότι παρουσιάζονται συγκεντρωτικά τα σφάλματα ταξινόμησης για την ανίχνευση (SED) και τον εντοπισμό των ηχητικών γεγονότων (DOA).



Σχήμα 7.7: Πορεία εκπαίδευσης για Baseline (SELD).



Σχήμα 7.8: Πορεία εκπαίδευσης για Groups (SELD).



Σχήμα 7.9: Πορεία εκπαίδευσης για VGG16 (SELD).

CNN Architecture	SED Error	F-score	DOA Error	Frame Recall	SELD Score
Baseline	0.57	0.7	14.66°	0.66	0.32
Groups	0.55	0.7	15.0°	0.67	0.31
VGG16	0.60	0.66	16.41°	0.66	0.34

Πίνακας 7.6: Αποτελέσματα κοινών επιπέδων CNN-RNN για ανίχνευση και εντοπισμό.

Στον Πίνακα 7.6 φαίνονται οι επιδόσεις των επιπέδων CNN στην αρχιτεκτονική#1. Σε αντίθεση με τα αποτελέσματα που παρουσιάστηκαν στην παράγραφο 7.5 (Πίνακας 7.5 τα Groups υπερτερούν των Baseline και VGG16. Συμπεραίνουμε όμως πως η αρχιτεκτονική#1 είχε πολύ χειρότερα αποτελέσματα σε όλες τις μετρικές σε σχέση τόσο με την αρχιτεκτονική#3 όσο και με την αρχιτεκτονική#2. Άρα η εκπαίδευση κοινών επιπέδων και για τα δύο υποπροβλήματα υπολείπεται των μεθόδων όπου η εκπαίδευση κάθε κλάδου πραγματοποιείται ξεχωριστά.

7.7 Τελικά Αποτελέσματα

Τα τελικά αποτελέσματα και για τα 4 τμήματα (folds) που είναι χωρισμένη η βάση δεδομένων, θα τα εξάγουμε με την αρχιτεκτονική#3 χρησιμοποιώντας τα Baseline επίπεδα CNN, που ήταν και η πιο αποδοτική μέθοδος αφού παρουσίασε τα καλύτερα αποτελέσματα σε όλες τις μετρικές.

Fold	SED Error	F-Score	DOA Error	Frame Recall	SELD Score
Fold 1	0.36	0.79	11.63°	0.79	0.21
Fold 2	0.45	0.74	21.88°	0.77	0.26
Fold 3	0.34	0.81	29.41°	0.83	0.21
Fold 4	0.38	0.78	22.26°	0.83	0.22
Average	0.38	0.78	21.29°	0.83	0.22
Baseline	0.35	0.80	30.8°	0.84	0.22

Πίνακας 7.7: Τελικά αποτελέσματα διασταυρωμένης επικύρωσης (Cross-Validation).

7.7.1 Αποτελέσματα Ανίχνευσης Ηχητικών Γεγονότων

Παρατηρώντας τα αποτελέσματα των προβλέψεων βλέπουμε διαφορές ανάλογα με τον αριθμό τον ενεργών κλάσεων. Είναι αρκετά λογικό ότι με περισσότερες κλάσεις ενεργές είναι πιο δύσκολο για το μοντέλο να τις αναγνωρίσει και να τις κατηγοριοποιήσει σωστά, πόσο μάλλον όταν από την φύση τους τα ηχητικά γεγονότα μοιάζουν μεταξύ τους. Στα Σχήματα 7.10 και 7.11 φαίνονται οι προβλέψεις με έντονο χρώμα και πιο αχνά οι ετικέτες αναφοράς. Σύμφωνα με τα παραδείγματα των ηχογραφήσεων που παρουσιάζονται στις παραπάνω εικόνες εξάγονται τα εξής συμπεράσματα για κάθε κλάση:

- χτύπημα (knock), δείκτης: 0. Πολύ Καλή επίδοση.
- συρτάρι (drawer), δείκτης: 1. Πολύ Καλή επίδοση.
- καθάρισμα λαιμού (clearthroat), δείκτης: 2. Καλή επίδοση, στιγμιαία σύγχυση με την κλάση 'ομιλία' (speech).
- τηλέφωνο (phone), δείκτης: 3. Καλή επίδοση, στιγμιαία σύγχυση με την κλάση **'κλειδιά' (keysdrop)**.

- **κλειδιά (keysdrop), δείκτης: 4**. Στο συγκεκριμένο παράδειγμα δεν υπάρχει η κλάση στις ετικέτες αναφοράς. Παρόλα αυτά παρατηρούμε ότι εμφανίζεται μαζί με τις κλάσεις 'τηλέφωνο' (phone) και 'πληκτρολόγιο' (keyboard).

- ομιλία (speech), δείκτης: 5. Καλή επίδοση, μικρή σύγχυση με την κλάση 'καθάρισμα λαιμού' (clearthroat).

