ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ - ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ

UNIVERSITY OF
PATRAS
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΤΟΜΕΑΣ: Τηλεπικοινωνιων και Τεχνολογιας Πληροφοριας

(Τ&ΤΠ)

ΕΡΓΑΣΤΗΡΙΟ: Ενσυρματης Επικοινωνιας

Διπλωματικη Εργασια

του φοιτητή του Τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών της Πολυτεχνικής Σχολής του Πανεπιστημίου Πατρών

Παναγιωτη Ζαχου του Σωτηριου

ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ: 1019336

Θέμα

Εντοπισμός Ακουστικής Πηγής με Βαθιά Νευρωνικά Δίκτυα Χρησιμοποιώντας Αμφιωτικές Παραμέτρους

Επιβλέπων

Καθηγητής Ιωάννης Μουρτζόπουλος

Αριθμός Διπλωματικής Εργασίας:

Πάτρα, Ιούλιος 2020

ΠΙΣΤΟΠΟΙΗΣΗ

Πιστοποιείται ότι η διπλωματική εργασία με θέμα

Εντοπισμός Ακουστικής Πηγής με Βαθιά Νευρωνικά Δίκτυα Χρησιμοποιώντας Αμφιωτικές Παραμέτρους

του φοιτητή του Τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών

Παναγιώτη Ζάχου του Σωτηρίου

(Α.Μ.: 1019336)

παρουσιάτηκε δημόσια και εξετάστηκε στο τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών στις

__/__/___

Ο Επιβλέπων	Ο Διευθυντής του Τομέα
Ιωάννης Μουρτζόπουλος	Ιωάννης Μουρτζόπουλος
Καθηγητής	Καθηγητής

Στοιχεία διπλωματικής εργασίας

Θέμα: **Εντοπισμός Ακουστικής Πηγής με Βαθιά Νευρωνικά Δίκτυα Χρησιμοποιώντας Αμφιωτικές Παραμέτρους**

Φοιτητής: **Παναγιώτης Ζάχος του Σωτηρίου**

Ομάδα επίβλεψης
Καθηγητής Ιωάννης Μουρτζόπουλος
Εργαστήρια
Ενσύρματης Επικοινωνίας

Περίοδος εκπόνησης της εργασίας:
Ιανουάριος 2019 - Ιούλιος 2020

Η εργασία αυτή γράφτηκε στο XeLaTeX και χρησιμοποιήθηκε η γραμματοσειρά GFS Didot του Greek Font Society.

Περίληψη

Στην παρούσα Διπλωματική Εργασία προτείνεται μια νέα μέθοδος συμπίεσης των διαστάσεων των αμφιωτικών παραμέτρων ILD και ITD. Ο στόχος της μεθόδου είναι η βέλτιστη αξιοποίηση αυτών των παραμέτρων για την εκπαίδευση Νευρωνικών Δικτύων καθώς και σε άλλες εφαρμογές μηχανικής μάθησης. Υλοποιούνται επίσης δύο μοντέλα Νευρωνικών Δικτύων, ένα πλήρως διασυνδεδεμένο με δομή που παρομοιάζει έναν πολυεπίπεδο Perceptron, καθώς και ένα CNN, με στόχο την εκτίμηση της γωνίας άφιξης μιας ακουστικής διέγερσης, σε πραγματικά δωμάτια με αντήχηση, από τις συμπιεσμένες παραμέτρους. Τα μοντέλα επιτυγχάνουν εξαιρετική ακρίβεια στην εκτίμηση με το μέσο λάθος να κυμαίνεται κάτω από 5^o, αξιοποιώντας τις παραμέτρους που έχουν συμπιεστεί κατά έναν παράγοντα $\sim 88\%$. Η προσέγγιση εκτίμησης της γωνίας άφιξης σε αυτή την εργασία, ξεπερνά παραδοσιακές μεθόδους εντοπισμού που αξιοποιούν τεχνικές μηχανικής μάθησης, τόσο στον χρόνο που χρειάζεται η εκπαίδευση του μοντέλου, όσο και στα αποτελέσματα που επιτυγχάνονται.

Λέξεις-κλειδιά: Αμφιωτικές παράμετροι, Νευρωνικά Δίκτυα, Εκτίμηση DOA, Συμπίεση Δεδομένων.

Ευχαριστίες

Έχοντας πλέον ολοκληρώσει τη διπλωματική μου, θα ήθελα να ευχαριστήσω όλα τα μέλη του Audiogroup για το εξαιρετικό κλίμα συνεργασίας στο εργαστήριο τον τελευταίο χρόνο μου ως προπτυχιακός φοιτητής.

