ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΙΧΝΕΥΣΗ ΚΑΙ ΕΝΤΟΠΙΣΜΟΣ ΗΧΗΤΙΚΩΝ ΣΥΜΒΑΝΤΩΝ ΣΕ ΠΕΡΙΒΑΛΛΟΝΤΑ ΚΑΤΕΥΘΥΝΤΙΚΗΣ ΠΑΡΕΜΒΟΛΗΣ

Διπλωματική Εργασία

Χάιδω Πουλιάνου

Επιβλέπων: Γεράσιμος Ποταμιάνος

Ιούλιος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΙΧΝΕΥΣΗ ΚΑΙ ΕΝΤΟΠΙΣΜΟΣ ΗΧΗΤΙΚΩΝ ΣΥΜΒΑΝΤΩΝ ΣΕ ΠΕΡΙΒΑΛΛΟΝΤΑ ΚΑΤΕΥΘΥΝΤΙΚΗΣ ΠΑΡΕΜΒΟΛΗΣ

Διπλωματική Εργασία

Χάιδω Πουλιάνου

Επιβλέπων: Γεράσιμος Ποταμιάνος

Ιούλιος 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

SOUND EVENT DETECTION AND LOCALIZATION WITH DIRECTIONAL INTERFERENCE

Diploma Thesis

Chaido Poulianou

Supervisor: Gerasimos Potamianos

July 2022

Εγκρίνεται από την Επιτροπή Εξέτασης:

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

Μέλος **Μπέλλας Νικόλαος**
Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Αργυρίου Αντώνιος**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Γεράσιμο Ποταμιάνο για το χρόνο του και την καθοδήγησή του σε οποιοδήποτε πρόβλημα είχα κατά τη διάρκεια εκπόνησης της διπλωματικής εργασίας. Ευχαριστώ επίσης τους καθηγητές της επιτροπής κ. Νικόλαο Μπέλλα και κ. Αντώνιο Αργυρίου για την παρεύρεσή τους και τη βοήθειά τους κατά τη διάρκεια των σπουδών μου. Ευχαριστώ επίσης όλους τους καθηγητές μου που κατά τη διάρκεια των σπουδών μου οι οποίοι δε δίστασαν να βοηθήσουν σε οποιαδήποτε απορία και δυσκολία αντιμετώπιζα. Τέλος, ευχαριστώ την οικογένεια και τους φίλους μου, οι οποίοι με υποστήριξαν στις πιο δύσκολες στιγμές και χωρίς αυτούς δε θα μπορούσα να τα βγάλω πέρα.

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».

Η Δηλούσα

Χάιδω Πουλιάνου

Διπλωματική Εργασία

ΑΝΙΧΝΕΥΣΗ ΚΑΙ ΕΝΤΟΠΙΣΜΟΣ ΗΧΗΤΙΚΩΝ ΣΥΜΒΑΝΤΩΝ ΣΕ ΠΕΡΙΒΑΛΛΟΝΤΑ ΚΑΤΕΥΘΥΝΤΙΚΗΣ ΠΑΡΕΜΒΟΛΗΣ

Χάιδω Πουλιάνου

Περίληψη

Το αντικείμενο έρευνας της διπλωματικής αφορά το πρόβλημα της αναγνώρισης ηχητικών συμβάντων και της εύρεσης των σφαιρικών συντεταγμένων τους στο χώρο. Για το σκοπό αυτό, χρησιμοποιήσαμε νευρωνικά δίκτυα, με στόχο να ξεπεραστεί σε ακρίβεια το μοντέλο βάσης που παρέχεται από το διαγωνισμό DCASE 2021. Η προσέγγιση, το αρχικό νευρωνικό μοντέλο, η εξαγωγή των χαρακτηριστικών των αρχείων ήχου καθώς και ο τρόπος καταμέτρησης της ακρίβειας του κάθε μοντέλου βασίστηκε στον περσινό διαγωνισμό DCASE 2021 Task 3. Τα μοντέλα που χρησιμοποιήθηκαν για να επιτευχθούν οι παραπάνω στόχοι βασίστηκαν στα μοντέλα ResNet και Conformer. Αξιοποιήθηκαν επιπλέον τεχνικές βελτιστοποίησης, όπως της επαύξησης των δεδομένων εισόδου και του συνδυασμού πολλαπλών μοντέλων. Το καλύτερο μοντέλο που βρέθηκε μετά από πειράματα κατάφερε να ξεπεράσει σε όλες τις μετρικές αξιολόγησης το μοντέλο βάσης.

Diploma Thesis

SOUND EVENT DETECTION AND LOCALIZATION WITH DIRECTIONAL INTERFERENCE

Chaido Poulianou

Abstract

The subject of this thesis concerns the problem of sound event detection and localization in spatial spherical coordinates. For this purpose, we employ neural networks, aiming to outperform the baseline model provided as part of the DCASE 2021 competition. The approach, the initial baseline neural network used, the feature extraction of the sound files, as well as the evaluation metrics were based on last year's DCASE 2021 Task 3 challenge. The neural networks used to achieve our goals were based on the ResNet and Conformer models. For further improvements, various methods were utilized such as data augmentation and average ensembling techniques. The best performing model in our experiments managed to surpass the baseline in all evaluation metrics.

Πίνακας περιεχομένων

Ευχαριστίες **ix**

Περίληψη **xii**

Abstract **xiii**

Πίνακας περιεχομένων **xv**

Κατάλογος σχημάτων **xvii**

Κατάλογος πινάκων **xix**

Συντομογραφίες **xxi**

1 Εισαγωγή **1**

1.1 Αντικείμενο της διπλωματικής 2

1.2 Σχετικές εργασίες 2

1.3 Συνεισφορά 4

1.4 Οργάνωση της διπλωματικής 5

2 Διαγωνισμός DCASE 2021 Task 3 **7**

2.1 Σύντομη περιγραφή του Task 3 8

2.2 Βάση δεδομένων 10

2.3 Μοντέλο βάσης 12

2.4 Μετρικές αξιολόγησης συστήματος 14

3 Προσέγγιση του προβλήματος **15**

3.1 Μοντέλα 16

3.1.1 Προ-επεξεργασία και υπερ-παράμετροι 16
3.1.2 Πρώτη προσέγγιση: ResNets 16
3.1.3 Δεύτερη προσέγγιση: Conformer 19
3.1.4 Τρίτη προσέγγιση: ResNet-Conformer 23
3.2 Περαιτέρω τεχνικές 24
3.2.1 Επαύξηση δεδομένων 24
3.2.2 Συγχώνευση μοντέλων 27

4 Πειράματα **29**
4.1 Πειράματα με CNN 30
4.2 Πειράματα με ResNet 31
4.3 Πειράματα με Conformer 36
4.4 Πειράματα με ResNet-Conformer 39
4.5 Πειράματα με συγχώνευση μοντέλων 40

5 Συμπεράσματα **45**

Βιβλιογραφία **47**

Κατάλογος σχημάτων

2.1 Επεξήγηση των διανυσμάτων ACCDOA 10
2.2 Γραφική απεικόνιση μιας συνθετικής ηχογράφησης. 11
2.3 Δομή του μοντέλου βάσης του διαγωνισμού DCASE 2021 Task 3, βασισμένη στο SELDnet (Εικόνα από [1]). 13

3.1 Η δομή του πρώτου SELD μοντέλου που χρησιμοποιήθηκε με ResNet18. . 18
3.2 Μπλοκ Conformer 21
3.3 (Πάνω) Μονάδα εμπρόσθιας τροφοδότησης. (Κάτω) Μονάδα συνέλιξης (Εικόνες από [2]). 21
3.4 Απεικόνιση των μετασχηματισμών του SpecAugment. 26
3.5 Απεικόνιση αλγόριθμου ανταλλαγής ηχητικών καναλιών 27

4.1 Απεικόνιση προβλέψεων μοντέλου βάσης. 34
4.2 Απεικόνιση προβλέψεων ResNet34v2. 35
4.3 Απεικόνιση προβλέψεων Squeeze-Conformer. 38
4.4 Απεικόνιση προβλέψεων ResNet34-Conformer. 41

Κατάλογος πινάκων

4.1 Πειράματα με CNN 31
4.2 Πειράματα πάνω στο ResNet18 32
4.3 Πειράματα πάνω στο ResNet34 32
4.4 Σύγκριση των βέλτιστων συστημάτων ResNet. 33
4.5 Πειράματα πάνω σε Conformer 37
4.6 Πειράματα με ResNet-Conformer 40
4.7 Πειράματα πάνω σε συγχώνευση μοντέλων 42

Συντομογραφίες

ACCDOA	Activity-Coupled Cartesian Direction of Arrival (Καρτεσιανή Κατεύθυνση της Άφιξης Δραστηριότητας σε Σύζευξη)
Bi-GRU	Bidirectional Gated Recurrent Network (Αμφίδρομο Επαναληπτικό Δίκτυο με Μηχανισμό Πύλης)
CNN	Convolutional Neural Network (Συνελικτικό Νευρωνικό Δίκτυο)
CRNN	Convolutional Recurrent Neural Network (Συνελικτικό Επαναληπτικό Νευρωνικό Δίκτυο)
DCASE	Detection and Classification of Acoustic Scenes and Events (Ανίχνευση και Ταξινόμηση Ακουστικών Σκηνών και Συμβάντων)
DNN	Deep Neural Network (Βαθύ Νευρωνικό Δίκτυο)
DOA	Direction of Arrival (Κατεύθυνση της Άφιξης)
FFN	Feed Forward Network (Δίκτυο Εμπρόσθιας Τροφοδότησης)
FFT	Fast Fourier Transform (Γρήγορος Μετασχηματισμός Fourier)
FOA	First-Order Ambisonic (Αμφισωνική Πρώτης Τάξης)
GCC-PHAT	Generalized Cross-Correlation Sequences (Γενικευμένες Ακολουθίες Ετεροσυσχέτισης)
GLU	Gated Linear Unit (Γραμμική Μονάδα με Μηχανισμό Πύλης)
IR	Impulse Response (Κρουστική Απόκριση)
LM	Language Model (Γλωσσικό Μοντέλο)
LSTM	Long Short-Term Memory (Μακριά Βραχυπρόθεσμη Μνήμη)
MHSA	Multi-Head Self-Attention (Πολυκέφαλη Αυτο-Προσοχή)
MIC	Microphone Array (Μικροφωνικό Δίκτυο)
MSE	Mean Square Error (Μέσο Τετραγωνικό Σφάλμα)
MUSIC	Multiple Signal Classification (Πολλαπλή Ταξινόμηση Σήματος)
ResNet	Residual Network (Υπολειμματικό Δίκτυο)

ReLU	Rectified Linear Unit (Διορθωτική Γραμμική Μονάδα)
RNN	Recurrent Neural Network (Επαναληπτικό Νευρωνικό Δίκτυο)
SED	Sound Event Detection (Ανίχνευση Ακουστικών Γεγονότων)
SELD	Sound Event Localization and Detection (Εντοπισμός και Ανίχνευση Ακουστικών Γεγονότων)
SGD	Stochastic Gradient Descent (Στοχαστική Κάθοδος Κλίσης)
SNR	Signal-to-Noise Ratio (Σηματο-Θορυβικός Δείκτης)
SRIR	Spatial Room Impulse Response (Κρουστική Απόκριση Χωρικού Δωματίου)
STFT	Short-Time Fourier Transform (Μετασχηματισμός Fourier Βραχέως Χρόνου)
SWA	Stochastic Weight Averaging (Στοχαστική Εύρεση Μέσου Όρου Βαρών)
TTA	Test-Time Augmentation (Επαύξηση κατά το Στάδιο Ελέγχου)

Κεφάλαιο 1

Εισαγωγή

Ο ήχος περιέχει αρκετή χρήσιμη πληροφορία, με την οποία μπορούμε να κατανοήσουμε καλύτερα το περιβάλλον και τι συμβαίνει σε αυτό (αναγνώριση σκηνών). Το πρόβλημα της αναγνώρισης ηχητικών συμβάντων και του εντοπισμού τους στο χώρο (SELD) απαρτίζεται από δύο υποπροβλήματα, την αναγνώριση ηχητικών συμβάντων (SED) και την εύρεση της κατεύθυνσης αφίξεώς τους (DOA). Το πρώτο στοχεύει στην αναγνώριση του είδους των ηχητικών συμβάντων και της εύρεση της χρονικής διάρκειας δραστηριότητάς τους που λαμβάνουν χώρα σε ένα ηχητικό αρχείο. Όσον αφορά το δεύτερο πρόβλημα, σκοπός του είναι η εύρεση των σφαιρικών συντεταγμένων της πηγής των ηχητικών αυτών συμβάντων. Έτσι η επίλυση του SELD δίνει ως αποτέλεσμα μια χωρο-χρονική περιγραφή του περιβάλλοντος στο οποίο εκτελείται, κάτι το οποίο μπορεί να χρησιμοποιηθεί για την ανάπτυξη ενός μεγάλου εύρους εφαρμογών. Τέτοιες περιλαμβάνουν αυτόματα συστήματα περιήγησης με περιορισμένο οπτικό πεδίο, αυτοματοποιημένα συστήματα ασφαλείας, συστήματα αναπαράστασης ακουστικών σκηνών, εφαρμογές ρομποτικής (βιομηχανικής και μη) [3], [4], με σκοπό την καλύτερη αλληλεπίδραση των ρομποτικών βοηθών με τον άνθρωπο, συστήματα εξωτερικής και εσωτερικής παρακολούθησης [5], καθώς και εφαρμογές σε έξυπνα σπίτια με την ενσωμάτωση συστημάτων SELD σε έξυπνους βοηθούς όπως η Alexa της Amazon [6].

Λόγω του μεγάλου εύρους εφαρμογών του, το πρόβλημα του SELD αποτελεί έναν ενεργό κλάδο έρευνας, ο οποίος έχει προσελκύσει το ενδιαφέρον του επιστημονικού χώρου ιδιαίτερα τα τελευταία χρόνια. Οι επιστήμονες μέχρι και την τελευταία δεκαετία προσέγγιζαν το θέμα επίλυσης του SELD ξεχωριστά για κάθε υποπρόβλημά του, αξιοποιώντας διάφορους αλγορίθμους μηχανικής μάθησης για την επίλυση του SED και παραμετρικές τεχνικές για την εύρεση του DOA. Ωστόσο, η συνεχής έρευνα στο θέμα απέδειξε την υπεροχή των με-

θόδων της βαθιάς μάθησης, συγκεκριμένα των νευρωνικών δικτύων, για την επίλυση του SELD.

1.1 Αντικείμενο της διπλωματικής

Ο σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι η δημιουργία ενός συστήματος που αξιοποιεί νευρωνικά δίκτυα, το οποίο θα μπορεί να χρησιμοποιηθεί για την αναγνώριση ηχητικών συμβάντων και εύρεση της κατεύθυνσης αφίξεώς τους σε περιβάλλοντα με παρεμβολές. Επιπλέον, το σύστημα θα πρέπει να ξεπερνάει σε επίδοση το μοντέλο βάσης που προσφέρεται από το διαγωνισμό του DCASE 2021 Task 3 [7]. Ο διαγωνισμός αυτός από το 2019 έχει ενσωματώσει σε ένα από τα προβλήματα των προκλήσεών του και την επίλυση του προβλήματος του SELD. Στόχος του διαγωνισμού είναι η εύρεση μεθόδων σε επίπεδο επεξεργασίας ήχου αλλά και νευρωνικών μοντέλων που να μπορούν να λύνουν ικανοποιητικά το πρόβλημα αυτό καθώς ταυτόχρονα και να ξεπερνάει σε ακρίβεια το μοντέλο βάσης που δίνεται από το διαγωνισμό. Το παρεχόμενο μοντέλο βασίζεται στο νευρωνικό δίκτυο SELDnet το οποίο παρουσιάσθηκε στο [8] και κάνει χρήση της ACCDOA μορφής εξόδου, η οποία περιγράφεται στο [9], για τις τελικές προβλέψεις.

Εκτός από την ανάπτυξη ενός ικανοποιητικού συστήματος SELD, χρησιμοποιήθηκε και η τεχνική επαύξησης δεδομένων για καλύτερη γενίκευση του αναπτυγμένου συστήματος. Επίσης, προσπαθήσαμε να συνδυάσουμε τα πολλαπλά μοντέλα που αναπτύξαμε με διάφορους τρόπους. Κάθε μοντέλο εκπαιδεύτηκε και αξιολογήθηκε στη βάση δεδομένων που παρέχεται από το διαγωνισμό DCASE 2021 Task 3. Στα δεδομένα αυτά εμφανίζονται επικαλυπτόμενα ηχητικά συμβάντα καθώς και παρεμβολές και θόρυβος.

1.2 Σχετικές εργασίες

Η λύση του προβλήματος SELD μέχρι πρόσφατα γινόταν ξεχωριστά για τα υποπροβλήματα του SED και DOA. Η αντιμετώπιση του SED συχνά επιλέγονταν να γίνει με διάφορους αλγορίθμους ταξινόμησης βασισμένους στη μηχανική μάθηση, όπως των μηχανών διανυσματικής στήριξης, των Γκαουσιανών μοντέλων μίξης, του κανόνα του πλησιέστερου γείτονα [10], [11], αλλά και με αλγορίθμους επεξεργασίας σημάτων [12]. Όσον αφορά την προσέγγιση του DOA, μέχρι και αυτή τη δεκαετία χρησιμοποιούνταν συχνά παραμετρικές μέ-

θοδοι [11]. Κάποιες από αυτές είναι οι αλγόριθμοι MUSIC και ESPRIT [13], [14]. Ωστόσο, υπάρχει ακόμα αρκετή ανάγκη για περαιτέρω μελέτη και έρευνα, καθώς σε πραγματικά περιβάλλοντα όπου υπάρχουν επικαλυπτόμενες πηγές ήχου και παρεμβολές με χαμηλό SNR, οι αλγόριθμοι αυτοί δεν λειτουργούν με ικανοποιητική ακρίβεια, με αποτέλεσμα να ταξινομούν λανθασμένα τα ηχητικά συμβάντα ή ακόμα και να μην τα ανιχνεύουν καν. Το πρόβλημα της αναγνώρισης ηχητικών συμβάντων σε τέτοια περίπλοκα περιβάλλοντα που εμφανίζουν επικάλυψη αναφέρεται και ως πολυφωνική αναγνώριση. Επίσης, ο υπολογισμός του DOA με βάση τις παραμετρικές μεθόδους απαιτεί μια πολύ καλή εκτίμηση του ακριβούς αριθμού παρόντων ηχητικών πηγών. Έτσι δεν αποδίδει ικανοποιητικά εκτιμήσεις DOA σε περίπλοκα περιβάλλοντα. Για αυτό το λόγο απαιτείται περαιτέρω έρευνα στο θέμα και η εύρεση αποτελεσματικών αλγορίθμων.

Για την αντιμετώπιση αυτού του προβλήματος, τα τελευταία χρόνια έχει αποδειχθεί ότι η χρήση μεθόδων που βασίζονται στη βαθιά μάθηση αντί των παραδοσιακών μεθόδων μπορούν να δώσουν καλύτερα αποτελέσματα. Μία από τις πρώτες τέτοιες προσεγγίσεις αναφέρεται στο [15]. Αυτή η προσέγγιση εμπνέεται από την παραμετρική μέθοδο του συστήματος MUSIC [13] αλλά χρησιμοποιεί DNN ως τη βάση αρχιτεκτονικής του συστήματος. Η χρήση DNN απεδείχθη πολύ καλύτερη σε επίδοση από το αρχικό σύστημα MUSIC. Αυτό οδήγησε την επιστημονική κοινότητα στην προσέγγιση του θέματος επίλυσης του SELD [11], ειδικά για προβλήματα πολλαπλών κλάσεων, σε ανάπτυξη συστημάτων που χρησιμοποιούν νευρωνικά δίκτυα. Η χρήση αυτών των υπολογιστικών συστημάτων για την επίλυση του προβλήματος SELD έχει αποδειχθεί ότι είναι αρκετά επιτυχής σε σχέση με τους αλγορίθμους μηχανικής μαθήσεως που χρησιμοποιούνταν ως τώρα [10].