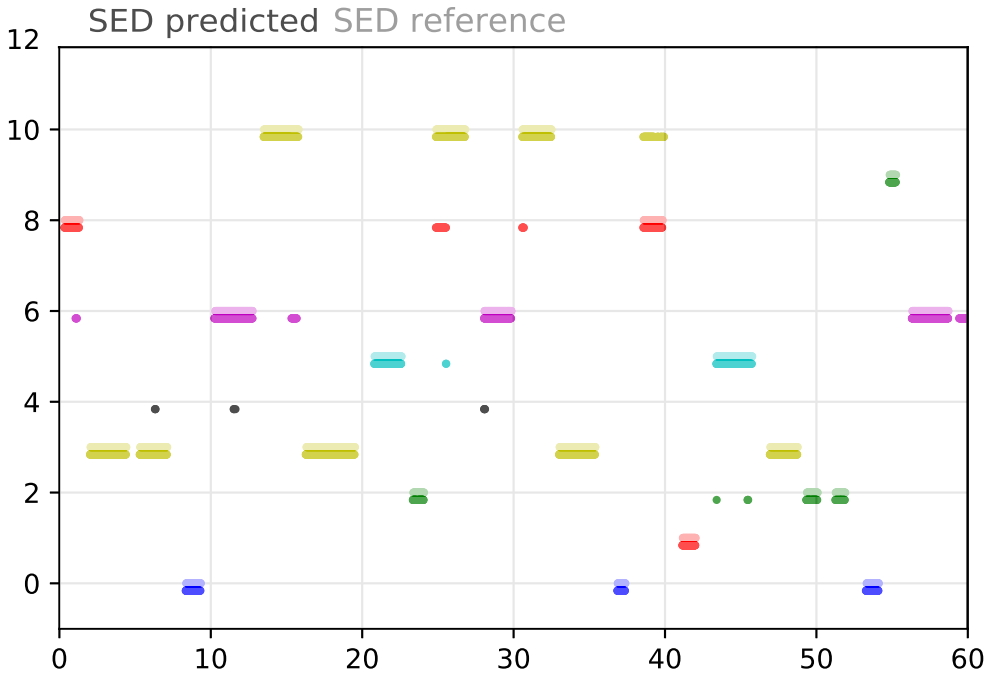
- πληκτρολόγιο (keyboard), δείκτης: 6. Μέτρια επίδοση, υπάρχει σύγχυση με τις κλάσεις 'βήχας' (cough) , 'γέλιο' (laughter) , **'κλειδιά' (keysdrop)**.

- γύρισμα σελίδας (pageturn), δείκτης:7. Στο παράδειγμα χωρίς επικαλυπτόμενα ηχητικά γεγονότα (Σχήμα 7.10) η συγκεκριμένη κλάση δεν εμφανίζεται ούτε ως ετικέτα αναφοράς ούτε ως πρόβλεψη. Επομένως σε αυτήν την περίπτωση είναι λιγότερη πιθανή η σύγχυσή της με άλλες κλάσεις. Στην περίπτωση όπου υπάρχει επικάλυψη ηχητικών γεγονότων (Σχήμα 7.11) ανιχνεύεται σωστά όταν υπάρχει στα δεδομένα αναφοράς, αλλά η λανθασμένη εμφάνισή της μαζί με τις υπόλοιπες κλάσεις είναι πολύ πιο συχνή.

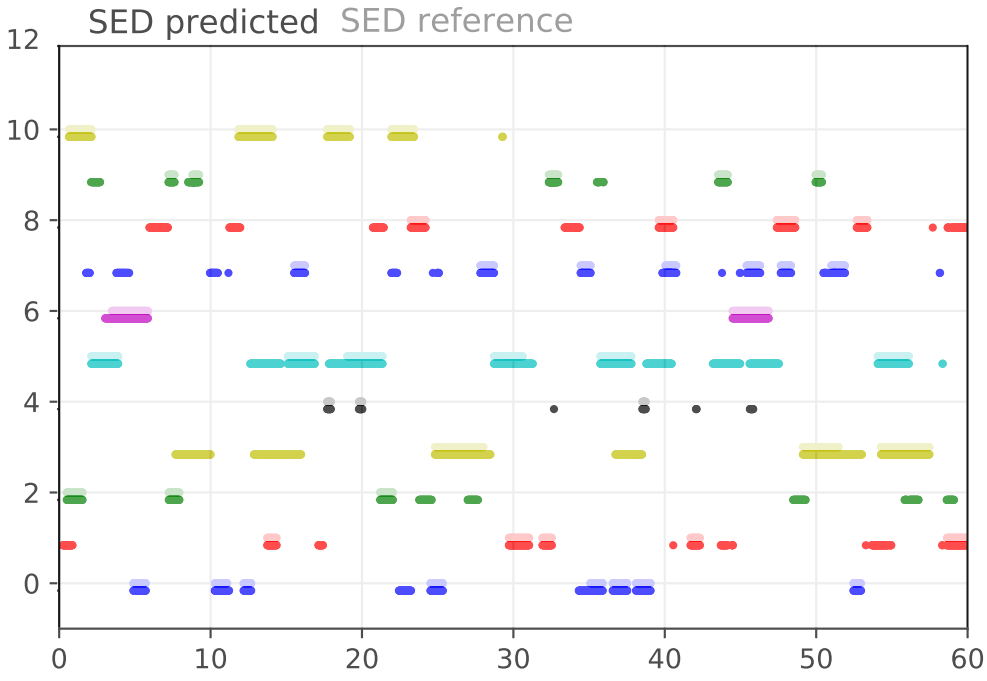
- χτύπημα πόρτας (doorslam), δείκτης:9. Πολύ καλή επίδοση.