Ιδιαιτέρως θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Ιωάννη Μουρτζόπουλο, ο οποίος με καθοδήγησε σε κάθε βήμα της εργασίας, δίνοντάς μου κίνητρα να εργάζομαι πάντα για το καλύτερο δυνατό αποτέλεσμα, για όλες τις ευκαιρίες που μου έδωσε να μάθω νέα πράγματα και να εξελιχθώ, και για την εμπιστοσύνη που μου έδειξε καθ' όλη την πορεία της διπλωματικής.

Ευχαριστώ στη συνέχεια τον Γαβριήλ για την προθυμία του να με βοηθήσει σε κάθε πρόβλημα που συναντούσα καθώς και τον Κωνσταντίνο για την καθοδήγησή του.

Ένα μεγάλο ευχαριστώ στην οικογένειά μου, που με στήριξε στις επιλογές μου, σε όλη την πορεία μου ως φοιτητής και σε κάθε εύκολη ή δύσκολη στιγμή.

Τέλος, θέλω να ευχαριστήσω όλους τους κοντινούς ανθρώπους που μου στάθηκαν, ιδιαίτερα τον Μηνά και τον Βασίλη, για όλες τις συζητήσεις και τις όμορφες αναμνήσεις.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Τα προβλήματα εντοπισμού γωνίας άφιξης (Direction of Arrival - DOA) βασισμένα σε παρατηρήσεις που προκύπτουν από το σήμα που έχει εκπεμφθεί, είναι μια ιδιαίτερα σημαντική λειτουργία της χωρικής ακοής των ανθρώπων. Συνήθως αντιμετωπίζονται με μικροφωνικές διατάξεις, αλλά σε πολλές περιπτώσεις, μόνο δύο μικρόφωνα είναι διαθέσιμα (λ.χ. τοποθετημένα σε ένα τεχνητό ή πραγματικό κεφάλι), το οποίο καθιστά την εκτίμηση της DOA σημαντικά δυσκολότερο πρόβλημα, ιδιαίτερα σε περιβάλλοντα με έντονη αντήχηση, παρόλο που ένας άνθρωπος προσαρμόζεται γρήγορα σε τέτοιες αντίξοες συνθήκες, εντοπίζοντας τους ήχους με έναν εύρωστο τρόπο. Η εκτίμηση της θέσης μιας ακουστικής πηγής, μέσω ενός εκπεμπόμενου ηχητικού σήματος, έχει ευρύ φάσμα εφαρμογών, όπως επόμενης γενιάς εμφυτεύματα κοχλία [1, 2], ακουστικά βοηθήματα [3], ρομπότ [4], συστήματα κατ' οίκον φροντίδας [5] και αναγνώρισης ομιλίας [6]. Ήδη υπάρχουσες στρατηγικές εντοπισμού πηγών, μπορούν να κατηγοριοποιηθούν με έναν ελαστικό τρόπο σε τρεις γενικές κατηγορίες: αυτές που βασίζονται στη μεγιστοποίηση της οδηγούμενης απόκρισης ισχύος (Steered

Response Power - SRP) ενός beamformer [7], τεχνικές που υιοθετούν έννοιες υψηλής φασματικής ανάλυσης [8, 9] και προσεγγίσεις που αξιοποιούν την πληροφορία που προέρχεται από τις διαφορές χρόνων άφιξης (Time Difference of Arrival - TDOA) σε ζεύγη μικροφώνων [10].

Οι περισσότερες μέθοδοι εκτίμησης DOA βασίζονται σε προσεγγίσεις που εντάσονται στην $3^{η}$ κατηγορία, δηλαδή στην αξιοποίηση του TDOA, με την μέθοδο Generalized Cross-Correlation PHAse Transform (GCC-PHAT) να είναι η επικρατέστερη [11]. Μια σύνοψη των τεχνικών TDOA βρίσκεται στο [10]. Στην πράξη, όλες αυτές οι προσεγγίσεις μπορεί να περιορίζονται από ένα, ή συνδυασμό κάποιων μειονεκτημάτων: υψηλό υπολογιστικό κόστος, μη ρεαλιστικές παραδοχές για τα σήματα ή/και τον θόρυβο, αναξιόπιστη απόδοση σε πραγματικά περιβάλλοντα, ειδικά σε δωμάτια με έντονη αντήχηση. Η αυξημένη διαθεσιμότητα υπολογιστικής δύναμης, έχει επιτρέψει την εμφάνιση νέων μεθόδων, που χρησιμοποιούν αλγορίθμους μηχανικής μάθησης (Machine Learning - ML) που μπορούν να αντιμετωπίσουν το πρόβλημα της εκτίμησης DOA.