Αρκετά μοντέλα προσεγγίζουν το πρόβλημα του SELD με την ταυτόχρονη επίλυση των SED και DOA αντί της ξεχωριστής αντιμετώπισής τους [8], [16], [17]. Τέτοιου είδους μοντέλα παράγουν εξαρχής κοινά χαρακτηριστικά για την κοινή πρόβλεψη των DOA και των αναγνωρισμένων συμβάντων. Ακόμα μια προσέγγιση είναι η χρήση κοινών παραμέτρων για ξεχωριστή πρόβλεψη SED και DOA [18], καθώς έχει μελετηθεί πως τα δύο υποπροβλήματα είναι αλληλοεξαρτώμενα και για καλύτερα αποτελέσματα απαιτείται ο συνδυασμός της εξαγόμενης πληροφορίας τους.

Όσον αφορά την αρχιτεκτονική των συστημάτων SELD, τα πιο επιτυχημένα μοντέλα είναι αυτά που χρησιμοποιούν βαθιά μοντέλα CRNN, τα οποία εκτός της χρήσης τους για προβλήματα επεξεργασίας βίντεο έχουν αποβεί αποτελεσματικά και στο πρόβλημα SELD.

Αν και τα μοντέλα αυτά πλεονεκτούν των απλών αλγορίθμων ταξινόμησης, σε προσομοιώσεις που αγγίζουν αρκετά πραγματικές συνθήκες απαιτείται ακόμα αρκετή έρευνα για ένα ικανοποιητικό σύστημα. Αυτό συμβαίνει γιατί σε τέτοια περιβάλλοντα το πρόβλημα της επικάλυψης και των παρεμβολών είναι αρκετά πιο συχνό. Ένας σημαντικός παράγοντας που έχει συνεισφέρει σε μεγάλο βαθμό πάνω στην έρευνα επίλυσης του SELD αποτελεί ο ετήσιος διαγωνισμός DCASE [19], στον οποίο βασίστηκε και το θέμα μελέτης της συγκεκριμένης διπλωματικής.

1.3 Συνεισφορά

Η συγκεκριμένη διπλωματική στοχεύει στη σχεδίαση ενός νευρωνικού μοντέλου που να επιλύει επιτυχώς το πρόβλημα SELD και πιο συγκεκριμένα να ξεπερνάει σε ακρίβεια το μοντέλο βάσης του διαγωνισμού DCASE 2021 Task 3. Για το σκοπό αυτό:

- Μελετήθηκαν σχετικά μοντέλα από έρευνες που εκπονήθηκαν στα πλαίσια του διαγωνισμού DCASE Task 3 των τελευταίων δύο χρόνων, οι οποίες εκτός από πρωτότυπα μοντέλα CRNN εκμεταλλεύτηκαν επίσης και την τεχνική της αυτο-προσοχής (self-attention) και αρχιτεκτονικές Transformer ή και Conformer. Πιο συγκεκριμένα, η έρευνα των [20], [21], [22], [23], [24], [25] αποτέλεσε έμπνευση για την προσέγγιση του προβλήματος με παραλλαγές των μοντέλων ResNet [23], [24], [25] και Conformer [20], [21], [22], [23], [26].

- Εφαρμόσαμε διάφορες αλλαγές και πειραματισμούς πάνω στις υπερ-παραμέτρους των μοντέλων και τεχνικές που επηρεάζουν την ακρίβεια στα αποτελέσματα ενός μοντέλου, όπως τεχνικές επαύξησης των δεδομένων εισόδου (data augmentation), αλλαγή των αποκωδικοποιητών του δικτύου (decoders) καθώς και τεχνικές συνδυασμού αυτών (ensembling).

- Αξιολογήσαμε κάθε μοντέλο που αναπτύξαμε πάνω στη βάση δεδομένων που παρέχεται από το διαγωνισμό DCASE 2021 Task 3 και συγκρίναμε την επίδοσή τους με το μοντέλο βάσης.

Πιο συγκεκριμένα, αναπτύχθηκαν τρία 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 προ-

βλήματα κατηγοριοποίησης εικόνων. Θα αναφερθεί η σημασία της αρχιτεκτονικής τους και ο λόγος που χρησιμοποιούνται και για προβλήματα κατηγοριοποίησης ηχητικών συμβάντων, καθώς και η σημασία του μεγέθους και του αριθμού στρωμάτων ενός τέτοιου μοντέλου και πως επηρεάζει την τελική ακρίβεια. Ως δεύτερο μοντέλο, γίνεται η χρήση Conformer [2], το οποίο αποτελεί την τελευταία λέξη σε προβλήματα επεξεργασίας ακολουθιών κειμένου και γλώσσας και χρησιμοποιείται συχνά σε εφαρμογές μετάφρασης. Τέλος, το τρίτο μοντέλο που αναπτύχθηκε αφορά το συνδυασμό των ResNet και Conformer όμοια με την έρευνα που αναφέρεται στο [27], με σκοπό την εκμετάλλευση των θετικών χαρακτηριστικών που μπορεί να προσφέρει το κάθε ένα ξεχωριστά. Όσον αφορά τις επιπλέον τεχνικές που χρησιμοποιήσαμε για την ανάπτυξη των τελικών μοντέλων, υιοθετήθηκαν μέθοδοι επαύξησης δεδομένων χρησιμοποιώντας τυχαία μετατόπιση του φάσματος, ανταλλαγή καναλιών [28], τεχνικές συγκάλυψης από τη μέθοδο του spec-augmentation [29] όπως στα [24], [25], [30], [31], [32] αλλά και πολλές μέθοδοι εξω- και ενδο-συγχωνεύσεων, όπως το SWA που αναφέρεται στα [22], [33].

1.4 Οργάνωση της διπλωματικής

Το υπόλοιπο της διπλωματικής αυτής εργασίας είναι δομημένο ως εξής:

- Κεφάλαιο 2: γίνεται περιγραφή του διαγωνισμού DCASE 2021 Task 3, του βασικού του μοντέλου, της βάσης δεδομένων, καθώς και αναφορά στις μετρικές αξιολόγησης.
- Κεφάλαιο 3: γίνεται παρουσίαση των μοντέλων και των τεχνικών βελτιστοποίησης που αναπτύχθηκαν για την επίλυση του προβλήματος SELD και περιγράφονται οι λόγοι επιλογής τους.
- Κεφάλαιο 4: γίνεται λεπτομερής παρουσίαση των αποτελεσμάτων για όλα τα πειράματα και γίνεται παρουσίαση των συμπερασμάτων πάνω σε αυτά.
- Κεφάλαιο 5: γίνεται μια σύνοψη της διπλωματικής και αναφορά σε μελλοντικές κατευθύνσεις έρευνας.

Κεφάλαιο 2

Διαγωνισμός DCASE 2021 Task 3

Το θέμα της αναγνώρισης και ταξινόμησης ηχητικών συμβάντων έχει προσελκύσει σημαντικά το ενδιαφέρον διάφορων ερευνητών με αποτέλεσμα τη δημιουργία του παγκόσμιου διαγωνισμού DCASE [19] για την περαιτέρω διερεύνησή του. Ο διαγωνισμός καλεί ενδιαφερόμενους ερευνητές από όλο τον κόσμο που ασχολούνται στα πεδία επεξεργασίας σημάτων και μηχανικής μάθησης με σκοπό την ανταλλαγή ιδεών και την έρευνα πάνω στην αντιμετώπιση προβλημάτων ταξινόμησης και αναγνώρισης ήχου σε διάφορες προκλήσεις. Ο διαγωνισμός του DCASE αποτελεί ευκαιρία για την εξέλιξη του κλάδου προσελκύοντας κάθε χρόνο όλο και περισσότερους συμμετέχοντες. Ο πρώτος διαγωνισμός έλαβε χώρα το 2013 και από το 2016 και μετά γίνεται ετησίως. Συγκεκριμένα, ο διαγωνισμός DCASE 2021 προσέφερε 6 κατηγορίες προκλήσεων (tasks) οι οποίες αφορούσαν:

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

Το θέμα της συγκεκριμένης διπλωματικής εργασίας αφορά το Task 3 (SELD) [7], το οποίο περιγράφεται στην επόμενη ενότητα.

2.1 Σύντομη περιγραφή του Task 3

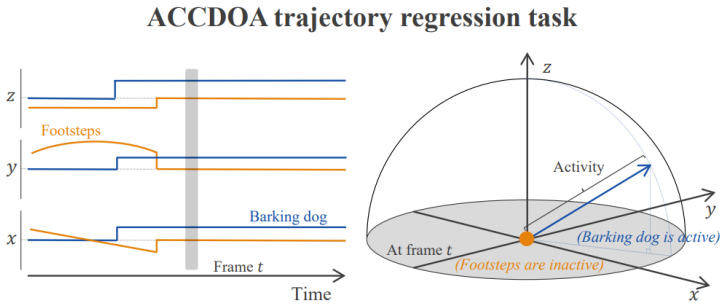
Ένα σύστημα SELD για το Task 3 θα πρέπει να αναγνωρίζει ποιες και πόσες κλάσεις είναι ενεργές σε κάθε χρονικό πλαίσιο του ηχητικού σήματος εισόδου (SED) καθώς και την θέση της πηγής κάθε κλάσης που εντοπίζεται στο χώρο σε σφαιρικές συντεταγμένες, δηλαδή μέσω της αζιμούθιας γωνίας τους και του ύψους τους από το μικρόφωνο εγγραφής (DOA). Για να καθορίσει τον εντοπισμό ενός συμβάντος, το σύστημα πρέπει να παρέχει ως έξοδο μια γραμμή που υποδηλώνει την τοπική του δραστηριότητα στο συγκεκριμένο χρονικό πλαίσιο μαζί με δύο ακόμα τροχιές που προσδιορίζουν την αζιμούθια γωνία και την τιμή ανύψωσής του για το συγκεκριμένο χρονικό πλαίσιο. Κάθε γραμμή χαρακτηρίζεται από κάποιο χρώμα, το οποίο προσδιορίζει σε ποια κλάση ανήκει το ανιχνευμένο συμβάν. Το τελικό γράφημα πρέπει να περιέχει όλα τα ενεργά συμβάντα σε κάθε χρονικό πλαίσιο μαζί με τη θέση τους για ένα δοθέν αρχείο ήχου. Στο διαγωνισμό DCASE 2021 Task 3 υπάρχουν 12 κλάσεις αναγνώρισης:

- Σειρήνα (0)
- Κλάμα μωρού (1)
- Ήχος πρόσκρουσης (2)
- Γάβγισμα σκύλου (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 προσθέτονται, εκτός από τον περιβαλλοντικό θόρυβο και την αντήχηση στα καταγεγραμμένα αρχεία ήχου, και ηχητικές παρεμβολές. Αυτές οι παρεμβολές είναι συμβάντα που δεν ανήκουν σε καμία από τις παραπάνω 12 κλάσεις. Σε κάθε αρχείο εισόδου, τα παρεμβαλλόμενα ηχητικά συμβάντα συντίθενται και τοποθετούνται στα χρονικά πλαίσια εισόδου με τον ίδιο τρόπο που γίνεται και για ένα οποιοδήποτε συμβάν-στόχο. Συγκεκριμένα, ο αριθμός ηχητικών δειγμάτων που ανήκει στις κλάσεις προς αναγνώριση και που χρησιμοποιήθηκαν για τη σύνθεση των αρχείων εισόδου ήταν 500, ενώ τα δείγματα που χρησιμοποιήθηκαν για τις παρεμβολές ήταν 400. Έτσι ο διαγωνισμός DCASE 2021 αποτελεί μεγαλύτερη πρόκληση, καθώς προσεγγίζει περιβάλλοντα υπό πραγματικές συνθήκες. Επίσης ο μέγιστος αριθμός των επικαλυπτόμενων ηχητικών συμβάντων σε σχέση με τον προηγούμενο διαγωνισμό αυξήθηκε από δύο σε τρία, με τη δυνατότητα επικάλυψης και συμβάντων ίδιας κλάσης σε κάθε χρονικό πλαίσιο.

Ωστόσο η πιο σημαντική αλλαγή στο διαγωνισμό DCASE 2021 είναι η σύμπτυξη των κλάδων εξόδου για τα υποπροβλήματα των SED και DOA, χρησιμοποιώντας μια νέα μορφή στόχου-εξόδου των προβλέψεων, την ACCDOA [9]. Σε αντίθεση με προηγούμενες χρονιές, το μοντέλο βάσης του διαγωνισμού DCASE 2021 δεν χρησιμοποιεί έναν ξεχωριστό κλάδο για την έξοδο του SED και έναν για την έξοδο του DOA στην τελική πρόβλεψη, αλλά τα συνδυάζει σε ένα μόνο κλάδο δίνοντας ως έξοδο μόνο ένα διάνυσμα σε μορφή ACCDOA που δίνει την πρόβλεψη όλων των κλάσεων για κάθε χρονικό πλαίσιο της εγγραφής εισόδου. Συγκεκριμένα, η μορφή ACCDOA ομαδοποιεί την δραστηριότητα ενός ηχητικού συμβάντος για κάποιο χρονικό πλαίσιο με τις συντεταγμένες του διανύσματος DOA του. Η πιθανότητα δραστηριότητας μίας κλάσης, κατά πόσο είναι δηλαδή πιθανό να είναι ενεργή ή όχι σε ένα χρονικό πλαίσιο, καθορίζεται από το μήκος του διανύσματος DOA της για κάποιο χρονικό πλαίσιο. Στο σύστημα βάσης του διαγωνισμού, αν το μήκος αυτό ξεπερνάει μια τιμή ορίου, τότε το συμβάν θεωρείται ενεργό. Για κάθε ενεργό συμβάν σε μορφή ACCDOA καθορίζεται επίσης η κατεύθυνσή του με βάση την κατεύθυνση του διανύσματος DOA του.

Η ομαδοποίηση της εξόδου που εμφανίζεται στο ACCDOA, μας δίνει τη δυνατότητα αντιμετώπισης του SELD προβλήματος ως ένα και μόνο πρόβλημα, σε αντίθεση με τις μορφές αναπαράστασης με δύο κλάδους. Με αυτόν τον τρόπο, μειώνονται οι παράμετροι που χρησιμοποιεί το δίκτυο και έτσι μειώνεται και το συνολικό μέγεθός του. Ένα ακόμα πλεονέκτημα σε σύγκριση με τη μέθοδο των δύο κλάδων, είναι πως με τη μορφή ACCDOA γλιτώνουμε το δίκτυο από το πρόβλημα πολλαπλών στόχων για την εύρεση ικανοποιητικών λύσεων SED



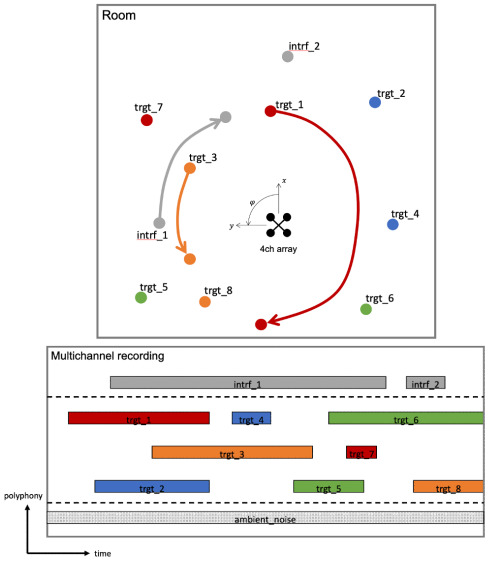
Σχήμα 2.1: Αναπαράσταση ACCDOA για 2 κλάσεις (γάβγισμα και βήματα) για ένα χρονικό πλαίσιο T, σκιαγραφημένο με γκρι πλαίσιο (αριστερά). Δεξιά φαίνεται πώς κάθε κλάση αναπαρίσταται με ένα διάνυσμα στο χώρο. Η κλάση «βήματα» δεν είναι ενεργή στο χρονικό πλαίσιο T οπότε το μέτρο του διανύσματος είναι μηδέν. Αντιθέτως, το «γάβγισμα» είναι ενεργό, και η κατεύθυνσή του φαίνεται με την κατεύθυνση του διανύσματος με βάση τις Καρτεσιανές συντεταγμένες DOA (Εικόνα από [9]).

και DOA. Με τη μορφή ACCDOA δεν απαιτείται η επίλυση αυτού του προβλήματος πολλαπλών στόχων και έτσι μειώνεται η πολυπλοκότητα του δικτύου.

Ένα παράδειγμα για πρόβλεψη δύο κλάσεων με διάνυσμα εξόδου ACCDOA παρουσιάζεται στο Σχήμα 2.1. Η ομαδοποιημένη αυτή μορφή εξόδου των SED και DOA έχει αποδειχθεί μέσω πειραμάτων [9] ότι δίνει καλύτερες προβλέψεις, ειδικά για τις μετρικές LE_{CD} και F_{20° (που περιγράφονται στην Ενότητα 2.4). Για αυτό το λόγο, η μορφή ACCDOA υιοθετήθηκε και στη δικιά μας προσέγγιση.

2.2 Βάση δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση και την αξιολόγηση του μοντέλου προέρχονται από τη βάση δεδομένων TAU-NIGENS Spatial Sound Events 2021 [34], η οποία αποτελείται από πολλαπλές χωρικές ηχογραφήσεις σκηνών και συμβάντων σε αυτές. Κάθε συμβάν εμφανίζεται με διαφορετική κατεύθυνση και απόσταση από το σημείο εγγραφής. Πιθανές τιμές της αζιμούθιας γωνίας είναι στο διάστημα $\phi \in (-180^\circ, 180^\circ]$ και ανύψωσης στο διάστημα $\theta \in [-45^\circ, 45^\circ]$. Η τοποθέτηση στο χώρο κάθε ηχητικού συμβάντος σε κάθε ηχητική σκηνή έγινε με χρήση των χωρικών κρουστικών αποκρίσεών τους (IR) και της μετέπειτα σύνθεσής τους για τα τελικά αρχεία. Κάθε IR ηχογραφήθηκε με τη βοήθεια ενός σφαιρικού μικροφώνου Eigenmike σε 13 διαφορετικές τοποθεσίες εσωτερικού χώρου (δωμάτια) στο Πανεπιστήμιο Tampere της Φινλανδίας. Κάθε ηχητικό αρχείο αντιστοιχεί σε ένα



Σχήμα 2.2: Γραφική απεικόνιση μιας συνθετικής ηχογράφησης. Τα χρωματισμένα αντικείμενα υποδεικνύουν τις κλάσεις-στόχους, τα γκρι τον θόρυβο και τα παρεμβαλλόμενα συμβάντα, και τα βέλη τα μη στατικά συμβάντα (Εικόνα από [7]).

από τα 13 δωμάτια και αποτελεί σύνθεση πολλαπλών IR (SRIR). Η διαδικασία συλλογής και σύνθεσης των SRIR περιγράφεται στο [1]. 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 τα ίδια δωμάτια, συνολικής διάρκειας 30 λεπτών της ώρας. Το μέγεθος του θορύβου επιλέχθηκε να ηχογραφηθεί με ομοιόμορφο SNR στο διάστημα 30-60db, δηλαδή περιβάλλον με αρκετό θόρυβο προς αθόρυβο περιβάλλον. Κάθε ηχογράφηση περιέχει με ίση πιθανότητα στατικά και μη στατικά συμβάντα. Τα μη στατικά συμβάντα, συντίθενται με γωνιακή ταχύτητα $10°$/sec, $20°$/sec ή $40°$/sec ανάλογα με το αν η πηγή κινείται αργά, με μέτρια ταχύτητα ή γρήγορα αντίστοιχα.