- γέλιο (laughter), δείκτης:10. Μέτρια επίδοση, σύγχυση με τις κλάσεις 'πληκτρολόγιο' (keyboard) και 'βήχας' (cough).

Συμπερασματικά, γίνονται κάποια λάθη στην σωστή ανίχνευση των ηχητικών γεγονότων και έχουν κυρίως την μορφή εισαγωγών (insertions) παρά αντικαταστάσεων (substitutions). Το μοντέλο είναι αρκετά αποτελεσματικό στο να ανιχνεύει σωστά μία κλάση όταν είναι ενεργή, ακόμη και στην περίπτωση όπου συνυπάρχουν δύο διαφορετικά γεγονότα. Μια παρενέργεια της παραπάνω συμπεριφοράς είναι πως πολλές φορές μαζί με αυτήν την κλάση, ανιχνεύεται και ένα ηχητικό γεγονός που δεν υπάρχει στην πραγματικότητα. Πολύ σπάνια εμφανίζεται ηχητικό γεγονός όταν δεν υπάρχει ενεργή κλάση. Πιο συγκεκριμένα και σύμφωνα με το Κεφάλαιο 5, τα ψευδώς θετικά (False Positives) είναι πολύ περισσότερα από τα ψευδώς αρνητικά (False Negatives).



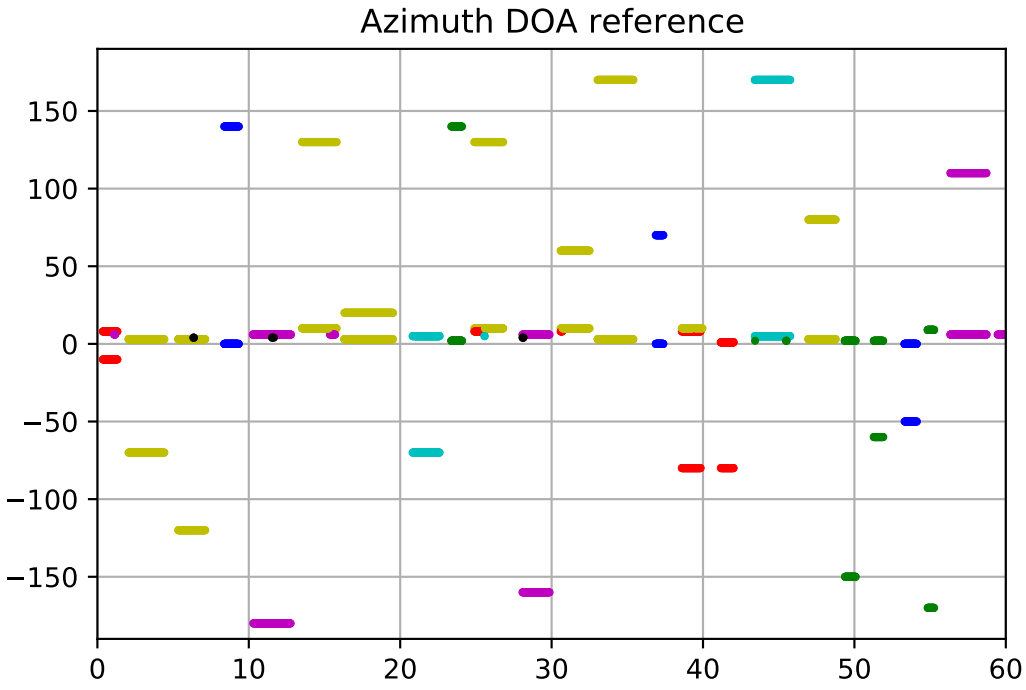
Σχήμα 7.10: Ανίχνευση χωρίς να επικαλύπτονται ηχητικά γεγονότα. Όνομα αρχείου 'split1_ir0_ov1_1'.