Τα συστήματα μηχανικής μάθησης που σχεδιάζονται για αυτό το σκοπό, χρησιμοποιούν διαφορετικά δεδομένα εκπαίδευσης, αρχιτεκτονικές, περιβάλλοντα δοκιμής και μετρικές αξιολόγησης. Η πιο ευρέως διαδεδομένη πλέον μέθοδος, είναι η επιβλεπόμενη μάθηση ή supervised machine learning, η οποία έχει ως στόχο, την εκμάθηση μιας συνάρτησης, ενώ ταυτόχρονα την βελτιστοποιεί, η οποία αντιστοιχίζει μια είσοδο σε μια έξοδο, βασιζόμενη σε παραδείγματα εισόδου-εξόδου που χρησιμοποιούνται κατά την εκπαίδευση. Η συνάρτηση εκτιμάται από labeled δεδομένα εκπαίδευσης, που αποτελούνται από ένα σύνολο παραδειγμάτων εκπαίδευσης [12, 13]. Πολλές έρευνες στις οποίες εφαρμόζονται τέτοιες τεχνικές ταξινόμησης DOA και αλγόριθμοι εντοπισμού, έχουν δείξει πως αυτή η προσέγγιση είναι έγκυρη και παρέχει εξαιρετικά αποτελέσματα, πχ χρησιμοποιώντας την στρατηγική random forest ensemble [14, 15, 16] καθώς και τεχνικές που αξιοποιούν νευρωνικά δίκτυα [17, 18, 19].

Σε αυτή την εργασία γίνεται προσπάθεια να αναπτυχθεί μια νέα προσέγγιση για τη μείωση δεδομένων που προέρχονται από αμφιωτικές παραμέτρους, με κατάλληλο τρόπο για εκτίμηση DOA με regression νευρωνικά δίκτυα. Οι εν λόγω αμφιωτικές παράμετροι, είναι η αμφιωτική διαφορά χρόνου άφιξης και η αμφιωτική διαφορά στάθμης ακουστικής πίεσης (Interaural Time Difference και Interaural Level Difference ITD και ILD αντίστοιχα). Συγκεκριμένα, οι αμφιωτικές παράμετροι προέρχονται από μετρήσεις Binaural Room Impulse Responses (BRIRs) από πραγματικά δωμάτια, όπου τα αντηχητικά αποτελέσματα μεγαλώνουν σημαντικά το μέγεθος του dataset ενώ ταυτόχρονα δεν υπάρχουν αρκετές βάσεις δεδομένων που να πληρούν τις απαιτήσεις που υπήρχαν ως προς την αξιοπιστία, το πλήθος των δεδομένων και την ομοιογένειά τους. Για τον σκοπό αυτό, εκτελέστηκε μια εκτενής συγκριτική μελέτη, κατά την οποία αξιολογήθηκε η απόδοση των νευρωνικών όταν χρησιμοποιήθηκαν οι προτεινόμενες συμπιεσμένες αμφιωτικές παράμετροι και οι ασυμπίεστες για τον ίδιο σκοπό, την εκτίμηση της γωνίας άφιξης. Δοκιμάστηκαν δύο εναλλακτικές αρχιτεκτονικές νευρωνικών: μία πλήρως διασυνδεδεμένη (FC), ουσιαστικά ένας Multilayer Perceptron, και μία συνελικτική (CNN).

Εδώ, οι μέθοδοι που δοκιμάστηκαν, επιχειρούν να προβλέψουν την συνάρτηση από τις μεταβλητές εισόδου, σε συνεχείς πραγματικές μεταβλητές δηλαδή την γωνία άφιξης, από τις αμφιωτικές παραμέτρους που χρησιμοποιούνται ως παραδείγματα εκπαίδευσης. Οι μετρικές που χρησιμοποιήθηκαν για να αξιολογηθεί η απόδοση του εκάστοτε μοντέλου είναι το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error - MSE), το οποίο χρησιμοποιήθηκε ως εκτιμήτρια συνάρτηση (Objective Function) κατά την εκπαίδευση, η ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error - RMSE) και το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE) που παρέχουν μια εκτίμηση του πόσο 'μακριά' είναι η προβλεπόμενη DOA, από την πραγματική.