Η βάση δεδομένων αποτελείται από 800 ηχογραφημένα αρχεία διαρκείας ενός λεπτού το καθένα, από τα οποία 600 αποτελούν τη βάση ανάπτυξης (development dataset) στην οποία θα εκπαιδευτεί το μοντέλο και τα υπόλοιπα 200 τη βάση αξιολόγησης (evaluation dataset). Οι ηχογραφήσεις της βάσης ανάπτυξης είναι καταγραφές που έγιναν στα 11 από τα 13 συνολικά δωμάτια και διαμοιράζονται σε 6 σύνολα δεδομένων (splits) των 100 ηχογραφήσεων το κάθε ένα. Οι ηχογραφήσεις των υπόλοιπων 2 δωματίων (200 συνολικά) ανήκουν στη βάση αξιολόγησης. Πιο συγκεκριμένα από τα 6 σύνολα της βάσης ανάπτυξης, τα 1-4 χρησιμοποιούνται για εκπαίδευση (training), το 5 για επαλήθευση (validation) και το 6 για έλεγχο (testing). Κάθε ηχογράφηση δίνεται μέσω δύο μορφών καταγραφής χώρου 4 καναλιών η

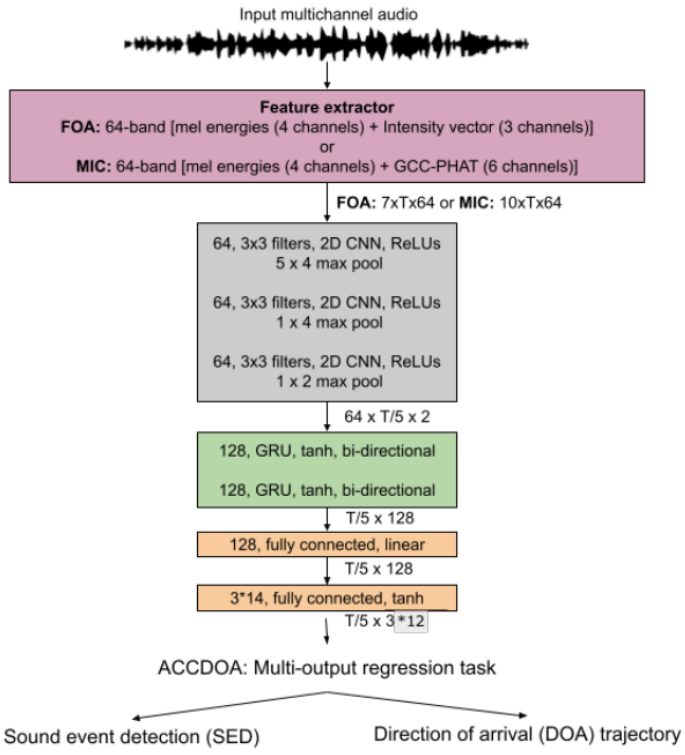
κάθε μία. Αυτές οι μορφές είναι το τετράεδρο δίκτυο μικροφώνων (MIC) και η Αμφισωνική πρώτης τάξης (FOA). Στα πλαίσια της διπλωματικής, επιλέχθηκε να ασχοληθούμε μόνο με τη μορφή MIC. Εκτός από τα 4 κανάλια των ηχογραφήσεων, προστίθενται και επιπλέον κανάλια ανάλογα με τη μορφή εισόδου για την περιγραφή της κατεύθυνσης. Για τη μορφή FOA προστίθενται 3 ακόμα κανάλια για τα διανύσματα ακουστικής εντάσεως (acoustic intensity vectors), ενώ για τη MIC προστίθενται ακόμα 6 για τις γενικευμένες ακολουθίες ετεροσυσχέτισης (GCC-PHAT). Για κάθε αρχείο έγινε υποδειγματοληψία σε συχνότητα 24kHz [1].

2.3 Μοντέλο βάσης

Η δομή και αρχιτεκτονική του μοντέλου βάσης για το διαγωνισμό DCASE 2021 Task 3 [35] (Σχήμα 2.3) είναι παρόμοια με της προηγούμενης χρονιάς. Υιοθετεί ένα CRNN νευρωνικό δίκτυο που βασίζεται στην αρχιτεκτονική SELDnet [8].

Όσον αφορά τη διαδικασία προ-επεξεργασίας των δεδομένων, γίνεται σύμπτυξη κάθε φασματογράμματος σε κλίμακα log-Mel του κάθε αρχείου ήχου μαζί με τα εξαγόμενα χαρακτηριστικά των διανυσμάτων έντασης ή GCC-PHAT, ανάλογα με την επιλεγμένη μορφή δεδομένων. Αρχικά, γίνεται η εξαγωγή των χαρακτηριστικών κατά την οποία το μοντέλο βάσης για κάθε αρχείο ηχογράφησης εισόδου εξάγει πολυκαναλικά φασματογράμματα σε 64 δέσμες σε κλίμακα Mel (Mel-bands) από FFT 1024 σημείων, χρησιμοποιώντας παράθυρο Hamming διάρκειας 40ms με 20ms μήκος άλματος (50% επικάλυψη) και συχνότητα δειγματοληψίας 24kHz. Σε πειράματα που δοκιμάστηκαν αργότερα, αποδείχθηκε ότι η αλλαγή της συχνότητας σε 16kHz και των δεσμών Mel σε 128 δεν βελτίωσε τα αποτελέσματα, οπότε υιοθετήσαμε τις ίδιες παραμέτρους εξαγωγής των ακουστικών χαρακτηριστικών με το μοντέλο βάσης και στα δικά μας πειράματα. Αυτή η διαδικασία δίνει μια ακολουθία από διανύσματα χαρακτηριστικών (έστω T στον αριθμό, ένα για κάθε χρονικό πλαίσιο) για κάθε αρχείο εισόδου στο δίκτυο.

Μετά την εξαγωγή των χαρακτηριστικών, γίνεται η είσοδός τους στο νευρωνικό δίκτυο από το οποίο και θα προκύψουν οι προβλέψεις της ταξινόμησης μέσω διανυσμάτων ACCDOA. Σε πρώτο στάδιο, επεξεργάζονται από ένα δίκτυο CNN τριών στρωμάτων. Ακολουθεί συνάρτηση ενεργοποίησης ReLU και χρονική υποδειγματοληψία μέσω μέγιστης ομαδοποίησης (max-pooling), όπου η ακολουθία εισόδου T διανυσμάτων μετατρέπεται σε ακολουθία με $T/5$ πλαίσια. Στο δεύτερο στάδιο του δικτύου, χρησιμοποιείται RNN δίκτυο, συγκεκρι-



Σχήμα 2.3: Δομή του μοντέλου βάσης του διαγωνισμού DCASE 2021 Task 3, βασισμένη στο SELDnet (Εικόνα από [1]).

μένα Bi-GRU, το οποίο χρησιμεύει στην εκμάθηση των χρονικών συσχετίσεων της εισόδου. Τέλος, τα δεδομένα περνάνε μέσα από δύο πλήρως διασυνδεδεμένα στρώματα και προκύπτει ως τελική έξοδος ένα διάνυσμα ACCDOA διαστάσεων $T/5 \times 3 \times 12$, όπου κάθε μία από τις 12 κλάσεις-στόχους αντιπροσωπεύεται από 3 διανύσματα που αντιστοιχούν στις συντεταγμένες τους x, y, z στο χώρο για κάθε ένα από τα $T/5$ χρονικά πλαίσια. Η τελική έξοδος αναγνωρίζει ένα γεγονός ως ενεργό και επιλέγεται η υπολογισμένη του DOA, εφόσον το μήκος του διανύσματος δραστηριότητας για τη συγκεκριμένη κλάση σε κάποιο από τα $T/5$ χρονικά πλαίσια ξεπερνάει ένα όριο τιμής 0.5.

Ωστόσο, το παρεχόμενο μοντέλο βάσης δεν είναι σε θέση να ταξινομεί επικαλυπτόμενα ηχητικά συμβάντα της ίδιας κλάσης και αδυνατεί στη σωστή εύρεση επικάλυψης τριών γεγονότων, αν και αυτές οι περιπτώσεις μπορούν να προκύψουν καθώς περιλαμβάνονται στη βάση δεδομένων. Στη συγκεκριμένη διπλωματική εργασία προσπαθήσαμε να αντιμετωπίσουμε αυτά τα προβλήματα με την αλλαγή του CNN με άλλες αρχιτεκτονικές. Ειδικότερα, η αλλαγή με ResNet-Conformer, τα πειράματα του οποίου περιγράφονται στην Ενότητα 4.4, κατάφερε να ανιχνεύσει τρία επικαλυπτόμενα γεγονότα ενώ έφτασε πολύ κοντά στην στιγ-

μιαία αναγνώριση επικαλυπτόμενου συμβάντος ίδιας κλάσης.

2.4 Μετρικές αξιολόγησης συστήματος

Για την αξιολόγηση της ακρίβειας κάθε μοντέλου χρησιμοποιούνται 4 μετρικές, από τις οποίες 2 επικεντρώνονται στην ακρίβεια του SED και οι υπόλοιπες 2 στην ακρίβεια χωρικής τοποθέτησης. Συγκεκριμένα, οι μετρικές που στοχεύουν στην αξιολόγηση του SED, επίσης αναφερόμενες και ως μετρικές με επίγνωση στην τοποθεσία (location-aware), πλέον εξαρτώνται από την τοποθεσία των συμβάντων. Οι μετρικές αυτές είναι το σφάλμα απόκλισης ER_{20° (error rate) και το F-σκορ F_{20° (F-score). Οι προβλέψεις που προκύπτουν από το μοντέλο πρέπει να είναι κάτω από ένα όριο απόκλισης 20° για να ληφθούν υπόψιν και να συνυπολογιστούν στο τελικό αποτέλεσμα αξιολόγησης.

Οι υπόλοιπες 2 μετρικές που αφορούν το πρόβλημα του DOA εξαρτώνται από τις εντοπισμένες κλάσεις, δηλαδή κάθε μετρική υπολογίζεται για κάθε κλάση και όχι για όλη την έξοδο συνολικά. Αυτές οι μετρικές είναι το σφάλμα τοποθέτησης LE_{CD} (location error), το οποίο εκφράζει τη μέση γωνιακή απόσταση μεταξύ των προβλέψεων και αναφορών στην ίδια κλάση, και η δεύτερη μετρική είναι η ανάκληση τοποθέτησης LR_{CD} (location recall), η οποία εκφράζει το ποσοστό των ορθώς υπολογισμένων τοποθεσιών που αναγνωρίσθηκε για μία κλάση σε σχέση με το συνολικό αριθμό των εντοπισμένων εμφανίσεών της. Το επίθεμα CD σημαίνει ότι η τιμή τους εξαρτάται από κάθε μία κλάση ξεχωριστά (class-dependent). Περισσότερα για τον τρόπο υπολογισμού τους περιγράφονται στους επίσημους κανονισμούς του διαγωνισμού [36]. Επίσης σημειώνεται πως κάθε μετρική υπολογίζεται σε μη-επικαλυπτόμενα χρονικά πλαίσια διαρκείας ενός δευτερολέπτου.

Μια ακόμα μετρική που συνδυάζει τις τιμές όλων των παραπάνω είναι η $SELD_{score}$, η οποία υπολογίζεται ως:

$$SELD_{score} = \frac{(SED_{score} + DOA_{score})}{2}$$

όπου:

$$SED_{score} = \frac{(ER_{20^\circ} + (1 - F_{20^\circ}))}{2}$$

$$DOA_{score} = \frac{(LE_{CD}/180 + (1 - LR_{CD}))}{2}$$

Γενικά ένα επιτυχημένο σύστημα προσπαθεί να μειώσει όσο περισσότερο γίνεται τις τιμές των ER_{20°, LE_{CD} και $SELD_{score}$ και να αυξήσει αυτές των F_{20° και LR_{CD}.

Κεφάλαιο 3

Προσέγγιση του προβλήματος

Μελετώντας τις τεχνικές αναφορές για τα συστήματα που αναπτύχθηκαν στο διαγωνισμό DCASE 2021 Task 3, παρατηρείται πως εκείνα που βρίσκονται πιο ψηλά στην τελική κατάταξη έχουν αναπτύξει διαφορετικές τεχνικές για την εξαγωγή των χαρακτηριστικών στο στάδιο προ-επεξεργασίας των δεδομένων [24] ή έχουν υιοθετήσει διαφορετικές μορφές εξόδου πέρα του ACCDOA [30]. Ωστόσο, η συγκεκριμένη διπλωματική επικεντρώνεται μόνο στο κομμάτι των νευρωνικών δικτύων και των τεχνικών μετα-επεξεργασίας των προβλέψεων για την ανάπτυξη των συστημάτων.

Από μία άποψη, το πρόβλημα της αναγνώρισης ήχου παρομοιάζει με το πρόβλημα της αναγνώρισης εικόνων, εφόσον υπάρχει σωστή προ-επεξεργασία των δεδομένων. Έτσι, μπορεί να χρησιμοποιεί και δίκτυα που συχνά συναντώνται στα προβλήματα αναγνώρισης εικόνων, όπως τα ResNets. Επιπλέον, ο ήχος είναι ένα σήμα με εξάρτηση της πληροφορίας του στο χρόνο, οπότε όμοια μπορεί να παρομοιαστεί και με προβλήματα πρόβλεψης ακολουθιών που κάνουν χρήση δικτύων γλωσσικής μοντελοποίησης, όπως οι Conformers και Transformers. Βασιζόμενοι σε αυτή τη λογική, ως πρώτο μοντέλο επιλέχθηκε να χρησιμοποιηθεί η αρχιτεκτονική των ResNets, η οποία εφαρμόζεται και στο σύστημα που βρίσκεται δεύτερο στην κατάταξη του διαγωνισμού [24]. Τέλος, λόγω του μεγάλου ενδιαφέροντος και της δημοτικότητας στα περισσότερα συστήματα για αρχιτεκτονικές που εφαρμόζουν την τεχνική της αυτο-προσοχής, επιλέχθηκε να μελετηθεί και η χρήση ενός δεύτερου μοντέλου που βασίζεται στο μοντέλο του Conformer.

Αξίζει να σημειωθεί πως σε όλα τα μοντέλα επιλέχθηκε η αντιμετώπιση του προβλήματος SELD με μικτή εκπαίδευση και πρόβλεψη με μορφή ACCDOA, όμοια με το σύστημα βάσης. Έτσι δεν υπάρχει διαχωρισμός του δικτύου σε δύο διαφορετικά συστήματα για την

επίλυση του SED και DOA ξεχωριστά, αλλά κάθε αρχιτεκτονική που εφαρμόστηκε αφορά τη μικτή πρόβλεψη και των δύο. Για περαιτέρω βελτιστοποίηση των αποτελεσμάτων, αξιοποιήθηκαν τεχνικές επαύξησης δεδομένων καθώς και σύμπτυξης των μοντέλων. Αξίζει να σημειωθεί ότι κάθε μοντέλο αναπτύχθηκε με χρήση των Keras και Tensorflow API [37], [38].

3.1 Μοντέλα

3.1.1 Προ-επεξεργασία και υπερ-παράμετροι

Τα δεδομένα υπέστησαν όμοια προ-επεξεργασία με το μοντέλο βάσης πριν εισαχθούν στο δίκτυο. Συγκεκριμένα, για την εξαγωγή των χαρακτηριστικών εισόδου σε επίπεδο μεθόδων επεξεργασίας σήματος, έγινε εξαγωγή των φασματογραμμάτων log-Mel κάθε αρχείου ήχου σε 64 δέσμες Mel με συχνότητα δειγματοληψίας στα 24kHz και STFT σε παράθυρο Hamming μήκους 40ms και άλματος 20ms. Επίσης, η συνάρτηση σφάλματος που επιλέχθηκε να χρησιμοποιηθεί στο τελευταίο στρώμα για την τελική πρόβλεψη κατά την εκπαίδευση του δικτύου είναι αυτή του σφάλματος ελαχίστων τετραγώνων, καθώς αποδείχθηκε ότι δίνει καλύτερα αποτελέσματα από αυτή της δυαδικής δι-εντροπίας. Ο βελτιστοποιητής του μοντέλου επίσης επιλέχθηκε να είναι ο αλγόριθμος Adam [39] μετά από σχετικές δοκιμές με εναλλακτικούς βελτιστοποιητές, όπως οι αλγόριθμοι SGD [40] και AdaBelief [41]. Τέλος, όμοια με το μοντέλο βάσης αξιοποιήθηκε η πρόωρη διακοπή της εκπαίδευσης σε περίπτωση μη βελτίωσης της τιμής του $SELD_{score}$ μέσα σε κάποιο αριθμό επαναλήψεων [8]. Επιλέχθηκε ο αριθμός επαναλήψεων να είναι 15.

3.1.2 Πρώτη προσέγγιση: ResNets

Μοντέλα CNN ή CRNN χρησιμοποιούνται στην πλειοψηφία των συστημάτων αναγνώρισης ήχου λόγω της ικανότητας που έχει η πράξη της συνέλιξης να αναγνωρίζει τις συσχετίσεις των τοπικών εξαρτήσεων μεταξύ των δεδομένων εισόδου. Καθώς οι αρχιτεκτονικές εξελίσσονται, περισσότερα στρώματα προστίθενται στις παραπάνω αρχιτεκτονικές για να αυξηθεί η ακρίβεια της κατηγοριοποίησης, ειδικά σε επικαλυπτόμενα ηχητικά συμβάντα όπου η είσοδος χρειάζεται πιο περίπλοκους υπολογισμούς. Όμως όσο το δίκτυο αυξάνει σε βάθος, παρουσιάζεται το αποκαλούμενο πρόβλημα της εξαφανιζόμενης κλίσης (vanishing gradient) [42].

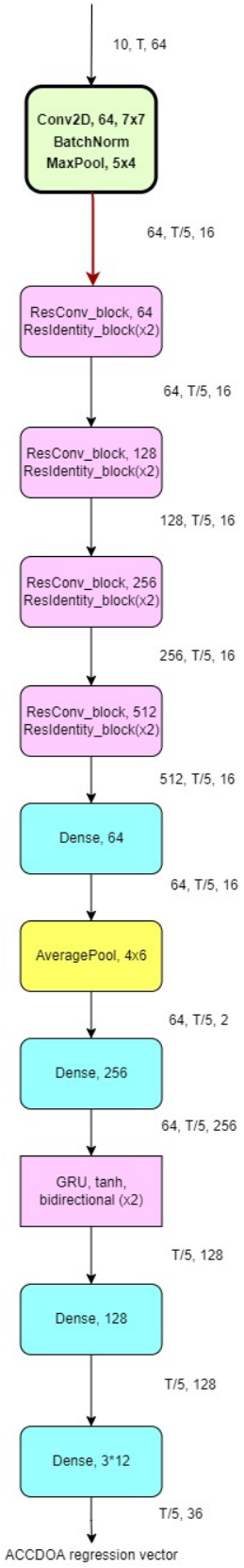
Μια λύση για αυτό το πρόβλημα βρέθηκε με την εισαγωγή παραλειπομένων (skip) συνδέσεων στο δίκτυο, δηλαδή του συνυπολογισμού της αρχικής εισόδου και κατά τον τελικό υπολογισμό της εξόδου σε ένα ολικό άθροισμα. Χρήση αυτής της προσέγγισης κάνουν τα ResNets [43]. Η πρωτοτυπία τους είναι στην εισαγωγή των υπολειμματικών μπλοκ (residual blocks), τα οποία εμπεριέχουν στη δομή τους παραλειπόμενες συνδέσεις. Πιο συγκεκριμένα, τα υπολειμματικά μπλοκ αποτελούνται από ομάδες ταυτοτικών και συνελικτικών μπλοκ στη σειρά. Κάθε ομάδα αποτελείται αρχικά από ένα συνελικτικό μπλοκ και από κάποιον αριθμό ταυτοτικών μπλοκ. Σε ένα συνελικτικό μπλοκ η παραλειπόμενη είσοδος συνελίσσεται μαζί με την είσοδο πριν προστεθούν μαζί για την τελική έξοδο, ενώ σε ένα ταυτοτικό η παραλειπόμενη σύνδεση προστίθεται στην αρχική είσοδο χωρίς να περάσει από συνέλιξη. Ο αριθμός ομάδων μπλοκ εξαρτάται από το βάθος του δικτύου, το οποίο φαίνεται από τον αριθμό στο όνομα του ResNet. Για παράδειγμα, ένα ResNet-34 αποτελείται από 34 στρώματα και πιο συγκεκριμένα από ένα συνελικτικό στρώμα 7×7, 4 ομάδες υπολειμματικών μπλοκ (6, 8, 12 και 6 αντίστοιχα σε κάθε ομάδα) και τέλος ένα στρώμα μέσης ομαδοποίησης (average pooling) για την υποδειγματοληψία των δεδομένων.