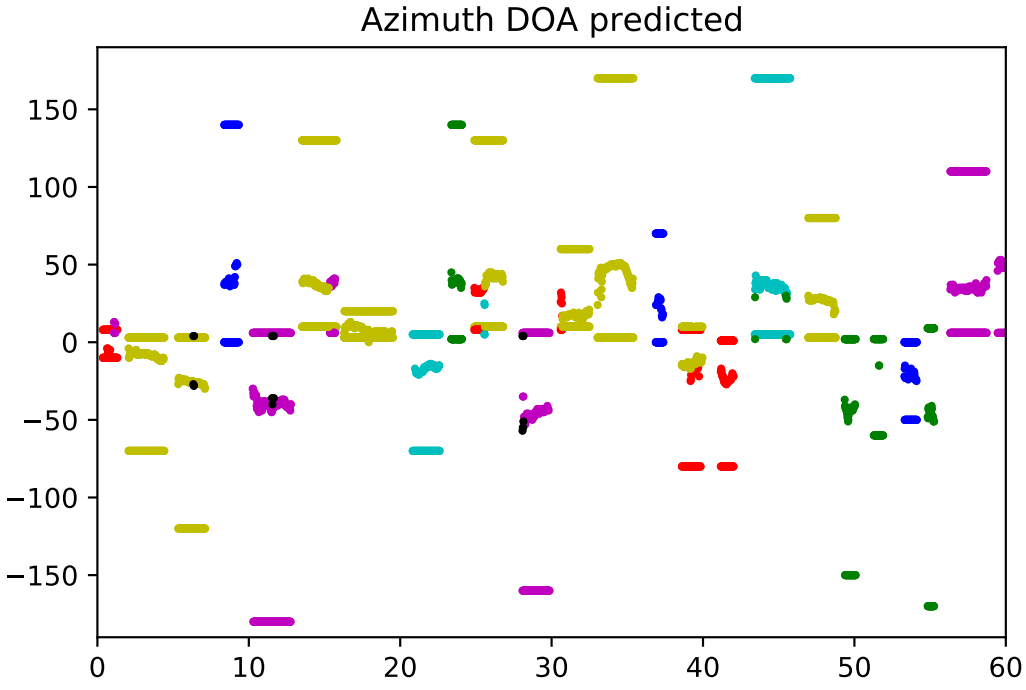
Σχήμα 7.11: Ανίχνευση όπου τα ηχητικά γεγονότα επικαλύπτονται. Όνομα αρχείου 'split1_ir1_ov2_38'.

7.7.2 Αποτελέσματα Εντοπισμού της Ηχητικής Πηγής

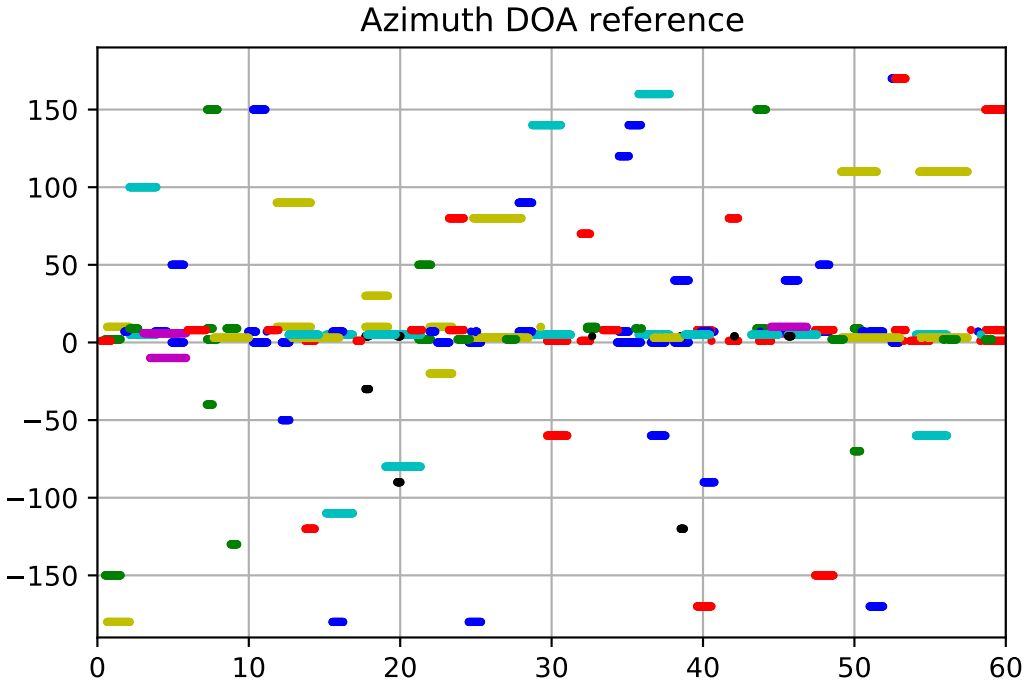
Στο πρόβλημα του εντοπισμού της ηχητικής πηγής το μοντέλο τα πήγε πολύ καλύτερα από την βασική μέθοδο κάτι που αποτυπώνεται κυρίως στην μετρική DOA Error στον Πίνακα 7.7. Φαίνεται πως η προσθήκη των χαρακτηριστικών GCC-PHAT συνέβαλε αρκετά σε αυτό το αποτέλεσμα. Στα Σχήματα 7.13, 7.15, 7.17 και 7.19 φαίνονται τα αποτελέσματα των εκτιμήσεων των συντεταγμένων αζιμούθιου και ανύψωσης. Όπως και στο πρόβλημα της ανίχνευσης οι επιδόσεις είναι πιο ικανοποιητικές στην περίπτωση που δεν επικαλύπτονται τα ηχητικά γεγονότα.