Όπως έχει αναφερθεί, η μέθοδος συμπίεσης δεδομένων που προτεί-

νεται εδώ, μειώνει δραματικά το πλήθος των διαστάσεων των αμφιωτικών παραμέτρων, επιτυγχάνοντας έναν εντυπωσιακό λόγο συμπίεσης έως και 97% καθώς και εξαιρετικά αποτελέσματα εκπαίδευσης των Νευρωνικών Δικτύων. Κατ' επέκταση αυτό σημαίνει πως δεν χρειάζεται να γίνει κανένας συμβιβασμός μεταξύ της ταχύτητας σύγκλισης του μοντέλου και της ακρίβειάς του.

Το μεγαλύτερο μέρος της επεξεργασίας των δεδομένων έγινε σε περιβάλλον MATLAB, χρησιμοποιώντας την εργαλειοθήκη Auditory Modeling Toolbox [20], ενώ η προεπεξεργασία προτού τα δεδομένα χρησιμοποιηθούν για την εκπαίδευση του Νευρωνικού επεξεργάστηκαν με τη γλώσσα Python, όπου έγινε και η εκπαίδευση χρησιμοποιώντας τη διεπαφή TensorFlow [21].

ΚΕΦΑΛΑΙΟ 2

Θεωρία - Μέθοδος

Στο κεφάλαιο αυτό αναφέρονται οι θεωρητικές έννοιες πάνω στις οποίες στηρίζεται η συγκεκριμένη εργασία. Αρχικά αναλύεται η διαδικασία της στερεοφωνικής και αμφιωτικής ακρόασης, η συνάρτηση μεταφοράς του κεφαλιού καθώς και η έννοια της αμφιωτικής κρουστικής απόκρισης δωματίου. Στη συνέχεια περιγράφονται οι διωτικές παράμετροι, ILD και ITD, καθώς και το μοντέλο του Dietz [22] το οποίο εφαρμόστηκε για την εξαγωγή τους και η έννοια του λευκού θορύβου. Στη συνέχεια προσεγγίζονται οι απαραίτητες έννοιες της μηχανικής μάθησης όπως και οι δομές των νευρωνικών που εκπαιδεύτηκαν, τα στοιχεία από τα οποία αυτές αποτελούνται, οι μετρικές αξιολόγησής τους και η σημασία της επιλογής ενός κατάλληλου βελτιστοποιητή (optimizer) στην ταχύτητα σύγκλισης του εκάστοτε μοντέλου.

2.1 Στερεοφωνική και Αμφιωτική ακρόαση

Σε αυτή την ενότητα δίνονται οι βασικές αρχές που διέπουν την αντιληπτική λειτουργία εντοπισμού της θέσης ακουστικών πηγών στο χώρο και οι σχετικές τεχνολογικές προσεγγίσεις που ακολουθούνται από τα ηχητικά συστήματα.

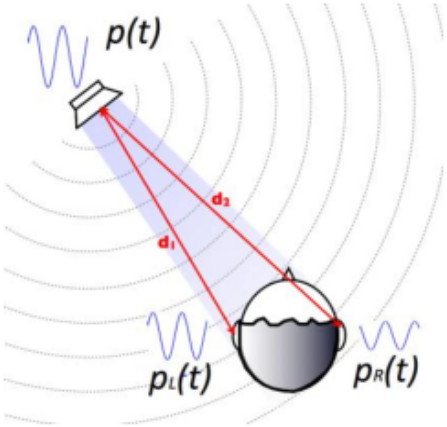
2.1.1 Στερεοφωνική ακρόαση

Ο πλέον διαδεδομένος τρόπος αναπαραγωγής του ήχου για οικιακή ή ατομική ακρόαση είναι η στερεοφωνία που βασίζεται στη χρήση δύο ανεξάρτητων καναλιών (αριστερού-L και δεξιού-R) στα οποία συνδυάζονται πολλαπλά κανάλια και πηγές κατά την ηχογράφηση και τα οποία αναπαράγονται από δύο ηχεία, κατάλληλα τοποθετημένα στο χώρο και σε σχέση με τον ακροατή ή και από ακουστικά κυρίως για ακρόαση από φορητές συσκευές.