Η χρήση των μοντέλων ResNet έχει παρατηρηθεί ότι ξεπερνά σε επιδόσεις αρχιτεκτονικές που κάνουν χρήση απλών CRNN, καθώς τα ResNets με τη χρήση των παραλειπομένων συνδέσεων διατηρούν την πληροφορία εισόδου που μπορεί να χαθεί σε βαθιά δίκτυα με το πρόβλημα της εξαφανιζόμενης κλίσης. Αυτή η βασική ιδέα των παραλειπομένων συνδέσεων έχει εμπνεύσει αρκετές επιτυχημένες αρχιτεκτονικές που χρησιμοποιούνται σε κλάδους της όρασης υπολογιστών και αυτόματης αναγνώρισης φωνής, όπως το DenseNet [44].

Ως πρώτη προσέγγιση στο πρόβλημα του Task 3, θεωρήσαμε αναγκαίο για να γίνει καλύτερη εξαγωγή πληροφορίας να προστεθούν επιπλέον στρώματα CNN στο πρώτο στάδιο του δικτύου. Όμως, λόγω του προβλήματος εξαφανιζόμενης κλίσης αποφασίστηκε η προσέγγιση του προβλήματος με ResNet. Το πρώτο μοντέλο που χρησιμοποιήθηκε βασίστηκε στην αρχιτεκτονική του ResNet18 και του ResNet34.

Αρχιτεκτονική

Η δομή του μοντέλου ResNet φαίνεται στο Σχήμα 3.1. Αρχικά, τα δεδομένα διαστάσεων (B, C, T, F) περνάνε από ένα στρώμα CNN για την εξαγωγή των χαρακτηριστικών τους και στη συνέχεια εφαρμόζεται σε αυτά υποδειγματοληψία για να έρθουν σε μορφή $(B, 64, T/5, F/4)$ και να επεξεργαστούν περαιτέρω από το ResNet, όπου B το μέγεθος ομά-



Σχήμα 3.1: Η δομή του πρώτου SELD μοντέλου που χρησιμοποιήθηκε με ResNet18.

δων δεδομένων που εισέρχονται για εκπαίδευση (batch size), C ο αριθμός καναλιών, T ο αριθμός χρονικών πλαισίων και F οι δέσμες Mel. Το ResNet βασίστηκε στο ResNet18 και αποτελείται από 4 ομάδες υπολειμματικών μπλοκ. Κάθε ομάδα περιέχει ένα συνελικτικό μπλοκ και ακολουθείται από δύο ταυτοτικά. Το ResNet ακολουθείται από ένα στρώμα μέσης ομαδοποίησης μεγέθους 4×6 και ακολουθεί πυκνό στρώμα (dense) για να φέρει τα δεδομένα στις διαστάσεις $(B, T/5, 256)$. Στη συνέχεια ακολουθεί το στάδιο εξαγωγής των χρονικών εξαρτήσεων, στο οποίο εφαρμόσαμε παρόμοιο RNN δίκτυο με αυτό του μοντέλου βάσης, δηλαδή δύο επίπεδα Bi-GRU μεγέθους 128 που μετατρέπουν τα δεδομένα στη μορφή $(B, T/5, 128)$. Για την τελική έξοδο, τα δεδομένα περνάνε από δύο γραμμικά πυκνά στρώματα μεγέθους 128 και από ένα τελικό πυκνό στρώμα μεγέθους 36 που φέρνει την έξοδο στις επιθυμητές διαστάσεις $(T/5, 36)$ με εξίσωση ενεργοποίησης $\tanh$. Η πρόβλεψη για την ύπαρξη μίας κλάσης σε ένα από τα $T/5$ πλαίσια επιλέγεται με βάση ένα όριο 0.5 για το μέτρο του μήκους του διανύσματος ACCDOA εξόδου, όπως στο μοντέλο βάσης.

Για να μελετηθεί η επιρροή των διαφόρων αριθμών στρωμάτων σε ένα ResNet δίκτυο, πέρα από το μοντέλο που κάνει χρήση ResNet18, αναπτύχθηκε όμοιο μοντέλο που αντικαθιστά το ResNet18 με ResNet34. Για το ResNet34 εφαρμόζονται 4 ομάδες μπλοκ με 3, 4, 6 και 3 υπολειμματικά μπλοκ αντίστοιχα. Κάθε ομάδα εκτός από την πρώτη χρησιμοποιεί και ένα συνελικτικό μπλοκ με μέγεθος φίλτρων 64, 128, 256 και 512 αντίστοιχα σε κάθε ομάδα. Τα πυκνά στρώματα και η μέση ομαδοποίηση παραμένουν ίδια.

3.1.3 Δεύτερη προσέγγιση: Conformer

Ως δεύτερη προσέγγιση για τη βελτίωση των αποτελεσμάτων επιλέχθηκε να δοκιμαστεί ο Conformer. Ο Conformer είναι συνδυασμός CNN δικτύου με τα χαρακτηριστικά ενός μοντέλου Transformer. Η ιδέα των Transformers παρουσιάσθηκε πρώτη φορά στο [45] όπου περιγράφεται ένα μοντέλο για μετάφραση οποιασδήποτε ακολουθίας εισόδου σε μια επιθυμητή ακολουθία εξόδου μέσω μηχανισμών προσοχής (attention) και υπολειμματικών συνδέσεων (residual connections). Η εισαγωγή των Transformers [46], [47], [48], [49] και των μηχανισμών προσοχής 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 μοντέλα RNN λόγω του ότι είναι ικανά να κατανοούν και να βρίσκουν τις σχέσεις εξαρτήσεων των δεδομένων εισόδου για μεγαλύτερη εμβέλεια και μήκος ακολουθίας. Η επιτυχία

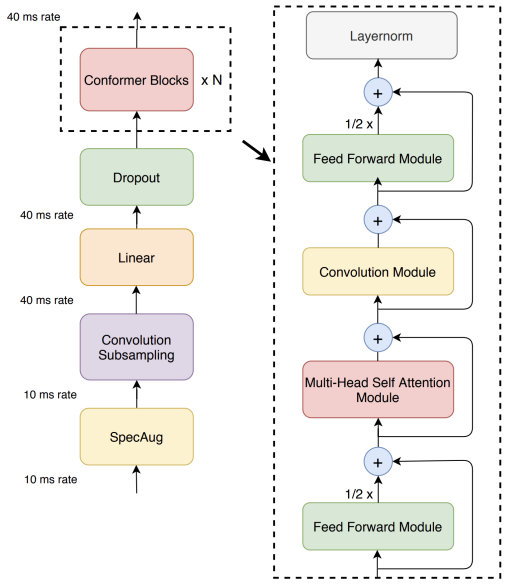
τους αυτή βασίζεται στο μηχανισμό αυτο-προσοχής όπου μπορεί να εφαρμόζει απεριόριστο σε μήκος παράθυρο αναφοράς για την αποκάλυψη σχέσεων στα δεδομένα εισόδου, σε αντίθεση με τα RNN, με αποτέλεσμα να μπορούν να βρουν το συνολικό πλαίσιο συνάφειας των δεδομένων εισόδου και εξόδου και να συγκλίνουν πιο γρήγορα σε κάποια πρόβλεψη.

Ο μηχανισμός προσοχής είναι η διαδικασία αντιστοίχησης 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 διανύσματος ζήτησης με το αντίστοιχο κλειδί του. Οι τιμές της εξόδου μαρτυρούν τη συσχέτιση μεταξύ τους και καθορίζουν ανάλογα πόσο να επικεντρωθεί το δίκτυο σε αυτές τις εξαρτήσεις, ή αλλιώς πόση προσοχή να δώσει. Όμοια, ο μηχανισμός της αυτο-προσοχής ανακαλύπτει το βαθμό εξαρτήσεως μεταξύ των δεδομένων εισόδου και εξάγει ένα διάνυσμα της πιθανότητας συσχέτισής τους.

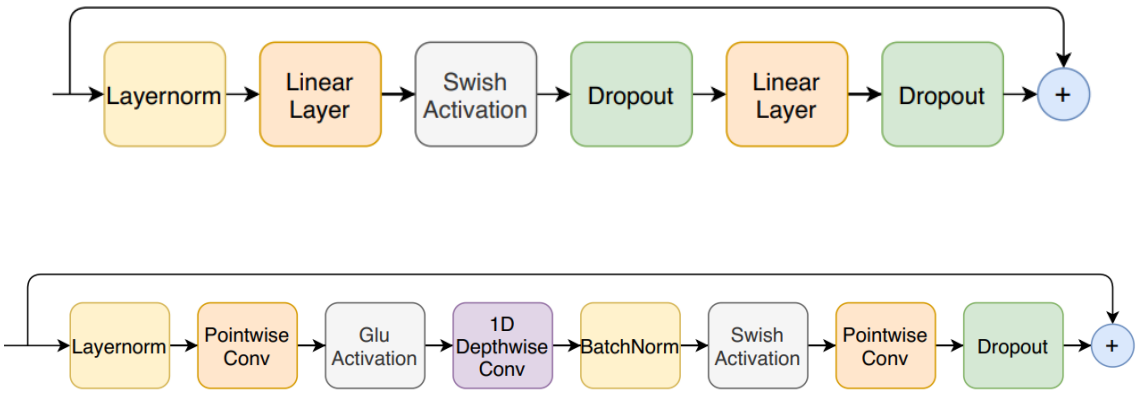
Σε αντίθεση με τους Transformers, τα CNN και RNN είναι καλύτερα στην εύρεση εξαρτήσεων ακολουθίας εισόδου σε τοπικό επίπεδο. Μάλιστα έρευνες και πειραματισμοί με διάφορα μοντέλα, όπως στο [50] όπου προτείνεται ο διαχωρισμός των δεδομένων εισόδου σε δύο ξεχωριστούς κλάδους που κάνουν χρήση Transformers και συνέλιξης ανεξάρτητα, έχουν αποδείξει ότι για καλύτερα αποτελέσματα 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 στο [2] και είναι κυρίως εμπνευσμένο από την πρόταση στο [50] στο πολυκλαδικό του μοντέλο. Αλλά αντίθετα με το μοντέλο που προτείνεται στο [50] όπου η τελική έξοδος των δεδομένων συμπτύσσεται αφού έχει διαχωριστεί και εφαρμοστεί ξεχωριστά σε αυτά αυτο-προσοχή και συνέλιξη, στο Conformer συνδυάζονται οι μηχανισμοί αυτοί μαζί με δίκτυα εμπρόσθιας τροφοδότησης για την εξαγωγή πληροφορίας ταυτόχρονα σε ένα κλάδο.

Αρχιτεκτονική

Η αρχιτεκτονική ενός Conformer μπορεί να ενσωματωθεί στο στάδιο κωδικοποίησης της πληροφορίας εισόδου ή και αποκωδικοποίησης ενός δικτύου. Στο [2] παρουσιάζεται η χρήση του για το στάδιο κωδικοποίησης σε μία αρχιτεκτονική η οποία αποτελείται από ένα αρχικό στρώμα συνέλιξης και από πολλαπλά στοιβαγμένα Conformer μπλοκ (Σχήμα 3.2). Ένα Conformer μπλοκ αποτελείται από τέσσερις διαφορετικές μονάδες αρχιτεκτονικών, οι οποίες είναι παραλλαγές των FNN, MHSA, CNN και ενός στρώματος κανονικοποίησης (normaliza-



Σχήμα 3.2: Απεικόνιση της αρχιτεκτονικής του μοντέλου κωδικοποιητή με Conformer που εμφανίζεται στο [2] (αριστερά) και ενός Conformer μπλοκ που το αποτελεί (δεξιά) (Εικόνα από [2]).



Σχήμα 3.3: (Πάνω) Μονάδα εμπρόσθιας τροφοδότησης. (Κάτω) Μονάδα συνέλιξης (Εικόνες από [2]).

tion layer), το οποίο κανονικοποιεί τις τιμές ενεργοποίησης των προηγούμενων στρωμάτων. Σε κάθε μονάδα αρχιτεκτονικής εφαρμόζεται και η τεχνική της υπολειπόμενης σύνδεσης, η οποία προστίθεται στην έξοδο κάθε μονάδας. Η μονάδα εμπρόσθιας τροφοδότησης (Σχήμα 3.3) ακολουθεί τη δομή παρόμοιων FFN αρχιτεκτονικών σε μοντέλα Transformers, μόνο που στο μπλοκ του Conformer χρησιμοποιεί επιπλέον συνάρτηση ενεργοποίησης swish για την κανονικοποίηση του δικτύου.

Η μονάδα MHSA είναι όμοια με αυτή που εφαρμόζεται σε έναν Transformer με μόνη

διαφορά στο κομμάτι της αυτο-προσοχής, όπου χρησιμοποιεί την εξίσωση της σχετικής κωδικοποίησης θέσης (relative positional encoding) [47], η οποία έχει καλύτερα αποτελέσματα στη γενίκευση του δικτύου για ανεξάρτητο μήκος ακολουθίας εισόδου. Η μονάδα συνέλιξης, εμπνευσμένη από το [50], περιέχει σημειακή συνέλιξη (pointwise convolution) και μονάδα GLU. Η σημειακή συνέλιξη λειτουργεί εφαρμόζοντας τη πράξη της συνέλιξης με φίλτρο 1×1 προς κάθε σημείο της εισόδου. Ακολουθεί μονοδιάστατη συνέλιξη κατά βάθος (depthwise convolution) με συνάρτηση ενεργοποίησης swish. Η συνέλιξη κατά βάθος αποσυνθέτει την είσοδο ως προς κάθε κανάλι της 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) [51], στο οποίο το μοντέλο ενώ μπορεί να είναι ακριβές για δεδομένα στα οποία έχει εκπαιδευτεί, δεν μπορεί να δώσει αντίστοιχα ακριβή αποτελέσματα όταν του παρουσιάζονται νέα δεδομένα που δεν περιέχονται στο σύνολο εκπαίδευσης.

Όσον αφορά την αρχιτεκτονική και δομή για την πρώτη μας δοκιμή ενός μοντέλου Conformer, υιοθετήσαμε στο στάδιο της εξαγωγής των χαρακτηριστικών παρόμοια δομή δικτύου CNN με το μοντέλο βάσης. Πιο συγκεκριμένα, η είσοδος κατά το πρώτο στάδιο συνελίσσεται μέσω δισδιάστατου στρώματος συνέλιξης με φίλτρο μεγέθους 64 και μεγέθους πλαισίου 3×3 και υποδειγματοληπτείται με στρώμα μέγιστης ομαδοποίησης φίλτρου μεγέθους 4 κατά το πεδίο της συχνότητας. Στη συνέχεια, τα δεδομένα περνάνε από ένα δεύτερο στρώμα συνέλιξης μεγέθους 256 και μεγέθους πλαισίου 3×3 και ακολουθεί όμοια υποδειγματοληψία τους μέσω μέγιστης ομαδοποίησης μεγέθους 4 στο πεδίο της συχνότητας. Τα δεδομένα υποδειγματολειπτούνται περαιτέρω με μέγιστη ομαδοποίηση μεγέθους 5 και 4 στο πεδίο του χρόνου και συχνότητας αντίστοιχα. Έτσι η είσοδος μεγέθους $(10, T, 64)$ μετατρέπεται σε $(256, T/5, 2)$, δηλαδή όμοια μορφή με το μοντέλο βάσης στην έξοδο του πρώτου σταδίου.

Το δεύτερο στάδιο αφορά την εύρεση των σχέσεων της εισόδου και χρησιμοποιεί Conformer. Συγκεκριμένα, η είσοδος μετατρέπεται σε δισδιάστατο διάνυσμα διαστάσεων $(T/5, 512)$ και εφαρμόζεται σε αυτή πυκνό στρώμα μεγέθους dim πριν την είσοδο στα στρώματα Conformer. Το μέγεθος dim βρέθηκε μετά από δοκιμές ότι είναι βέλτιστο όταν τίθεται σε τιμή 256. Αφού μετατρέψουμε την είσοδο σε δισδιάστατη μορφή για να μπορεί να περαστεί από τον Conformer, ακολουθούν δύο στοιβαγμένα στρώματα Conformer στη σειρά. Κάθε στρώμα Conformer περιέχει 8 κεφαλές προσοχής στο στάδιο του MHSA, όπου η κάθε μία έχει μέγεθος 64. Οι τιμές αυτές προέκυψαν από δοκιμές και πειραματισμούς με άλλα

νούμερα και βρέθηκαν ότι αυτές είναι βέλτιστες.

Τέλος, το τρίτο στάδιο αφορά την εύρεση των χρονικών συσχετίσεων της εισόδου και μετά από πειράματα αποφασίστηκε να μην αφαιρεθεί το RNN δίκτυο (όμοια με [2], [22]) και να εφαρμοστεί Bi-GRU και πυκνά στρώματα όμοια με το μοντέλο βάσης για την τελική έξοδο. Αυτό έγινε γιατί αν και αρκετές έρευνες έχουν αποδείξει ότι η αντικατάσταση των RNN δικτύων με Conformer για την εύρεση των εξαρτήσεων της εισόδου αποδίδει καλύτερες προβλέψεις, μέσω των δοκιμών μας ένα τέτοιο μοντέλο απαιτεί αρκετά παραπάνω επαναλήψεις εκπαίδευσης σε σχέση με την ενσωμάτωση ενός Bi-GRU για να το ξεπεράσει σε ακρίβεια. Έτσι για να έχουμε πιο δίκαιη σύγκριση, εφαρμόστηκε Bi-GRU και τρέξαμε όλα τα μοντέλα για 40 επαναλήψεις. Επιπλέον, δοκιμάστηκε ένα ακόμα μοντέλο Conformer, το οποίο χρησιμοποιεί στο στάδιο της εξαγωγής διαφορετικό δίκτυο CNN. Συγκεκριμένα, το CNN του μοντέλου βάσης αντικαθίσταται από ένα δίκτυο DenseNet με δύο DenseNet μπλοκ [44]. Σχετικά με το προγραμματιστικό κομμάτι της υλοποίησης του στρώματος Conformer στα μοντέλα μας, επηρεαστήκαμε από την υλοποίηση και έρευνα των [52], [53].