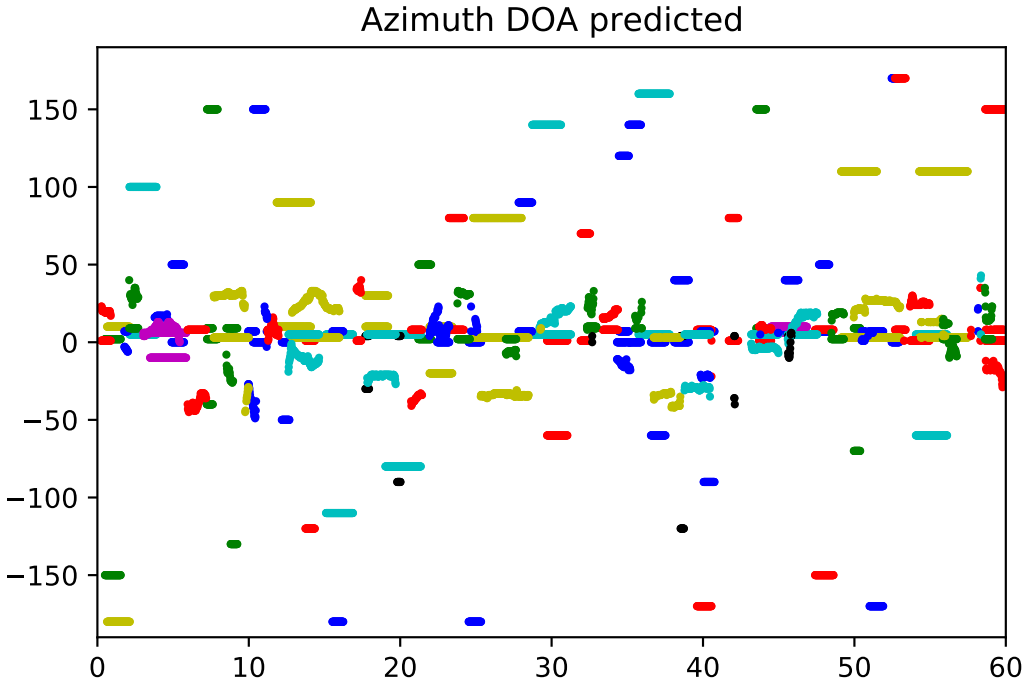
Σχήμα 7.12: Αζιμούθιο, τιμές αναφοράς χωρίς επικαλυπτόμενα γεγονότα.



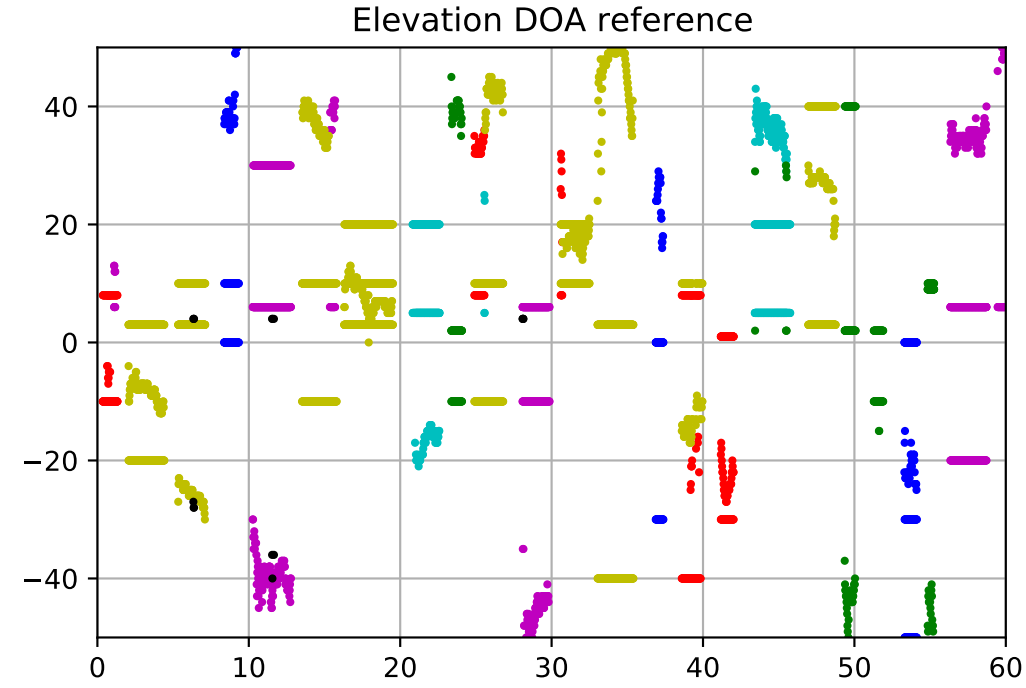
Σχήμα 7.13: Αζιμούθιο, εκτίμηση τιμών χωρίς επικαλυπτόμενα γεγονότα.