3.1.4 Τρίτη προσέγγιση: ResNet-Conformer

Από τα αποτελέσματα των πειραμάτων αποδείχθηκε ότι κάθε ένα από τα δύο μοντέλα ResNet και Conformer έχουν διαφορετικά πλεονεκτήματα. Τα ResNet κάνουν καλύτερη εξαγωγή των τοπικών σχέσεων που είναι αμετάβλητες σε μετατοπίσεις των δεδομένων εισόδου, σε σύγκριση με ένα απλό CNN που χρησιμοποιείται στο μοντέλο Conformer και όπως είναι αυτό του μοντέλου βάσης. Επίσης ένας Conformer προσφέρει καλύτερη εκμάθηση των τοπικών και ολικών συσχετίσεων εισόδου σε σύγκριση με ένα RNN. Έτσι αποφασίστηκε η δημιουργία ενός μοντέλου το οποίο να συνενώνει τα δύο αυτά για την ταυτόχρονη εκμετάλλευση των πλεονεκτημάτων που προσφέρουν. Η ιδέα για ένα τέτοιο μοντέλο ερευνήθηκε και αναπτύχθηκε στο [28] και αποδείχθηκε ότι ξεπερνούσε σε επίδοση κάθε ένα από τα δύο μοντέλα ξεχωριστά. Στη δικιά μας αρχιτεκτονική, προσθέτουμε πέρα από το κομμάτι των Conformer και ένα δίκτυο RNN για την εξαγωγή των εξαρτήσεων, καθώς από πειράματα αποδείχθηκε ότι συνέκλινε σε καλύτερη πρόβλεψη για 40 επαναλήψεις. Ωστόσο για παραπάνω τρεξίματα, η αφαίρεση του RNN είχε καλύτερα αποτελέσματα, αλλά λόγω της καταμέτρησης της ακρίβειας όλων των μοντέλων σε 40 επαναλήψεις για πιο δίκαιη αξιολόγηση, επιλέχθηκε η πρόσθεση δύο στρωμάτων Bi-GRU στο τελικό δίκτυο.

Αρχιτεκτονική

Ως πρώτο μοντέλο ResNet με Conformer, χρησιμοποιήθηκε η υλοποίηση της δομής ResNet34 από την πρώτη προσέγγιση. Το ResNet34 ακολουθείται από δύο στρώματα Conformer με $dim = 256$, αριθμό κεφαλών 8 και βάθους 64. Ουσιαστικά, αποτελεί παραλλαγή του πρώτου μοντέλου προσέγγισης Conformer, με αντικατάσταση του αρχικού CNN του σταδίου εξαγωγής χαρακτηριστικών από ένα δίκτυο ResNet34. Η έξοδος του Conformer περνάει από ένα δίκτυο Bi-GRU όμοιο με το αρχικό μοντέλο και για την τελική έξοδο εφαρμόζεται πυκνό στρώμα μεγέθους 36 με συνάρτηση ενεργοποίησης $tanh$.

Ως δεύτερη αρχιτεκτονική για τη προσέγγιση ResNet-Conformer, χρησιμοποιήθηκε η ίδια δομή με το μοντέλο ResNet34-Conformer αλλά με την αντικατάσταση του ResNet34 από ένα υπολειμματικό δίκτυο με λιγότερα στρώματα. Πιο συγκεκριμένα, το δίκτυο αυτό όμοια με το ResNet34 απαρτίζεται από 4 ομάδες μπλοκ. Σε αντίθεση όμως με το ResNet34, κάθε ομάδα μπλοκ αποτελείται από ένα συνελικτικό υπολειμματικό στρώμα και ένα ταυτοτικό υπολειμματικό στρώμα μόνο. Η δομή αυτού του ResNet βασίστηκε στο [27].

3.2 Περαιτέρω τεχνικές

3.2.1 Επαύξηση δεδομένων

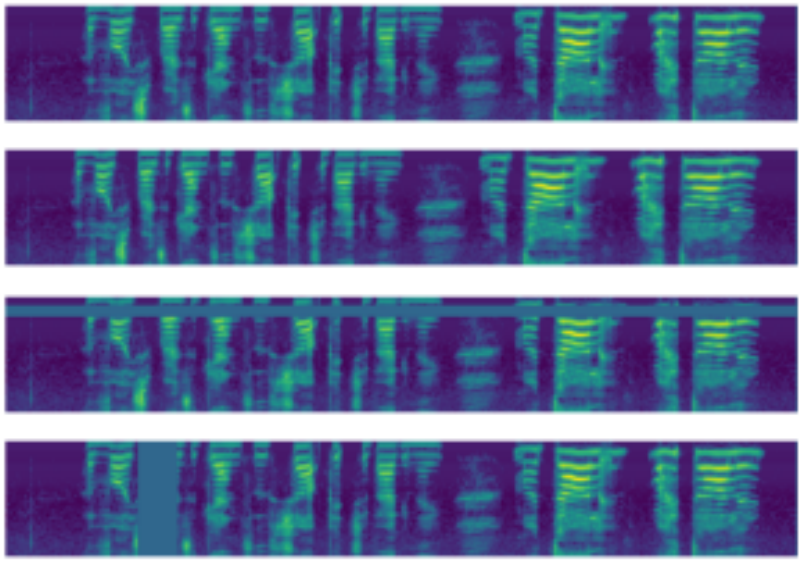
Ένα επιτυχημένο νευρωνικό δίκτυο οφείλει να μπορεί να προσαρμόζεται και σε δεδομένα στα οποία δεν έχει εκπαιδευτεί. Αυτή η ικανότητα των μοντέλων, να μπορούν δηλαδή να δίνουν ακριβή αποτελέσματα και σε νέες εισόδους ονομάζεται γενίκευση (generalization). Αν ένα δίκτυο δεν είναι ικανό να γενικεύει καλά, οδηγείται σε υπερ-μοντελοποίηση [51], [54]. Ένας από τους τρόπους που μπορεί ένα δίκτυο να βελτιωθεί ως προς την ικανότητα γενίκευσής του είναι η εκπαίδευση σε αυξημένο όγκο δεδομένων. Ωστόσο, η βάση δεδομένων που παρέχεται στο διαγωνισμό DCASE 2021 Task 3 για την εκπαίδευση των μοντέλων είναι αρκετά περιορισμένη σε μέγεθος, περιέχοντας μόλις 600 δείγματα των ακουστικών κλάσεων. Επίσης απαγορεύεται από τους κανονισμούς η προσθήκη επιπλέον δεδομένων στην παρεχόμενη βάση από εξωτερικές πηγές. Έτσι, για να αποφύγουμε το πρόβλημα της υπερ-μοντελοποίησης χρησιμοποιώντας μόνο τη δοθείσα βάση δεδομένων, οδηγούμαστε στη μέθοδο της επαύξησης δεδομένων (data augmentation). Η μέθοδος αυτή χρησιμοποιεί διάφορες τεχνικές μετασχηματισμών στα δεδομένα εισόδου, με σκοπό τη δημιουργία

νέων δεδομένων με διαφορετική μορφή από την αρχική. Τα μετασχηματισμένα αυτά νέα δεδομένα προστίθενται στην αρχική βάση και χρησιμοποιούνται στη συνολική εκπαίδευση του δικτύου. Οι μετασχηματισμοί μπορούν να εφαρμοστούν είτε στα αρχικά δεδομένα, είτε στα εξαχθέντα χαρακτηριστικά τους μετά το στάδιο της προ-επεξεργασίας. Στα πλαίσια της διπλωματικής αξιοποιήθηκαν και οι δύο τρόποι. Συνολικά έγινε εφαρμογή τριών τεχνικών επαύξησης δεδομένων, συγκεκριμένα: (α) της τυχαίας μετατόπισης του φάσματος του σήματος εισόδου (random shift), (β) της χρήσης μασκών στο πεδίο συχνότητας και χρόνου (time and frequency masking) και (γ) της ανταλλαγής καναλιών ήχου της εισόδου (audio channel swap). Η τελευταία ανήκει και στις τεχνικές επαύξησης δεδομένων χώρου (spatial augmentation).

Η πρώτη μέθοδος μετατοπίζει τυχαία το εισαγόμενο φασματόγραμμα του σήματος κατά ένα τυχαίο αριθμό δεσμών Mel προς τα πάνω ή προς τα κάτω, δηλαδή κατά μήκος του άξονα της συχνότητας.

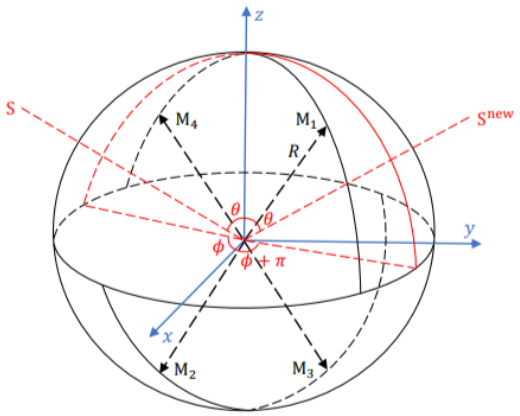
Η δεύτερη μέθοδος βασίζεται στις τεχνικές του SpecAugment [29]. Ουσιαστικά για κάθε φασματόγραμμα Mel επιλέγεται τυχαία μία λωρίδα μηδενισμού στον άξονα του χρόνου ή και της συχνότητας η οποία θα λειτουργήσει ως μάσκα μηδενίζοντας τις τιμές στην περιοχή που θα εφαρμοστεί. Η τεχνική αυτή στο πεδίο της συχνότητας λειτουργεί επιλέγοντας τυχαία συνεχόμενες δέσμες Mel στο διάστημα $[f_0, f_0 + \Delta f)$ στις οποίες θα εφαρμοστεί η μάσκα μηδενισμού, όπου f_0 ένας τυχαίος αριθμός δέσμης Mel. Όμοια για το πεδίο του χρόνου, το διάστημα στο οποίο εφαρμόζεται η μάσκα είναι στο $[t_0, t_0 + \Delta t)$, επιλέγοντας τυχαία έναν αριθμό πλαισίων από μια ομοιόμορφη κατανομή από το t_0 έως μια δοσμένη μέγιστη τιμή πλαισίων. Ένα παράδειγμα εφαρμογής της μεθοδολογίας αυτής εμφανίζεται στο Σχήμα 3.4, όπου οι περιοχές μασκών που μηδενίζονται φαίνονται στο φασματόγραμμα με μια πιο ανοιχτή μπλε απόχρωση.

Οι παραπάνω τεχνικές είναι αποδοτικές όσον αφορά τη βελτίωση του SED, αλλά δεν συμβάλουν αρκετά στην καλύτερη εκτίμηση του DOA. Οι τεχνικές που στοχεύουν στην επαύξηση των δεδομένων που αντιπροσωπεύουν τις DOA ανήκουν στην κατηγορία των χωρικών επαυξήσεων (spatial augmentation). Αυτές οι τεχνικές είναι δύο: η πολυ-καναλική προσομοίωση (multi-channel simulation) [28] και η ανταλλαγή καναλιών ήχου (audio channel swapping). Στα πλαίσια της διπλωματικής χρησιμοποιήθηκε η δεύτερη τεχνική για χωρική επαύξηση. Η ανταλλαγή καναλιών ήχου ανταλλάσσει τυχαία τα κανάλια της εισόδου MIC (κανάλια M1, M2, M3, M4) με βάση τους μετασχηματισμούς που αναφέρονται στο



Σχήμα 3.4: Στην πάνω εικόνα εμφανίζεται το αρχικό φασματόγραμμα Mel πριν την εφαρμογή κάποιου μετασχηματισμού. Η τρίτη από πάνω εικόνα δείχνει το φασματόγραμμα αφού εφαρμοστεί σε αυτό μάσκα συχνότητας και η τέταρτη χρονική μάσκα. Η δεύτερη εικόνα αφορά έναν τρίτο μετασχηματισμό SpecAugment που δεν εφαρμόστηκε στα πλαίσια της διπλωματικής, την παραμόρφωση χρόνου (Εικόνα από [29]).

[28]. Για τη μορφή MIC υπάρχει περιορισμένος αριθμός μετασχηματισμών που μπορούν να εφαρμοστούν στα κανάλια για να μη χρειαστεί να αλλοιωθεί η χωρική πληροφορία. Με βάση τους περιορισμούς για τη μέγιστη αζιμούθια γωνία και το πεδίο τιμών ανύψωσης, οι μετασχηματισμοί αυτοί είναι συνολικά 8. Η τεχνική αυτή δεν αλλάζει μόνο τα χαρακτηριστικά της εισόδου, αλλά και τις ετικέτες προβλέψεων ώστε να συμπίπτουν οι διαστάσεις των καναλιών μετά την ανταλλαγή τους. Ένα παράδειγμα ανταλλαγής καναλιών MIC φαίνεται στο Σχήμα 3.5. Στο σχήμα αυτό, με S απεικονίζεται το διάνυσμα DOA πριν κάποιο μετασχηματισμό, το οποίο χαρακτηρίζεται από αζιμούθια γωνία ϕ και ανύψωση θ. Έστω ότι επιλέγεται να εφαρμοστεί ανταλλαγή καναλιών και συγκεκριμένα ο μετασχηματισμός $(\phi \leftarrow \phi + \pi, \theta \leftarrow \theta)$. Σύμφωνα με τον Πίνακα 1 στο [28] για να γίνει ο μετασχηματισμός αυτός, θα γίνει ανταλλαγή του καναλιού M1 με το M4 και του M2 με το M3. Έτσι δημιουργείται ένα νέο διάνυσμα DOA S_{new} με νέες γωνίες $(\phi + \pi, \theta)$. Το νέο αυτό διάνυσμα προστίθεται με το αρχικό S για την εκπαίδευση του δικτύου. Αν και οι συνολικοί μετασχηματισμοί για δεδομένα MIC είναι 8, επιλέχθηκε να αξιοποιηθούν μόνο 2 από αυτούς, καθώς με την εφαρμογή όλων των 8 δεν υπήρξε αρκετά παραπάνω βελτίωση για τις προβλέψεις τοποθεσίας. Συγκεκριμένα οι μετασχηματισμοί που επιλέχθηκαν είναι για γωνίες $(-\phi + \pi/2, \theta)$ και $(-\phi - \pi/2, \theta)$.



Σχήμα 3.5: Εφαρμογή της τεχνικής ανταλλαγής καναλιών για το μετασχηματισμό ($\phi \leftarrow \phi + \pi, \theta \leftarrow \theta$). Φαίνεται η αρχική DOA S πριν το μετασχηματισμό σε σχέση με την τελική S_{new}. (Εικόνα από [28]).

Όσον αφορά την προγραμματιστική υλοποίηση των τεχνικών επαύξησης δεδομένων, βασίστηκαμε στην υλοποίηση του [55].

3.2.2 Συγχώνευση μοντέλων

Από τα αποτελέσματα των πειραμάτων (Κεφάλαιο 4) είναι εμφανές ότι διαφορετικά μοντέλα πλεονεκτούν σε κάποια μετρική αξιολόγησης από κάποια άλλα. Για παράδειγμα, τα ResNet μοντέλα παρουσίασαν αρκετά καλά αποτελέσματα για τις μετρικές SED, ER_{20° και F_{20°, ενώ τα Conformer είχαν καλύτερες τιμές για τις μετρικές που αφορούν το DOA, τις LE_{CD} και LR_{CD}. Επειδή κάθε μοντέλο χαρακτηρίζεται από ξεχωριστές ικανότητες και υπερτερεί σε διαφορετικά σημεία από τα υπόλοιπα, για τη δημιουργία της τελικής εξόδου αποφασίστηκε η αξιοποίηση της μεθόδου συγχώνευσης (ensembling) για τη δημιουργία ενός μοντέλου, το οποίο να μπορεί να αξιοποιεί όλα τα διαφορετικά πλεονεκτήματα των επιμέρους μοντέλων.

Οι μέθοδοι συγχώνευσης χρησιμοποιούνται συχνά σε αλγορίθμους μηχανικής μάθησης [56], [57], καθώς μπορούν τις περισσότερες φορές να ξεπεράσουν τους περιορισμούς επίδοσης σε σύγκριση με την χρήση του κάθε αλγορίθμου ξεχωριστά. Ένα ακόμα πλεονέκτημα της χρήσης συγχωνευμένου μοντέλου είναι πως μειώνει την πιθανότητα της υπερμοντελοποίησης του δικτύου, καθώς δεν βασίζεται μόνο στα βάρη ενός μοντέλου. Οι μέθοδοι συγχώνευσης μπορούν να χωριστούν σε δύο κατηγορίες, σε αυτές που βασίζονται σε τεχνικές ενδο-συγχώνευσης (intra-ensembling) [57] και σε τεχνικές εξω-συγχώνευσης (inter-

ensembling).

Οι πρώτες αφορούν υπολογισμούς και μετατροπές των στοχαστικών μεταβλητών ενός μοντέλου που μπορούν να λάβουν χώρα κατά ή και έπειτα από το στάδιο εκπαίδευσής του. Μια τεχνική ενδο-συγχώνευσης είναι η στοχαστική μέση ανανέωση βαρών (SWA) [58], [59]. Αυτή η μέθοδος εφαρμόζεται κατά την εκπαίδευση του δικτύου και η λειτουργία της είναι να ανανεώνει τα βάρη του δικτύου με το μέσο όρο τους, ανάλογα με τον αριθμό επαναλήψεων που βρίσκεται. Η μέθοδος SWA αξιοποιήθηκε και στα πλαίσια της συγκεκριμένης διπλωματικής. Η διαδικασία ξεκινάει μετά από τη δέκατη επανάληψη και καλείται ανά 2 επαναλήψεις.

Όσον αφορά την κατηγορία των εξω-συγχωνεύσεων υπάρχουν 4 βασικές τεχνικές: Bagging, Stacking, Boosting και Averaging [60]. Εμείς ασχοληθήκαμε με τη συγχώνευση μοντέλων σύμφωνα με την τεχνική του Simple Averaging. Σε αυτήν, κάθε μοντέλο δημιουργεί μία πρόβλεψη για κάθε αρχείο εισόδου και η τελική πρόβλεψη του συγχωνευμένου μοντέλου δημιουργείται με τον υπολογισμό του μέσου όρου όλων των προβλέψεων αυτών ανά μοντέλο.

Κατά τα πειράματά μας δοκιμάσαμε 8 μοντέλα συγχώνευσης, τα οποία συνδυάζουν όλα τα μοντέλα που αναπτύξαμε. Πιο συγκεκριμένα, τα μοντέλα συγχώνευσης αξιοποιούν διάφορους συνδυασμούς μοντέλων CNN, ResNet, Conformer και ResNet-Conformer. Πιο αναλυτικά θα αναπτύξουμε από τι αποτελείται το κάθε μοντέλο συγχώνευσης στο επόμενο κεφάλαιο.

Κεφάλαιο 4

Πειράματα

Στο κεφάλαιο αυτό περιγράφουμε τα διάφορα πειράματα που διεξήχθησαν για κάθε ένα από τα επιλεγμένα μοντέλα και συγκρίνουμε τα αποτελέσματα των μετρικών αξιολόγησής τους για την εύρεση της βέλτιστης αρχιτεκτονικής. Σχολιάζονται και αναδεικνύονται οι διαφορές στα αποτελέσματα των πειραμάτων που κάνουν χρήση διαφορετικών τεχνικών βελτιστοποίησης. Τα πειράματα για κάθε μοντέλο έγιναν με τις ίδιες τιμές παραμέτρων στο στάδιο προ-επεξεργασίας των δεδομένων. Συγκεκριμένα έγινε δειγματοληψία με συχνότητα 24kHz και με παράθυρο Hamming μήκους 40ms με 20ms άλμα. Κάθε μοντέλο εκτός από περιπτώσεις που θα αναφερθούν, εκπαιδεύτηκε με βελτιστοποιητή Adam με ρυθμό εκπαίδευσης 0.001 και για τον υπολογισμό του σφάλματος έγινε χρήση σφάλματος ελαχίστου τετραγώνου. Κάθε μοντέλο εκπαιδεύτηκε στα σύνολα 1-4 της βάσης ανάπτυξης για 40 επαναλήψεις. Το μέγεθος ομάδων δεδομένων εκπαίδευσης (batch size) επιλέχθηκε να είναι 64. Τα αποτελέσματα των πειραμάτων αφορούν τη βάση αξιολόγησης. Τα πειράματα αφορούν:

- Αλλαγή δομής του CNN του σταδίου εξαγωγής χαρακτηριστικών.
- Χρήση ResNet με διαφορετικό αριθμό στρωμάτων.
- Αλλαγή αποκωδικοποιητών του δικτύου
- Αλλαγή του αριθμού δεσμών δεδομένων.
- Χρήση διαφορετικής διάστασης dim για το μηχανισμό αυτο-προσοχής.
- Αλλαγή στη δομή και στον αριθμό στρωμάτων των μοντέλων.
- Χρήση τεχνικών επαύξησης δεδομένων.

- Συγχώνευση μοντέλων.

4.1 Πειράματα με CNN

Αρχικά δοκιμάστηκαν πειράματα πάνω στην αρχική αρχιτεκτονική CNN του μοντέλου βάσης. Δοκιμάσαμε να αλλάξουμε τον βελτιστοποιητή Adam με SGD και με την συνάρτηση σφάλματος MSE με δυαδική δι-εντροπία που συνηθίζεται σε προβλήματα ταξινόμησης. Αλλαγή στον αριθμό ομάδων δεδομένων εκπαίδευσης δεν έδειξε σημαντική αλλαγή στα αποτελέσματα και δεν παρουσιάζεται. Επιπλέον αλλαγές έγιναν στο ρυθμό εκπαίδευσης, όπου δοκιμάστηκε η εκπαίδευση του δικτύου για αρχικό ρυθμό 0.1 και 0.00001, καθώς και αλλαγή στον αποκωδικοποιητή σε LSTM αντί GRU. Επίσης, δοκιμάσαμε τη χρήση 3 παράλληλων κλάδων για την έξοδο, χρησιμοποιώντας 3 πυκνά στρώματα ταυτόχρονα με διαφορετική αρχικοποίηση βαρών. Κάθε στρώμα αντιπροσωπεύει μια από τς 3 συντεταγμένες (x, y, z) ACCDOA της τελικής πρόβλεψης. Ακόμα, δοκιμάστηκε να αλλάξουμε το αρχικό CNN και να χρησιμοποιήσουμε DenseNet υπό προκαθορισμένες συνθήκες. Όλα τα αποτελέσματα των μετρικών για τα διαφορετικά πειράματα πάνω στα δίκτυα CNN φαίνονται στον Πίνακα 4.1.

Στον πίνακα αυτόν με lr συμβολίζουμε τον αρχικό ρυθμό εκπαίδευσης και με DA συμβολίζουμε τη μέθοδο επαύξησης δεδομένων. Αν DA=1 έχει εφαρμοστεί τυχαία μετατόπιση φάσματος σήματος εισόδου, για DA=2 έχει γίνει χρήση μασκών στο πεδίο συχνότητας και χρόνου και για DA=3 έγινε ανταλλαγή καναλιών ήχου της εισόδου.

Από τα αποτελέσματα φαίνεται πως όταν εφαρμόζεται κάποια μέθοδος επαύξησης δεδομένων, αυτό έχει αρνητική επιρροή στην επίδοση του αρχικού δικτύου. Το μόνο πλεονέκτημα που προσέφερε η εφαρμογή μεθόδων επαύξησης δεδομένων είναι η βελτίωση του LR_{CD} με τυχαία μετατόπιση φάσματος του σήματος εισόδου, αλλά είχε χειρότερο ολικό $SELD_{score}$. Επίσης παρατηρήθηκε ότι η χρήση GRU είναι καλύτερη του LSTM. Όσον αφορά το ρυθμό εκπαίδευσης αρχικά δοκιμάσαμε να θέσουμε την αρχική του τιμή σε 0.1, ωστόσο το δίκτυο δεν συνέκλινε σε κάποιο αποτέλεσμα και έδινε υπερβολικά μεγάλο σφάλμα. Αυτό σημαίνει ότι ο ρυθμός εκπαίδευσης ήταν αρκετά μεγάλος, και οδήγησε το σύστημα σε απότομη αύξηση των κλίσεων (exploding gradient), καθιστώντας το ανίκανο να συγκλίνει σε κάποιο αποτέλεσμα.

Μια μεγάλη βελτίωση στα αποτελέσματα είχαμε με την αντικατάσταση του αρχικού CNN με ένα DenseNet. Το DenseNet που χρησιμοποιήσαμε αποτελείται από 2 στρώματα

Πίνακας 4.1: Πειράματα με CNN

Μοντέλο	$ER_{20^\circ}(^\circ)$	$F_{20^\circ}(^\circ)$	LE_{CD}	LR_{CD}	$SELD_{score}$
Baseline, 50 epochs	0.74	24.7	30.9	38.2	0.57
Baseline	0.75	22.4	33.8	39.9	0.58
Baseline, lr=0.01	-	-	-	-	-
Baseline, lr=0.00001	0.74	22.3	38.5	33.3	0.60
Baseline, LSTM	0.78	22.4	36.4	35.3	0.60
Baseline, SGD	0.85	16.2	56.2	21.7	0.75
Baseline, SGD, Binary Cross-Entropy	0.87	15.6	55.0	20.2	0.76
Baseline, DA=1	0.8	15.5	39.8	30.0	0.64
Baseline, DA=2	0.84	21.4	36.8	42.5	0.60
Baseline, DA=1+2	0.82	16.7	40.0	27.3	0.63
Baseline, parallel	0.89	11.4	36.1	14.8	0.71
DenseNet	**0.73**	**28.0**	**27.6**	**41.7**	**0.55**

με βάθος 3 για κάθε DenseNet μπλοκ. Το δίκτυο αυτό όπως φαίνεται από τα αποτελέσματα ξεπερνάει σε κάθε μετρική το αρχικό μοντέλου βάσης.

4.2 Πειράματα με ResNet

Ως πρώτο μοντέλο αποφασίσαμε να χρησιμοποιήσουμε ένα δίκτυο ResNet18, το οποίο συχνά εφαρμόζεται σε προβλήματα ταξινόμησης εικόνων. Ωστόσο, για να πειραματιστούμε με την επιρροή του αριθμού στρωμάτων ενός ResNet, χρησιμοποιήσαμε και ένα ακόμα μοντέλο υπολειμματικών δικτύων με περισσότερα στρώματα βάθους, το ResNet34. Τα αποτελέσματα των δοκιμών για τα δύο αυτά μοντέλα παρουσιάζονται στους Πίνακες 4.2, 4.3.

Για κάθε πείραμα αξιοποιήθηκαν περαιτέρω η τεχνική του SWA για κάθε μοντέλο ξεχωριστά, καθώς και ένας ελεγκτής βαθμού εκμάθησης, ο οποίος μειώνει την τιμή του βαθμού εκμάθησης κατά 10% αν η τετραγωνική ρίζα της τιμής σφάλματος απόκλισης της τελευταίας επανάληψης δεν ξεπερνάει ένα όριο τιμής 0.3. Κάθε μοντέλο εκπαιδεύτηκε με βελτιστοποιητή Adam και σφάλμα ελαχίστου τετραγώνου ως συνάρτηση σφάλματος.

Στα δίκτυα ResNet34v2 και ResNet18v2 αλλάξαμε τον αριθμό και τις διαστάσεις κάποιων πυκνών στρωμάτων πριν την τελική έξοδο. Με βάση τα αποτελέσματα, τα ResNet34v2

Πίνακας 4.2: Πειράματα πάνω στο ResNet18

Μοντέλο	ER_{20°	F_{20°	LE_{CD}	LR_{CD}	$SELD_{score}$
Baseline	**0.75**	22.4	33.8	39.9	0.58
ResNet18	0.89	7.9	43.3	13.5	0.73
ResNet18, DA=1	0.86	9.1	44.6	25.3	0.70
ResNet18, DA=2	0.85	9.4	47.2	25.5	0.69
ResNet18, DA=1+2	0.8	13.8	40.9	28.1	0.67
ResNet18v2, DA=1+2	**0.75**	**24.0**	**33.4**	**42.4**	**0.57**

Πίνακας 4.3: Πειράματα πάνω στο ResNet34

Μοντέλο	ER_{20°	F_{20°	LE_{CD}	LR_{CD}	$SELD_{score}$
Baseline	0.75	22.4	33.8	39.9	0.58
ResNet34	0.79	20.1	29.5	28.7	0.63
ResNet34, DA=1	0.77	20.1	37.6	37.3	0.60
ResNet34, DA=2	0.77	21.7	35.1	38.5	0.59
ResNet34v2, DA=2	0.73	29.5	29.8	**48.9**	0.53
ResNet34, DA=1+2	0.76	22.3	34.6	37.9	0.58
ResNet34v2, DA=1+2+3	**0.68**	**34.8**	**24.1**	45.1	**0.50**

και ResNet18v2 ξεπέρασαν σε επιδόσεις τα αρχικά ResNet34 και ResNet18 που χρησιμοποιήσαμε αντίστοιχα, ενώ το ResNet34v2 κατάφερε να δώσει καλύτερα αποτελέσματα ακόμα και από το μοντέλο βάσης που κάνει χρήση CNN και της παραλλαγής του DenseNet.

Για το μοντέλο ResNet34v2 εφαρμόσαμε επίσης δύο τεχνικές μετα-επεξεργασίας, την επαύξηση στο χρονικό στάδιο ελέγχου (TTA) και δυναμικά όρια (dynamic thresholds). Η τεχνική του TTA εφαρμόζει την τεχνική της επαύξησης δεδομένων αλλά κατά το στάδιο της αξιολόγησης, αφού το δίκτυο έχει ήδη εκπαιδευτεί. Συγκεκριμένα, εφαρμόζει στη βάση ελέγχου τη μέθοδο ανταλλαγής καναλιών ήχου που εφαρμόστηκε στο στάδιο εκπαίδευσης για 3 μετασχηματισμούς.

Για την επιλογή των δυναμικών ορίων υιοθετήσαμε τις τιμές που εφαρμόστηκαν στο [22], οι οποίες είναι για κάθε κλάση αντίστοιχα $[0.35, 0.36, 0.3, 0.4, 0.65, 0.6, 0.45, 0.55, 0.3, 0.3, 0.45, 0.3]$ αντί της σταθερής τιμής 0.5 που ήταν αρχικά για όλες τις κλάσεις. Η τεχνική

αυτή εφαρμόζεται γιατί η συχνότητα εμφάνισης κάποιων ηχητικών συμβάντων τείνει να είναι μικρότερη από κάποια άλλα και λόγω του υψηλού ορίου να προκύπτουν λανθασμένες προβλέψεις. Με τη χρήση ορίων προσαρμοσμένων για κάθε κλάση, έχουμε μικρότερη πιθανότητα να παραληφθεί η πρόβλεψη ενός λιγότερου δημοφιλούς γεγονότος ή να θεωρηθεί λανθασμένα ενεργό ένα περισσότερο δημοφιλές.

Από τα αποτελέσματα φαίνεται ότι ενώ οι τεχνικές επαύξησης δεδομένων δεν είχαν θετικό αποτέλεσμα όταν εφαρμόστηκαν στο αρχικό CNN, στα δίκτυα ResNet εμφανίζουν καλύτερες τιμές σχεδόν σε κάθε μετρική και παρουσιάζουν βέλτιστο ολικό σκορ.

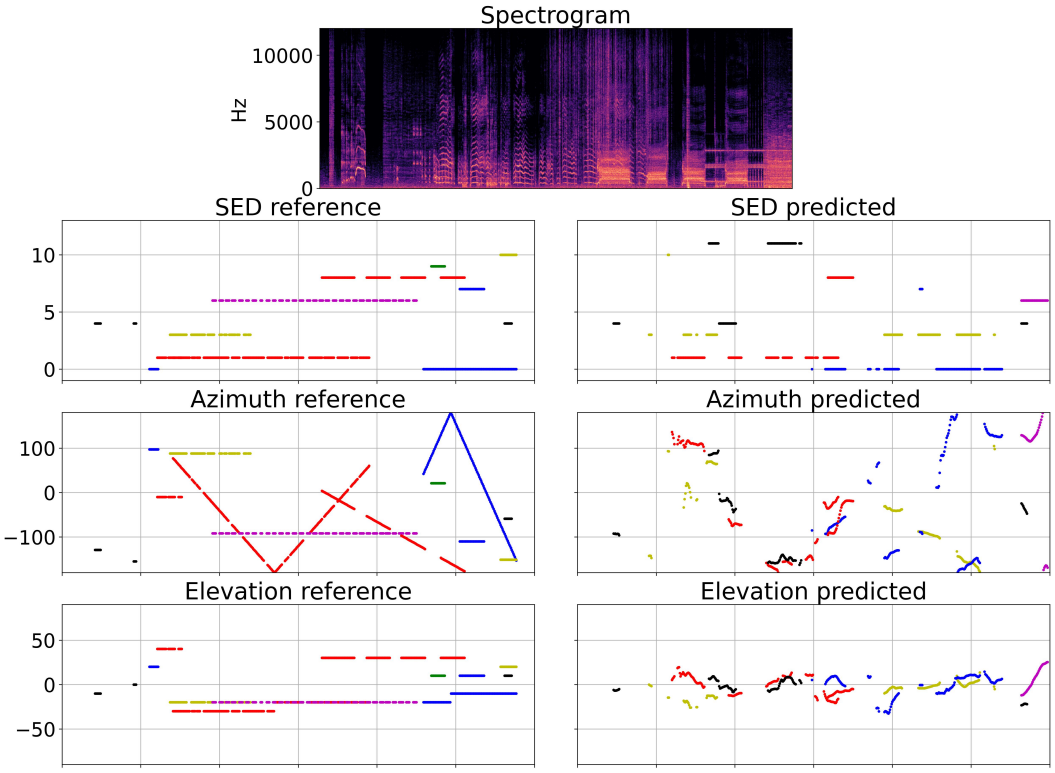
Συγκρίνοντας όλα τα μοντέλα ResNet που δημιουργήσαμε μεταξύ τους, προκύπτει ο Πίνακας 4.4 με τα καλύτερα από αυτά. Από τον Πίνακα αυτό μπορούμε να διαπιστώσουμε πως για περισσότερα στρώματα υπολειμματικών μπλοκ έχουμε μεγαλύτερη ακρίβεια στις προβλέψεις, καθώς το ResNet34 εμφάνισε καλύτερες τιμές στις μετρικές του σε σχέση με το ResNet18. Έτσι μπορεί να δει κανείς ότι το δίκτυο ResNet34v2 με DA=1+2+3 αποτελεί το βέλτιστο σύστημα μέχρι στιγμής.

Μια απεικόνιση της πρόβλεψης του μοντέλου βάσης και του ResNet34v2 σε ένα τυχαίο αρχείο ήχου φαίνεται στο Σχήμα 4.1 και 4.2 αντίστοιχα. Τα γραφήματα στα Σχήματα αυτά αντιπροσωπεύουν τα ενεργά ηχητικά συμβάντα για ένα αρχείο, καθώς και την αντίστοιχη DOA τους, σηματοδοτώντας κάθε πρόβλεψη με διαφορετικό χρώμα. Το αρχείο που επιλέγεται να γίνει η απεικόνιση είναι το fold_room1mix001, καθώς εμφανίζει κάθε δυσκολία που προσπαθούμε να επιλύσουμε, δηλαδή επικαλυπτόμενα γεγονότα ίδιας κλάσης, επικάλυψη με 3 γεγονότα, εμφάνιση γεγονότος με μεγάλη και επαναλαμβανόμενη συχνότητα εμφάνισης και παρεμβολές.

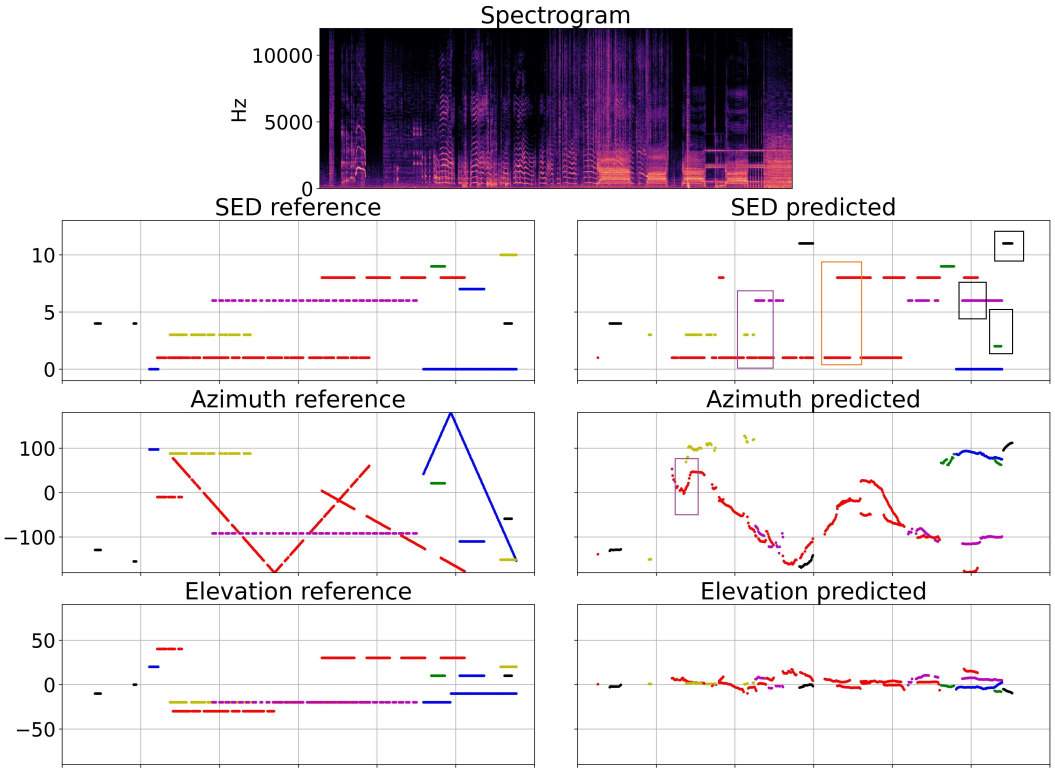
Όπως φαίνεται στο Σχήμα 4.2, στα χρονικά πλαίσια που περικλείονται από το μωβ παραλληλόγραμμο στο γράφημα των SED, έχουμε επικάλυψη τριών γεγονότων. Το δίκτυο ResNet34v2 είναι ικανό να αναγνωρίζει ταυτόχρονα και τα τρία αυτά γεγονότα. Ωστόσο για

Πίνακας 4.4: Και τα δύο μοντέλα ResNet αξιοποιούν τις τρεις τεχνικές επαύξησης δεδομένων. Το μοντέλο βάσης (baseline) δεν αξιοποιεί κάποια μέθοδο επαύξησης δεδομένων.

Μοντέλο	ER_{20°	F_{20°	LE_{CD}	LR_{CD}	$SELD_{score}$
Baseline	0.75	22.4	33.8	39.9	0.58
ResNet18v2	0.75	24.0	33.4	42.4	0.57
ResNet34v2	**0.68**	**34.8**	**24.1**	**45.1**	**0.50**



Σχήμα 4.1: Γραφική αναπαράσταση των προβλέψεων του μοντέλου βάσης πάνω στο αρχείο ήχου fold6_room1_mix001. Αριστερά απεικονίζονται οι αληθινές τιμές για το αρχείο αυτό, ενώ δεξιά παρουσιάζεται η πρόβλεψη που έγινε από το μοντέλο βάσης. Σε κάθε πλευρά, το πρώτο γράφημα από πάνω προς τα κάτω απεικονίζει τα ηχητικά συμβάντα που αναγνωρίζονται για το συγκεκριμένο αρχείο ήχου, το δεύτερο την αζιμούθια γωνία κάθε συμβάντος και το τελευταίο την ανύψωση.



Σχήμα 4.2: Γραφική αναπαράσταση των προβλέψεων του ResNet34v2 πάνω στο αρχείο ήχου fold6_room1_mix001. Αριστερά απεικονίζονται οι αληθινές τιμές για το αρχείο αυτό, ενώ δεξιά παρουσιάζεται η πρόβλεψη που έγινε από το ResNet34v2. Σε κάθε πλευρά, το πρώτο γράφημα από πάνω προς τα κάτω απεικονίζει τα ηχητικά συμβάντα που αναγνωρίζονται για το συγκεκριμένο αρχείο ήχου, το δεύτερο την αζιμούθια γωνία κάθε συμβάντος και το τελευταίο την ανύψωση.