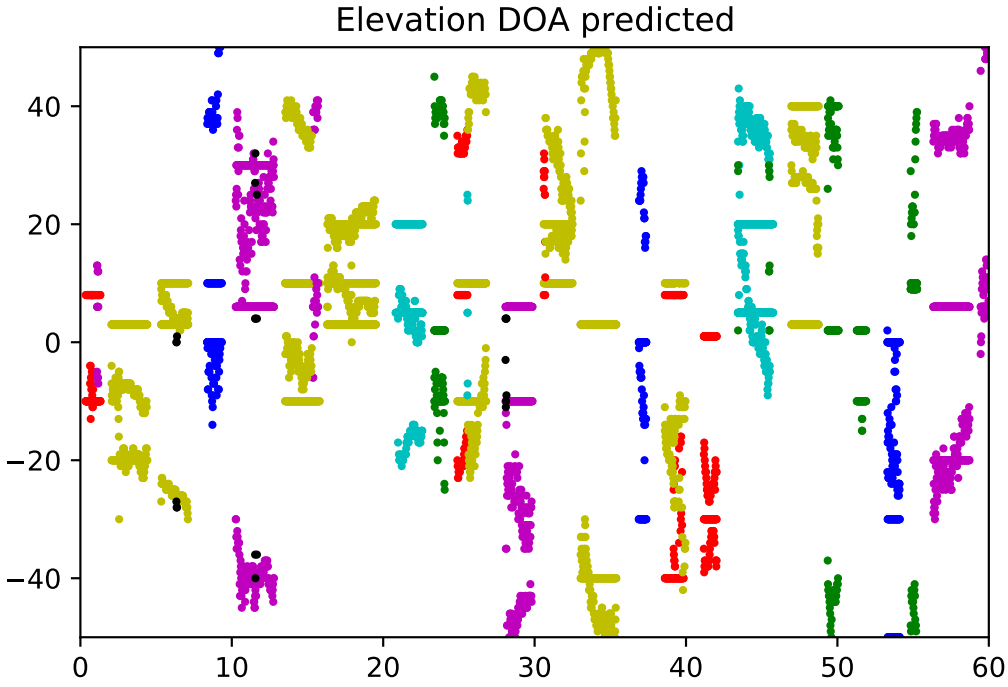
Σχήμα 7.14: Αζιμούθιο, τιμές αναφοράς με επικαλυπτόμενα γεγονότα.



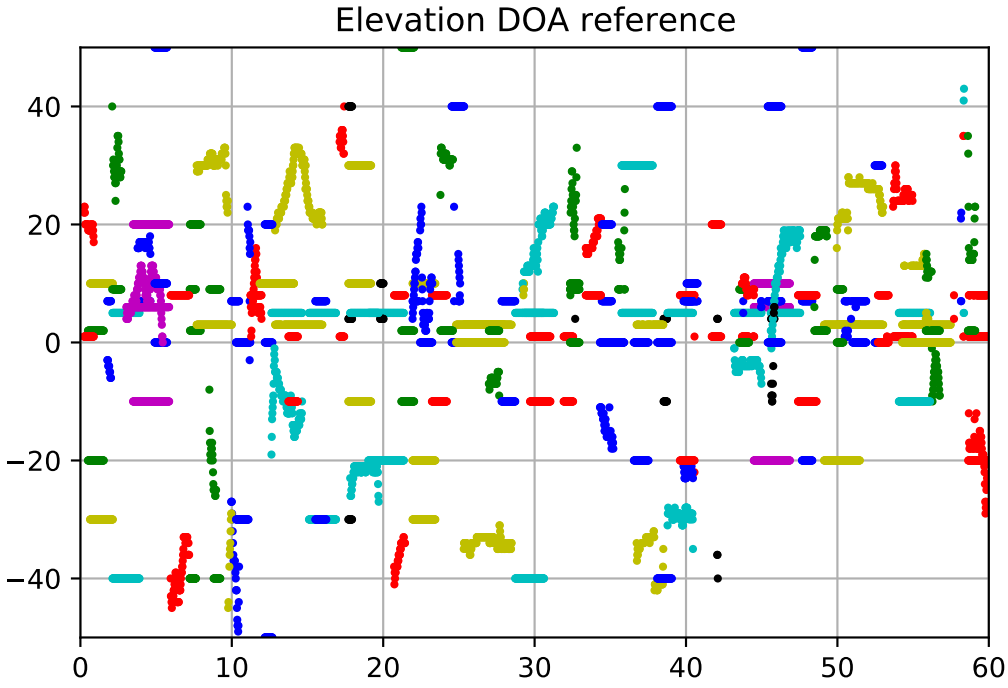
Σχήμα 7.15: Αζιμούθιο, εκτίμηση τιμών με επικαλυπτόμενα γεγονότα.