το γεγονός κλάσης 3 με ετικέτα «γάβγισμα σκύλο» που αντιστοιχεί στην κίτρινη λωρίδα κάνει λανθασμένη πρόβλεψη λίγα πλαίσια πριν ως το γεγονός κλάσης 1 με ετικέτα «κλάμα μωρού» (κόκκινη τεθλασμένη γραμμή). Επίσης λανθασμένη πρόβλεψη έχουμε προς τα τελευταία μαύρα παραλληλόγραμμα, όπου αναγνωρίζει τις κλάσεις 6 «ήχος βημάτων» με μωβ χρώμα αντί της πραγματικής κλάσης 7 «χτύπημα πόρτας» με μπλε χρώμα, την 9 «ανδρική κραυγή» με πράσινο αντί της 7 με μπλε και την 4 «γυναικεία κραυγή» με μαύρο αντί της 6 με κίτρινο. Επιπλέον, δεν καταφέρνει να προβλέψει την επικάλυψη γεγονότος ίδιας κλάσης (κλάσης 1 με κόκκινο) που εμφανίζεται στο μωβ παραλληλόγραμμο στο γράφημα της αζιμούθιας γωνίας. Πέρα από αυτό, το σχήμα για την αζιμούθια γωνία προκύπτει αρκετά ακριβές, με το κόκκινο γεγονός που είναι μη στατική πηγή να αναγνωρίζεται αρκετά κοντά στις πραγματικές τιμές. Οι τιμές της ανύψωσης ωστόσο φαίνονται πως δεν είναι ακριβείς.

4.3 Πειράματα με Conformer

Αν και με τη χρήση υπολειμματικών δικτύων καταφέραμε να λάβουμε ικανοποιητικά αποτελέσματα, οφείλαμε να δοκιμάσουμε και ένα δίκτυο που να ενσωματώνει Conformer για να επιβεβαιώσουμε τη δημοφιλία του σε παρόμοια προβλήματα ταξινόμησης. Ο Conformer μπορεί να εφαρμοστεί στο στάδιο κωδικοποίησης αλλά και αποκωδικοποίησης, αντικαθιστώντας αρχιτεκτονικές RNN που είχαμε μέχρι στιγμής. Ωστόσο, δοκιμάσαμε να χρησιμοποιήσουμε ένα δίκτυο CNN-Conformer χωρίς RNN στο στάδιο της αποκωδικοποίησης, αλλά δεν παρατηρήθηκε κάποια σημαντική σύγκλιση μέσα σε 40 επαναλήψεις. Για αυτό το λόγο, για το πρώτο μοντέλο Conformer χρησιμοποιήσαμε δίκτυο CNN-Conformer-GRU παρόλο την αύξηση των παραμέτρων και του χρόνου υπολογισμού. Επίσης για κάθε μοντέλο Conformer στα πειράματα αξιοποιήθηκαν 2 στρώματα Conformer στοιβαγμένα σε σειρά. Τα δεδομένα εισόδου για να επεξεργαστούν από το στρώμα του Conformer μετατρέπονται σε μορφή (B, T, dim), όπου B το batch size, T τα χρονικά πλαίσια και $dim = C \cdot F$ με τα C και F να συμβολίζουν τον αριθμό καναλιών και δεσμών Mel αντίστοιχα. Στις δοκιμές που ακολουθούν παρουσιάζονται μοντέλα που αξιοποιούν διαφορετικές τιμές για τη διάσταση dim.

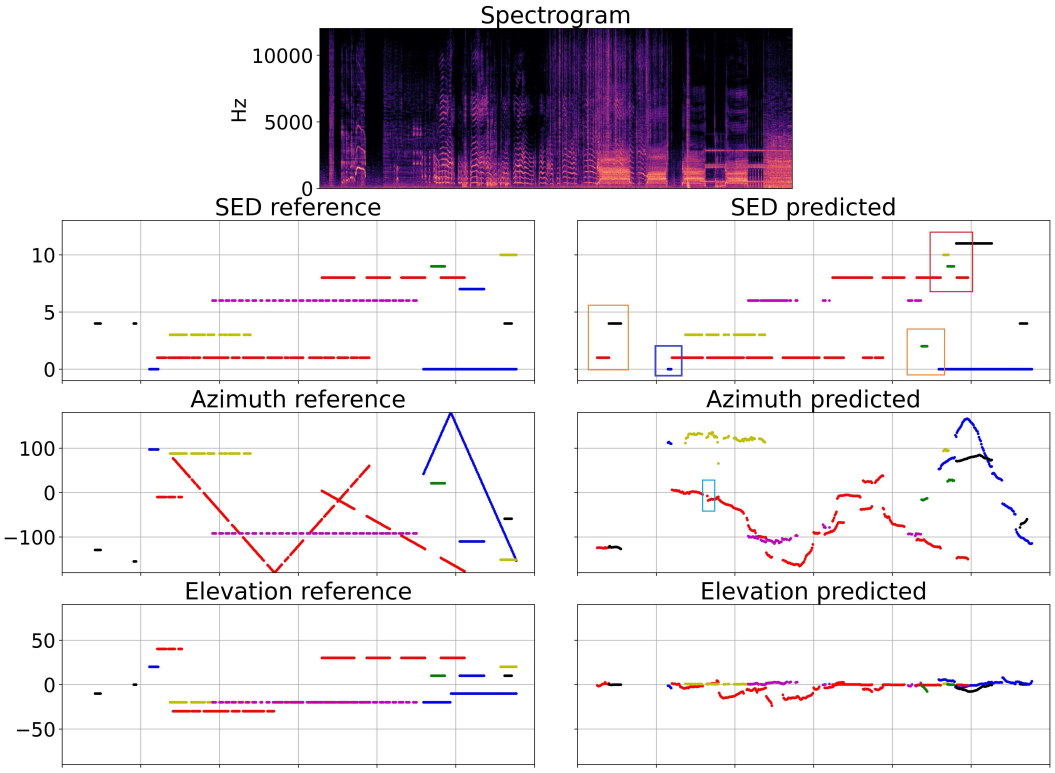
Για κάθε πείραμα αξιοποιήθηκαν περαιτέρω η τεχνική του SWA για κάθε μοντέλο ξεχωριστά, καθώς και ο ίδιος ελεγκτής βαθμού εκμάθησης που εφαρμόστηκε και για τα υπολειμματικά μοντέλα. Κάθε μοντέλο, εκτός αν αναφέρεται διαφορετικά, εκπαιδεύτηκε με βελτι-

Πίνακας 4.5: Πειράματα πάνω σε Conformer

Μοντέλο	dim	ER_{20°	F_{20°	LE_{CD}	LR_{CD}	$SELD_{score}$
Baseline	-	0.75	22.4	33.8	39.9	0.58
Conformer	60	0.87	10.1	48.9	25.1	0.70
Conformer, AdaBelief	60	0.89	7.6	45.7	24.4	0.72
Conformer	256	0.76	24.2	29.2	36.1	0.58
Conformer, heads=4, depth=24	256	0.79	21.2	31.5	34.5	0.57
Conformer, DA=1	256	0.74	28.7	29.0	44.9	0.54
Conformer, DA=2	256	0.73	29.3	28.6	45.2	0.53
Conformer, MEL-BANDS=128, DA=1	256	0.73	29.1	30.4	47.7	0.53
Conformer, MEL-BANDS=128, DA=2	256	0.74	26.8	36.2	43.1	0.54
Conformer, MEL-BANDS=128, DA=1+2	256	0.74	25.7	35.0	46.4	0.54
Conformer, DA=1+2+3	256	0.73	29.0	29.8	42.4	0.55
Conformer, different-CNN, DA=2	256	0.75	26	28.4	40.4	0.55
DenseNet-Conformer, DA=1+2+3	256	0.71	31.2	29.4	46.0	0.53
Squeeze-Conformer, DA=1+2+3	256	**0.65**	**39.4**	**27.5**	**55.1**	**0.47**

στοποιητή Adam αρχικού ρυθμού εκπαίδευσης 0.00001 και σφάλμα ελαχίστου τετραγώνου για τη συνάρτηση σφάλματος. Επίσης, εκτός αν επισημαίνεται, κάθε μοντέλο εκπαιδεύτηκε για δεδομένα με 64 δέσμες Mel με 8 κεφαλές βάθους 64 το κάθε στρώμα. Το μέγεθος φίλτρου για τη μονάδα συνέλιξης του Conformer επιλέχθηκε να είναι 31. Τα τελικά αποτελέσματα για τα μοντέλα Conformer φαίνονται στον Πίνακα 4.5.

Από τα αποτελέσματα φαίνεται πως ο αριθμός των δεσμών Mel δεν επηρέασε σημαντικά την ακρίβεια των μετρικών. Επίσης φαίνεται πως για μεγαλύτερο dim καθώς και για περισσότερες κεφαλές έχουμε και καλύτερα αποτελέσματα. Η χρήση AdaBelief αντί για το τυπικό Adam χειροτέρεψε τις μετρικές. Τα τελευταία 3 μοντέλα που παρουσιάζονται στον Πίνακα αφορούν την αλλαγή του αρχικού CNN μέρους του δικτύου. Για κάθε μοντέλο Conformer, χρησιμοποιούμε 2 στρώματα CNN μεγέθους 64 και 256 αντίστοιχα, τα οποία ακολουθούνται από 2 στρώματα μέσης ομαδοποίησης σε σειρά για να έρθει η είσοδος στη μορφή $(B, 256, T/5, 2)$. Αντιθέτως για το μοντέλο Conformer different-CNN χρησιμοποιήσαμε ένα δίκτυο CNN παρόμοιο με αυτό του μοντέλου βάσης. Το μοντέλο DenseNet-Conformer αλ-



Σχήμα 4.3: Γραφική αναπαράσταση των προβλέψεων του Squeeze-Conformer πάνω στο αρχείο ήχου fold6_room1_mix001. Αριστερά απεικονίζονται οι αληθινές τιμές για το αρχείο αυτό, ενώ δεξιά παρουσιάζεται η πρόβλεψη που έγινε από το μοντέλο Squeeze-Conformer. Σε κάθε πλευρά, το πρώτο γράφημα από πάνω προς τα κάτω απεικονίζει τα ηχητικά συμβάντα που αναγνωρίζονται για το συγκεκριμένο αρχείο ήχου, το δεύτερο την αζιμούθια γωνία κάθε συμβάντος και το τελευταίο την ανύψωση.

λάζει το CNN με το ίδιο δίκτυο DenseNet που χρησιμοποιήθηκε μεμονωμένα και φαίνεται η επίδοσή του στον Πίνακα 4.1. Τέλος για το Squeeze-Conformer, το οποίο απέδωσε και τα βέλτιστα αποτελέσματα, αλλάξαμε το αρχικό CNN με ένα όμοιο δίκτυο CNN 3 στρωμάτων όπως το μοντέλο βάσης, αλλά προσθέτωντας ένα χωρικό και καναλικό στρώμα πίεσης-διέγερσης (squeeze-excitation) όπως στο [61].

Η απεικόνιση της πρόβλεψης του Squeeze-Conformer παρουσιάζεται στο Σχήμα 4.3. Η πρόβλεψη έγινε για το ίδιο αρχείο εισόδου fold6_room1_mix001. Το δίκτυο αναγνωρίζει λανθασμένα την μαύρη κλάση που αντιστοιχεί σε «γυναικεία κραυγή» για συμβάν της κόκκινης κλάσης, δηλαδή για «κλάμα μωρού», όπως φαίνεται στο πρώτο πορτοκαλί παραλληλόγραμμο, αλλά επιδιορθώνει το λάθος καθώς συνεχίζει δεξιά στον οριζόντιο άξονα. Ωστόσο καταφέρνει να προβλέψει σωστά 3 επικαλυπτόμενα γεγονότα καθώς επίσης προσπαθεί να προβλέψει σωστά και την επικάλυψη ίδιας κλάσης στιγμιαία, όπως φαίνεται από το μπλε παραλληλόγραμμο στο γράφημα της αζιμούθιας γωνίας. Η πρόβλεψη για την κλάση 6 που αναπαρίσταται με κίτρινο παραμένει σταθερή χωρίς λάθος πρόβλεψη όπως στο ResNet. Επίσης σε αντίθεση με το ResNet, όπως φαίνεται στο τελευταίο πορτοκαλί παραλληλόγραμμο, προβλέπει σωστά για το πράσινο γεγονός κλάσης 8 και γενικά είναι πιο ακριβές στο να προβλέπει συμβάντα μικρής χρονικής διάρκειας. Ωστόσο, για το γεγονός κλάσης 6 με μωβ που αποτελεί συμβάν με μεγάλη επαναλαμβανόμενη συχνότητα, φαίνεται πως δυσκολεύεται να προβλεφθεί ολόκληρο για όλη τη διάρκειά του, όπως έγινε και για το ResNet. Οι προβλέψεις της ανύψωσης παραμένουν μη ακριβείς, αλλά καλύτερες από τα προηγούμενα μοντέλα.

4.4 Πειράματα με ResNet-Conformer

Από την έρευνα στο [28] εμπνεύστηκαμε να ανταλλάξουμε το CNN στάδιο στα μοντέλα Conformer που είχαμε ως τώρα με ResNet. Δοκιμάσαμε αυτή την αρχιτεκτονική για ResNet34 και για ακόμα ένα ResNet με 1 ταυτοτικό μπλοκ ανά στάδιο. Τα αποτελέσματα για τα ResNet-Conformer μοντέλα φαίνονται στον Πίνακα 4.6. Από τον πίνακα αυτό μπορούμε να δούμε πως το μοντέλο που χρησιμοποιεί ResNet34 και έχει περισσότερα υπολειμματικά μπλοκ, δίνει και καλύτερα αποτελέσματα. Φαίνεται επίσης πόσο βελτιώνονται ακόμα τα αποτελέσματα αν γίνει χρήση επαύξησης δεδομένων και TTA, όπου συγκεκριμένα για το ResNet34-Conformer βελτιώνει κάθε μετρική του. Ωστόσο τα υβριδικά αυτά μοντέλα δεν καταφέρνουν να ξεπεράσουν σε επιδόσεις τα αντίστοιχα μοντέλα που το αποτελούν ξεχω-

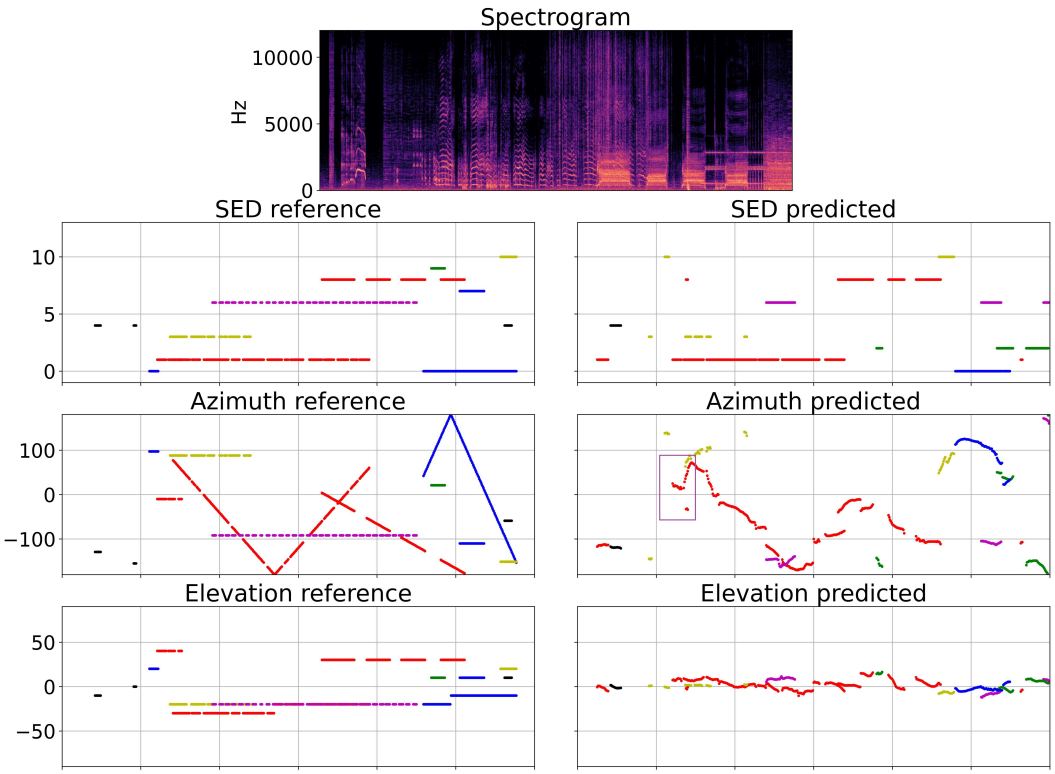
Πίνακας 4.6: Πειράματα με ResNet-Conformer

Μοντέλο	ER_{20°	F_{20°	LE_{CD}	LR_{CD}	$SELD_{score}$
Baseline	0.75	22.4	33.8	39.9	0.58
ResNet34-Conformer	0.76	19.2	35.1	36.7	0.59
ResNet34-Conformer, DA=1+2+3, TTA	**0.72**	**29.7**	**29.7**	**45.3**	**0.53**
ResNet1ID-Conformer, DA=1+2+3	0.78	21.6	37.7	44.8	0.58
ResNet1ID-Conformer, DA=1+2+3, TTA	0.73	25.6	31.8	41.8	0.55

ριστά. Η απεικόνιση των προβλέψεων για το αρχείο fold6_room1_mix001 για τα μοντέλα ResNet34-Conformer με λιγότερα στρώματα φαίνεται στο Σχήμα 4.4. Σε αυτή φαίνεται πως το δίκτυο ResNet-Conformer καταφέρνει στιγμιαία να προβλέψει το επικαλυπτόμενο συμβάν της ίδιας κλάσης 1, το οποίο περιβάλλεται από το μωβ παραλληλόγραμμο. Σε αυτό το πλαίσιο, αναδεικνύεται η ταυτόχρονη παρουσία δύο διαφορετικών κόκκινων γραμμών, το οποίο μεταφράζεται στην πρόβλεψη ενός συμβάντος ίδιας κλάσης στα ίδια χρονικά πλαίσια που εμφανίζεται και στην πραγματική πρόβλεψη στο αριστερό γράφημα. Η κλίση της γωνίας των συμβάντων δεν είναι ακριβής, αλλά επιτυγχάνεται η γενική αναγνώρισή τους. Επίσης παρατηρείται ότι η πρόβλεψη των δύο επικαλυπτόμενων γεγονότων ίδιας κλάσης φαίνεται πιο ακριβής σε σχέση με αυτήν που καταφέρνει το Squeeze-Conformer, καθώς έχουμε μια ευθεία γραμμή για το πρώτο γεγονός όμοια με τις αληθινές τιμές που φαίνονται στο αριστερό γράφημα. Όμως οι τιμές της ανύψωσης δεν έχουν καταφέρει να βελτιωθούν.

4.5 Πειράματα με συγχώνευση μοντέλων

Για τα μοντέλα συγχώνευσης συνδυάσαμε αρχιτεκτονικές που απεδείχθησαν βέλτιστες σε συγκεκριμένες μετρικές. Για παράδειγμα, το μοντέλο Squeeze-Conformer απέδωσε την καλύτερη τιμή για τις μετρικές ER_{20° και LR_{CD}. Το ResNet34 αντιθέτως είχε καλύτερο LE_{CD} από τα μοντέλα Conformer. Γενικά από τις αρχιτεκτονικές παρατηρούμε καλές τιμές στις μετρικές που αφορούν το DOA για μοντέλα Conformer, ενώ τα μοντέλα ResNet υπερτερούν σε τιμές των μετρικών SED, ειδικά στο F_{20°. Έτσι, επειδή κάθε μοντέλο έχει τα δικά του χαρακτηριστικά και υπερτερεί σε διαφορετικούς τομείς από τα υπόλοιπα, συνδυάζουμε τα πλεονεκτήματα όλων των διαφορετικών μοντέλων συγχωνεύοντάς τα σε νέες



Σχήμα 4.4: Γραφική αναπαράσταση των προβλέψεων του ResNet34-Conformer πάνω στο αρχείο ήχου fold6_room1_mix001. Αριστερά απεικονίζονται οι αληθινές τιμές για το αρχείο αυτό, ενώ δεξιά παρουσιάζεται η πρόβλεψη που έγινε από το μοντέλο ResNet34-Conformer. Σε κάθε πλευρά, το πρώτο γράφημα από πάνω προς τα κάτω απεικονίζει τα ηχητικά συμβάντα που αναγνωρίζονται για το συγκεκριμένο αρχείο ήχου, το δεύτερο την αζιμούθια γωνία κάθε συμβάντος και το τελευταίο την ανύψωση. Στο δεξί γράφημα, φαίνεται στο μωβ παραλληλόγραμμο η σωστή ανίχνευση και διαχωρισμός δύο συμβάντων ίδιας κλάσης με διαφορετικό DOA.

αρχιτεκτονικές. Ως πρώτο μοντέλο συγχώνευσης επιλέχθηκε η συγχώνευση των καλύτερων Conformer μοντέλων με κάποιες δομές CNN. Συγκεκριμένα:

- Το μοντέλο Ensemble 1 συνδυάζει δύο CNN, το αρχικό CNN του μοντέλου βάσης και ένα DenseNet, μαζί με δύο Conformer.
- Το Ensemble 2 συνδυάζει CNN μοντέλα με ResNet-Conformer.
- Το Ensemble 3 συνδυάζει CNN, Conformer και ResNet.
- Το Ensemble 4 συνδυάζει ResNet και τρεις Conformer.
- Το Ensemble 5 συνδυάζει ResNet, τρεις Conformer και ένα ResNet-Conformer.

Πίνακας 4.7: Πειράματα πάνω σε συγχώνευση μοντέλων

Μοντέλο	ER_{20°	F_{20°	LE_{CD}	LR_{CD}	$SELD_{score}$
Baseline	0.75	22.4	33.8	39.9	0.58
Ensemble 1	**0.65**	38.4	24.5	48.5	0.48
Ensemble 2	0.73	28.6	24.4	35.1	0.55
Ensemble 3	0.67	**39.8**	**22.5**	**48.7**	**0.47**
Ensemble 4	0.68	36.6	23.7	44.1	0.50
Ensemble 5	0.66	38.4	23.7	44.1	0.49
Ensemble 6	0.67	37.4	23.1	43.9	0.50
Ensemble 7	0.68	35.7	24.4	44.3	0.50
Ensemble 8	0.66	38.5	23.1	45.8	0.49

- Το Ensemble 6 συνδυάζει τρεις διαφορετικούς Conformer με ένα CNN.
- Το Ensemble 7 συνδυάζει δύο ResNet-Conformers και δύο Conformer.
- Το Ensemble 8 συνδυάζει ένα CNN, ένα ResNet, ένα Conformer και ένα ResNet-Conformer, και συγκεκριμένα ένα DenseNet, ένα ResNet34, ένα Squeeze-Conformer και ένα ResNet34-Conformer.

Για το Ensemble 8 αξίζει να σημειωθεί ότι χρησιμοποιεί τα βέλτιστα μοντέλα από κάθε αρχιτεκτονική.

Τα αποτελέσματα για όλα τα μοντέλα συγχώνευσης φαίνονται στον Πίνακα 4.7. Τα τελικά μοντέλα συγχώνευσης προέκυψαν μετά από διάφορες δοκιμές των μοντέλων που τα απαρτίζουν, όσον αφορά τις παραμέτρους που αυτά εκπαιδεύτηκαν και τον αριθμό του κάθε ενός που συμμετέχει στη συγχώνευση. Επιπλέον, για τα μοντέλα αυτά αξιοποιήθηκε κατά το στάδιο της πρόβλεψης η τεχνική TTA.

Από τα αποτελέσματα φαίνεται ότι με τα μοντέλα συγχώνευσης δεν έχουμε κάποια σημαντική βελτίωση για το γενικό $SELD_{score}$. Ωστόσο, παρατηρείται για τη μετρική του LE_{CD} μείωση σε σύγκριση με τα ανεξάρτητα μοντέλα. Συγκεκριμένα, η χαμηλότερη τιμή που είχε προκύψει για το LE_{CD} στα ανεξάρτητα μοντέλα ήταν για το ResNet34v2 με τιμή 24.1, ενώ τώρα με τη μέθοδο της συγχώνευσης η τιμή αυτή έχει πέσει στο 22.4. Επίσης ελάχιστη βελτίωση υπάρχει στο F_{20°, με την τιμή του να αυξάνεται από 39.4 που επιτευχθεί για το μοντέλο του Squeeze-Conformer στο 39.8. Παρόλα αυτά, η τιμή του LR_{CD} δεν βελτιώθηκε,

αντιθέτως μειώθηκε σε σχέση με τα ανεξάρτητα μοντέλα. Τέλος, η τιμή ER_{20°, αν και καλύτερη από τις γενικές τιμές που προέκυψαν από τα περισσότερα μοντέλα, δεν σημείωσε καλύτερη τιμή από αυτή του ανεξάρτητου μοντέλου Squeeze-Conformer. Γενικά αν και τα αποτελέσματα των μοντέλων συγχώνευσης είναι καλύτερα από τα περισσότερα μοντέλα που συμμετέχουν στη συγχώνευση, δεν καταφέρνουν να ξεπεράσουν σε επίδοση τα δύο βέλτιστα μοντέλα ResNet34v2 και Squeeze-Conformer. Ωστόσο, επιτυγχάνουν καλύτερες τιμές για τις μετρικές F_{20° και LE_{CD} από κάθε άλλο μοντέλο.

Κεφάλαιο 5

Συμπεράσματα

Σε αυτή τη διπλωματική παρουσιάσαμε διάφορα νευρωνικά δίκτυα για την επίλυση του προβλήματος που πραγματεύεται ο διαγωνισμός DCASE 2021 Task 3, με θέμα τον εντοπισμό και την ανίχνευση ακουστικών γεγονότων (SELD). Συγκεκριμένα, ασχοληθήκαμε με αρχιτεκτονικές που χρησιμοποιούνται για αναγνώριση εικόνων, τα ResNet, καθώς και σε γλωσσικά μοντέλα, τους Conformer, για την ταυτόχρονη αντιμετώπιση των υποπροβλημάτων της ταξινόμησης και της εύρεσης της θέσης των ακουστικών συμβάντων. Παράλληλα, αξιοποιήσαμε και αξιολογήσαμε διάφορες τεχνικές βελτίωσης των μοντέλων όπως η επαύξηση των δεδομένων καθώς και μεθόδους συνδυασμού τους. Το βέλτιστο μοντέλο βρήκαμε ότι είναι το Squeeze-Conformer, ένα δίκτυο με δύο στοιβαγμένα στρώματα Conformer το οποίο χρησιμοποιεί squeeze-excitation CNN στο στάδιο της εξαγωγής των χαρακτηριστικών. Το μοντέλο αυτό βρέθηκε ότι ξεπερνάει το μοντέλο βάσης που δίνεται από το διαγωνισμό σε όλες τις μετρικές αξιολόγησης και στο γενικό SELD σκορ. Όσον αφορά τις τεχνικές επαύξησης δεδομένων βρέθηκε ότι ενώ δεν βελτιώνει την επίδοση του αρχικού CNN μοντέλου βάσης, η θετική επίδρασή του στα υπόλοιπα μοντέλα ResNet και Conformer είναι εμφανής, βελτιώνοντας κάθε μετρική αξιολόγησης. Ειδικότερα, η μέθοδος τυχαίας μετατόπισης φάσματος φάνηκε να δίνει καλύτερα αποτελέσματα από αυτή της χρήσης μάσκας στο πεδίο του χρόνου και της συχνότητας και βελτιώνει σημαντικά κάθε μετρική, ιδιαίτερα τις ER_{20°, F_{20° και LR_{CD}. Η τεχνική ανταλλαγής καναλιών ήχου, ως τεχνική χωρικής επαύξησης, σημείωσε σημαντική βελτίωση των μετρικών που αφορούν τον εντοπισμό τοποθεσίας των συμβάντων, ειδικά σημείωσε την καλύτερη τιμή για τη μετρική LE_{CD}. Ακόμα, η χρήση της τεχνικής επαύξησης κατά το στάδιο ελέγχου έδωσε σημαντική βελτίωση στις μετρικές ER_{20° και F_{20° σε σχέση με μοντέλα που δεν την αξιοποίησαν.

Αν και το μοντέλο Squeeze-Conformer που αναπτύξαμε κατάφερε να επιτύχει τους στόχους που τέθηκαν, θα μπορούσαν να χρησιμοποιηθούν περαιτέρω τεχνικές για ένα ακόμα καλύτερο μοντέλο. Για το λόγο αυτό, ως μελλοντική επιδίωξη θα μπορούσαμε να επικεντρωθούμε πέρα από την επιρροή των νευρωνικών δικτύων και στον τρόπο της επεξεργασίας των δεδομένων εισόδου, όπου για παράδειγμα μπορεί να γίνεται η πρόβλεψη σε παράθυρα μικρότερης διάρκειας και η γενική συγχώνευση όλων των προβλέψεων για το τελικό αποτέλεσμα. Αυτή η τεχνική υιοθετήθηκε από κάποιες υψηλά καταταγμένες προσπάθειες στον διαγωνισμό DCASE 2021 Task 3. Επίσης, θα μπορούσαμε να ασχοληθούμε με τεχνικές που αλλάζουν τη μορφή της εξόδου ACCDOA σε μια μορφή πολλαπλών κομματιών για την επιτυχή αναγνώριση επικαλυπτόμενων συμβάντων ίδιας κλάσης. Για μια τέτοια προσέγγιση θα μπορούσαμε να υιοθετήσουμε ένα σύστημα ανεξάρτητων συμβάντων [62]. Επιπλέον, θα ήταν καλή ιδέα να δούμε την επίδοση συστημάτων που κάνουν ξεχωριστή πρόβλεψη των SED και DOA με τη μελέτη και εφαρμογή μεθόδων όπως του παράλληλου μοιράσματος της πληροφορίας [32]. Τέλος, η τεχνική αναζήτησης νευρωνικής αρχιτεκτονικής (neural architecture search) [63], [64] θα μπορούσε να χρησιμοποιηθεί για την πιο εύκολη εύρεση διαφορετικών μοντέλων και το βέλτιστο συνδυασμό υπερ-παραμέτρων για καλύτερα μοντέλα συγχώνευσης.

Βιβλιογραφία

[1] Archontis Politis, Sharath Adavanne, Daniel Krause, Antoine Deleforge, Prerak Srivastava, and Tuomas Virtanen. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection. In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, pages 125–129, Barcelona, Spain, 2021.

[2] Conformer: Convolution-augmented Transformer for Speech Recognition. https://arxiv.org/abs/2005.08100.

[3] Sound Source Localization for Robotic Applications. https://elib.dlr.de/137625/1/Sound_Source_Localization_for_Robotic_Application.release.pdf.

[4] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210, 2017.

[5] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. Audio surveillance: a systematic review. https://arxiv.org/abs/1409.7787, 2014.

[6] Alexa Amazon Wiki. https://en.wikipedia.org/wiki/Amazon_Alexa#cite_note-2.

[7] DCASE 2021 Task 3. https://dcase.community/challenge2021/task-sound-event-localization-and-detection.

[8] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. http://arxiv.org/abs/1807.00129, 2018.

[9] Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, 2021.

[10] 7 Types of Classification Algorithms. `https://analyticsindiamag.com/7-types-classification-algorithms/`.

[11] Sound Event Localization and Tracking. `https://www.aane.in/research/computational-audio-scene-analysis-casa/sound-event-localization-and-tracking`.

[12] Srđan Kitić and Alexandre Guérin. TRAMP: tracking by a real-time ambisonic-based particle filter. `https://arxiv.org/abs/1810.04080`, 2018.

[13] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[14] Estimation of Signal Parameters via Rotational Invariance Techniques. `https://en.wikipedia.org/wiki/Estimation_of_signal_parameters_via_rotational_invariance_techniques`.

[15] Ryu Takeda and Kazunori Komatani. Sound source localization based on deep neural networks with directional activate function exploiting phase information. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 405–409, 2016.

[16] 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.

[17] Annamaria Mesaros, Sharath Adavanne, Archontis Politis, Toni Heittola, and Tuomas Virtanen. Joint measurement of localization and detection of sound events. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 333–337, 2019.

[18] Jinbo Hu, Yin Cao, Ming Wu, Qiuqiang Kong, Feiran Yang, Mark D. Plumbley, and Jun Yang. A track-wise ensemble event independent network for polyphonic sound event localization and detection. `https://arxiv.org/abs/2203.10228`, 2022.

[19] DCASE Community. `https://dcase.community/`.

[20] Huang Daolang and Perez Ricardo. SSELDnet: a fully end-to-end sample-level framework for sound event localization and detection. Technical report, DCASE2021 Workshop, 2021.

[21] Sooyoung Park, Youngho Jeong, and Taejin Lee. Self-attention mechanism for sound event localization and detection. Technical report, DCASE2021 Workshop, 2021.

[22] Daniel Rho, Seungjin Lee, JinHyeock Park, Taesoo Kim, Jiho Chang, and Jonghwan Ko. A combination of various neural networks for sound event localization and detection. Technical report, DCASE2021 Workshop, 2021.

[23] Yuxuan Zhang, Shuo Wang, Zihao Li, Kejian Guo, Shijin Chen, and Yan Pang. Data augmentation and class-based ensembled CNN-Conformer networks for sound event localization and detection. Technical report, DCASE2021 Workshop, 2021.

[24] Thi Ngoc Tho Nguyen, Karn Watcharasupat, Ngoc Khanh Nguyen, Douglas L. Jones, and Woon Seng Gan. DCASE 2021 Task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection. Technical report, DCASE2021 Workshop, 2021.

[25] Jisheng Bai, Zijun Pu, and Jianfeng Chen. DCASE 2021 Task 3: SELD system based on ResNet and random segment augmentation. Technical report, DCASE2021 Workshop, 2021.

[26] Koichi Miyazaki, Tatsuya Komatsu, Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, and K. Takeda. Conformer-based sound event detection with semi-supervised learning and data augmentation. Technical report, DCASE2020 Workshop, 2020.

[27] Qing Wang, Huaxin Wu, Zijun Jing, Feng Ma, Yi Fang, Yuxuan Wang, Tairan Chen, Jia Pan, Jun Du, and Chin-Hui Lee. A model ensemble approach for sound event localization and detection. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5, 2021.

[28] Qing Wang, Jun Du, Huaxin Wu, Jia Pan, Feng Ma, and Chin-Hui Lee. A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection. https://arxiv.org/abs/2101.02919.

[29] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le. SpecAugment: a simple data augmentation method for automatic speech recognition. https://arxiv.org/abs/1904.08779, 2019.

[30] Kazuki Shimada, Naoya Takahashi, Yuichiro Koyama, Shusuke Takahashi, Emiru Tsunoo, Masafumi Takahashi, and Yuki Mitsufuji. Ensemble of ACCDOA- and EINV2-based systems with D3Nets and impulse response simulation for sound event localization and detection. Technical report, DCASE2021 Workshop, 2021.

[31] Nelson Yalta, Takashi Sumiyoshi, and Yohei Kawaguchi. The Hitachi DCASE 2021 Task 3 system: Handling directive interference with self attention layers. Technical report, DCASE2021 Workshop, 2021.

[32] Sang-Hoon Lee, Jung-Wook Hwang, Sang-Buem Seo, and Hyung-Min Park. Sound event localization and detection using cross-modal attention and parameter sharing for DCASE 2021 challenge. Technical report, DCASE2021 Workshop, 2021.

[33] Xiao-Lei Zhang and DeLiang Wang. Boosting contextual information for deep neural network based voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):252–264, 2016.

[34] TAU Dataset 2021. https://zenodo.org/record/4844825#.YoIjknXP1D8.

[35] Baseline Model for DCASE 2021 Task 3. https://github.com/sharathadavanne/seld-dcase2021.

[36] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen. Overview and evaluation of sound event localization and detection in DCASE 2019. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:684–698, 2020.

[37] Keras API. https://keras.io/.

[38] TensorFlow. https://www.tensorflow.org/resources/learn-ml?gclid=Cj0KCQjw-JyUBhCuARIsANUqQ_L-fCdD69_LrqdHBx2Dby_X4kfOlt5Eiyn5tQFFw9i9VAC_ze74zMcaAkb1EALw_wcB.

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980, 2014.

[40] Stochastic Gradient Descent Wiki. https://en.wikipedia.org/wiki/Stochastic_gradient_descent.

[41] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James S. Duncan. AdaBelief optimizer: Adapting step-sizes by the belief in observed gradients. https://arxiv.org/abs/2010.07468, 2020.

[42] Chi-Feng Wang. The Vanishing Gradient Problem. https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. http://arxiv.org/abs/1512.03385, 2015.

[44] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. http://arxiv.org/abs/1608.06993, 2016.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. https://arxiv.org/pdf/1706.03762.pdf, 2017.

[46] Transformer Explanation and Implementation. https://colab.research.google.com/github/tensorflow/text/blob/master/docs/tutorials/transformer.ipynb?hl=ro.

[47] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: attentive language models beyond a fixed-length context. https://arxiv.org/abs/1901.02860, 2019.

[48] Transformer Neural Networks - EXPLAINED! (Attention is all you need). https://www.youtube.com/watch?v=TQQlZhbC5ps&list=PLCN0oQdYUUCujGJuN_cFhSgFyH6TVY1h1&index=6.

[49] Illustrated Guide to Transformers Neural Network: A Step by Step Explanation. https://www.youtube.com/watch?v=4Bdc55j80l8&list=PLCN0oQdYUUCujGJuN_cFhSgFyH6TVY1h1&index=27&t=17s.

[50] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. https://arxiv.org/abs/2004.11886, 2020.

[51] Υπερμοντελοποίηση στην Μηχανική Μάθηση. 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/.

[52] Implementation of Conformer: Convolution-augmented Transformer for Speech Recognition. https://github.com/lucidrains/conformer, 2021.

[53] Soohwan Kim. Conformer: PyTorch implementation of conformer: convolution-augmented transformer for speech recognition. https://github.com/sooftware/conformer, 2021.

[54] Overfitting in a Neural Network Explained. https://deeplizard.com/learn/video/DEMmkFC6IGM.

[55] Thi Ngoc Tho Nguyen, Karn N. Watcharasupat, Ngoc Khanh Nguyen, Douglas L. Jones, and Woon-Seng Gan. SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1749–1762, 2022.

[56] Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results. https://www.toptal.com/machine-learning/ensemble-methods-machine-learningensemble.

[57] Yuan Gao, Zixiang Cai, and Lei Yu. Intra-ensemble in neural networks. https://arxiv.org/abs/1904.04466, 2019.

[58] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. https://arxiv.org/abs/1803.05407, 2018.

[59] Hao Guo, Jiyong Jin, and Bin Liu. Stochastic weight averaging revisited. https://arxiv.org/abs/2201.00519, 2022.

[60] A Gentle Introduction to Ensemble Learning Algorithms. https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/.

[61] Javier Naranjo-Alcazar, Sergi Perez-Castanos, Maximo Cobos, Francesc J. Ferri, and Pedro Zuccarello. Sound event localization and detection using squeeze-excitation residual CNNs. Technical report, DCASE2021 Workshop, 2021.

[62] Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, and Mark D. Plumbley. An improved event-independent network for polyphonic sound event localization and detection. https://arxiv.org/abs/2010.13092, 2020.

[63] Neural Architecture Search. https://en.wikipedia.org/wiki/Neural_architecture_search.

[64] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. https://arxiv.org/abs/2003.13678, 2020.