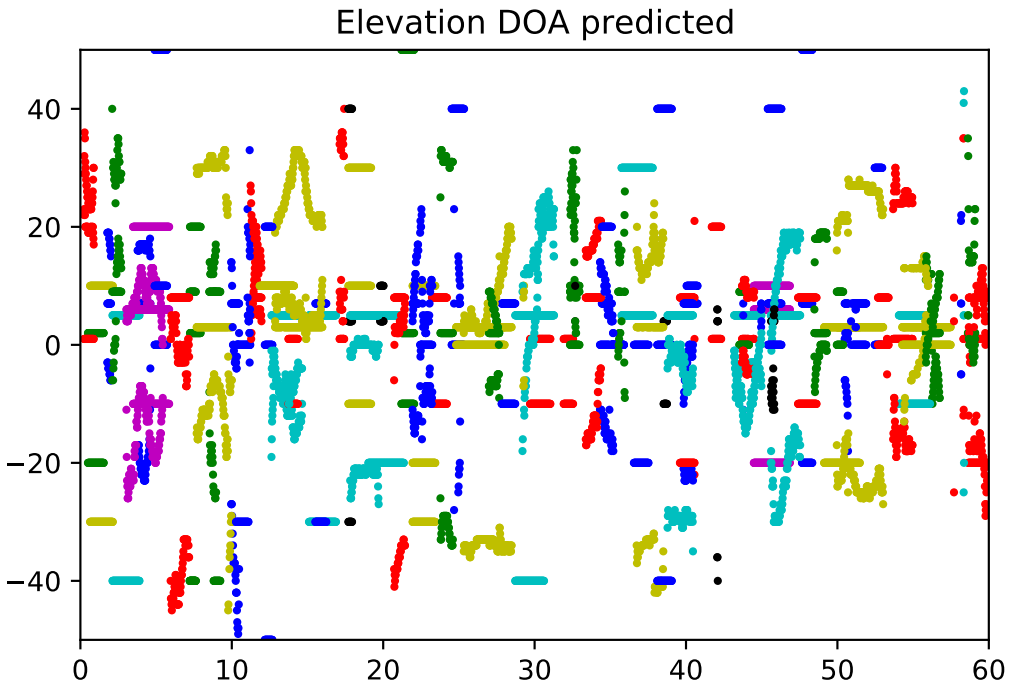
Σχήμα 7.16: Ανύψωση, τιμές αναφοράς χωρίς επικαλυπτόμενα γεγονότα.



Σχήμα 7.17: Ανύψωση, εκτίμηση τιμών χωρίς επικαλυπτόμενα γεγονότα.



Σχήμα 7.18: Ανύψωση, τιμές αναφοράς με επικαλυπτόμενα γεγονότα.



Σχήμα 7.19: Ανύψωση, εκτίμηση τιμών με επικαλυπτόμενα γεγονότα.

Κεφάλαιο 8

Συμπεράσματα

Στα πλαίσια αυτής της διπλωματικής εργασίας μελετήσαμε την αναγνώριση και τον εντοπισμό ηχητικών γεγονότων με την χρήση συνελικτικών και επαναληπτικών νευρωνικών δικτύων. Επίσης πειραματιστήκαμε με τα χαρακτηριστικά εισόδου και συγκρίναμε διαφορετικές αρχιτεκτονικές βαθιάς μάθησης για την αντιμετώπιση του προβλήματος.

8.1 Σύνοψη

Με τον συνδυασμό CNN, RNN δικτύων και την χρήση των χαρακτηριστικών GCC-PHAT καταφέραμε να πλησιάσουμε την απόδοση της βασικής (Baseline) μεθόδου και να βελτιώσουμε σημαντικά την μετρική του σφάλματος εντοπισμού (DOA Error). Σε απόδοση υπερίσχυσαν οι αρχιτεκτονικές όπου οι κλάδοι για κάθε υποπρόβλημα εκπαιδεύονται ξεχωριστά (Σχήματα 6.2 και 6.3). Επιπλέον η μεταφορά των βαρών του επιπέδου CNN από τον κλάδο SED στον κλάδο DOA είχε θετικό αντίκτυπο αν και οι διαφορές ήταν μικρές.

8.2 Μελλοντικές Επεκτάσεις

Τα περιθώρια βελτίωσης των μοντέλων για προβλήματα εντοπισμού και ανίχνευσης ηχητικών γεγονότων είναι τεράστια και υπάρχει πλήθος επιλογών για πειραματισμό. Ένα μεγάλο πρόβλημα που αντιμετωπίσαμε ήταν η υπερπροσαρμογή (overfitting) του μοντέλου [29] λόγω του μικρού μεγέθους των δεδομένων για την φύση του προβλήματος. Κρίνεται επομένως απαραίτητη η αύξηση του όγκου τους και η μελέτη διάφορων μεθόδων για ενίσχυσή τους (data augmentation) [30]. Τέλος, στα πλαίσια των ετήσιων DCASE workshops

η πολυπλοκότητα του προβλήματος αυξάνεται συνεχώς, ώστε να προσομοιώνονται καλύτερα οι συνθήκες που επικρατούν στην καθημερινή ζωή. Επεκτείνεται έτσι σε περιπτώσεις με περισσότερες από 2 ενεργές κλάσεις, με ηχητικά γεγονότα που εκπέμπουν ταυτόχρονα από διαφορετικά σημεία στον χώρο και που μπορούν να κινούνται ελεύθερα σε αυτόν.

Βιβλιογραφία

[1] DCASE 2019, Task 3 Sound Event Localization and Detection. https://dcase.community/challenge2019/task-sound-event-localization-and-detection.

[2] Virtanen Tuomas, Plumbley Mark D., and Ellis Dan. *Computational Analysis of Sound Scenes and Events*. Springer Cham, 1st edition, 1993.

[3] Νευρώνας - Βικιπαίδεια. https://el.wikipedia.org/wiki/Νευρώνας.

[4] Akshay L Chandra. McCulloch-Pitts Neuron — Mankind's first mathematical model of a biological neuron. https://towardsdatascience.com/mcculloch-pitts-model-5fdf65ac5dd1.

[5] Akshay L Chandra. Perceptron: The Artificial Neuron (An essential upgrade to the McCulloch-Pitts neuron). https://towardsdatascience.com/perceptron-the-artificial-neuron-4d8c70d5cc8d.

[6] Koutroumbas S., Theodoridis K. *Αναγνώριση Προτύπων*. Π.Χ. ΠΑΣΧΑΛΙΔΗΣ, Αθήνα, 4η edition, 2009.

[7] Jitendra Tembhurne and Tausif Diwan. Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. *Multimedia Tools and Applications*, 80:1–40, 02 2021.

[8] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *CoRR*, abs/1511.08458, 2015.

[9] https://www.diagrams.net/.

[10] https://mhacoustics.com/.

[11] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162, 2016.

[12] Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, and Mark Plumbley. Polyphonic sound event detection and localization using a two-stage strategy. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York University, 2019.

[13] Sainath Adapa. Urban sound tagging using convolutional neural networks. *CoRR*, abs/1909.12699, 2019.

[14] H. Wang and P. Chu. Voice source localization for automatic camera pointing system in videoconferencing. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 187–190, 1997.

[15] Selina Chu, Shrikanth Narayanan, and C.-C. Jay Kuo. Environmental sound recognition with time-frequency audio features. *Trans. Audio, Speech and Lang. Proc.*, 17(6):1142–1158, 2009.

[16] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, and Tuomas Virtanen. Sound event detection in multichannel audio using spatial and harmonic features. *CoRR*, abs/1706.02293, 2017.

[17] Janghoon Cho, Sungrack Yun, Hyoungwoo Park, Jungyun Eum, and Kyuwoong Hwang. Acoustic scene classification based on a large-margin factorized CNN. *CoRR*, abs/1910.06784, 2019.

[18] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, and Xavier Serra. Audio tagging with noisy labels and minimal supervision. *CoRR*, abs/1906.02975, 2019.

[19] W. McCulloch and W. Piits. A logical calculus and the ideas of immanent in the nervous activity. *Bulletin of Mathematical Biophysic*, 5:115–133, 1943.

[20] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark B. Sandler. A comparison on audio signal preprocessing methods for deep neural networks on music tagging.

[21] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. A multi-room reverberant dataset for sound event localization and detection. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pages 10–14, New York University, NY, USA, October 2019.

[22] DCASE 2016, Task 2 Sound Event Detection in synthetic audio. `https://dcase.community/challenge2016/task-sound-event-detection-in-synthetic-audio#citation`.

[23] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1462–1466. IEEE, 2018.

[24] George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, 12(1):49–57, 2010.

[25] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[26] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, March 2018.

[27] Keras: The python deep learning API. `https://keras.io/`.

[28] Rui Lu and Zhiyao Duan. Bidirectional GRU for sound event detection. Technical report, DCASE2017 Challenge, September 2017.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[30] Janek Ebbers and Reinhold Häb-Umbach. Convolutional recurrent neural network and data augmentation for audio tagging with noisy labels and minimal supervision. In

Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019). New York University, 2019.

[31] Κ.Διαμαντάρας. *Τεχνητά Νευρωνικά Δίκτυα*. Κλειδάριθμος, Αθήνα, 2007.

[32] Amidi Afshine and Amidi Shervine. A detailed example of how to use data generators with keras. `https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly`.

[33] Utilities for detection and classification of acoustic scenes. `https://dcase-repo.github.io/dcase_util/index.html`.

[34] Sotirios Panagiotis Chytas and Gerasimos Potamianos. Hierarchical detection of sound events and their localization using convolutional neural networks with adaptive thresholds. Technical report, DCASE2019 Challenge, June 2019.