Σε ακρόαση με φυσικό τρόπο, η αντίληψη που δημιουργείται από την ύπαρξη μιας ακουστικής πηγής στο ελεύθερο ακουστικό πεδίο ή και χώρο, οφείλεται στα συνδυασμένα ερεθίσματα από τα δύο αυτιά του ακροατή που επιτρέπουν τον προσδιορισμό της θέσης της πηγής. Η λειτουργία του μηχανισμού χωρικής αντίληψης του ήχου στηρίζεται εν πολλοίς στην αποκαλούμενη ***δυϊκή θεωρία*** (duplex theory). Σύμφωνα με τη θεωρία αυτή, το κάθε αυτί, λόγω της διαφορετικής απόστασής του από την ηχητική πηγή d_1 και d_2 όπως φαίνεται στο Σχήμα 2.1, λαμβάνει διαφορετικές τιμές ηχητικής πίεσης $p_L(t)$ και $p_R(t)$ λόγω:

1. Της εξασθένησης της τιμής της πίεσης συναρτήσει της απόστασης.
2. Της διαφοράς φάσης ή και σχετικής καθυστέρησης λόγω του διαφορετικού χρόνου άφιξης του ήχου σε κάθε αυτί.
3. Της πρόσθετης εξασθένησης που δημιουργείται στην ακουστική πίεση στο αυτί που 'καλύπτεται' ακουστικά από το κεφάλι, που δημιουργεί σε περιοχή των συχνοτήτων φαινόμενα ηχητικής σκίασης.

Αναλόγως με τη συχνότητα, το ποιος από τους παραπάνω λόγους είναι σημαντικότερος για την εξασθένιση και αναλύεται περαιτέρω στην ενότητα 2.4, αλλά γενικότερα οι παραπάνω μηχανισμοί λειτουργούν επί το πλείστον συμπληρωματικά.



Σχήμα 2.1: Απεικόνιση του μηχανισμού στερεοφωνικής ακρόασης

Οι ακροατές εντοπίζουν και διαφοροποιούν με ακρίβεια πηγές που δεν εμφανίζουν σχετικές διαφορές ηχοστάθμης ή/και φάσης στα δύο αυτιά (πχ μια πηγή που είναι ακριβώς μπροστά από το παρατηρητή και μια που είναι ακριβώς από πίσω του στην ίδια απόσταση), στην αντίληψη αξιοποιούνται επιπλέον φαινόμενα ανακλάσεων από τον άνω κορμό (ώμος, στήθος κλπ.) καθώς επίσης και την επίδραση του εξωτερικού πτερυγίου του αυτιού που συλλέγει και διαφοροποιεί τα ηχητικά κύματα που φτάνουν σε αυτό ανάλογα με τις διαφορετικές γωνίες πρόσπτωσης και στα δύο επίπεδα.

2.1.2 Αμφιωτική ακρόαση

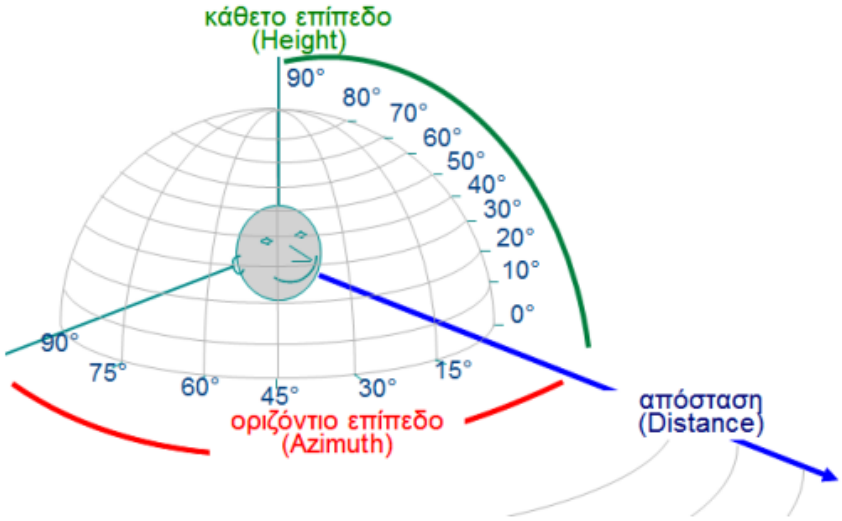
Ο άνθρωπος ως ακουστικός δέκτης παρουσιάζει εξαιρετικές ικανότητες στην αναγνώριση και εντοπισμό ηχητικών πηγών στον τρισδιάστατο χώρο. Οι βασικές αρχές αντίληψης της θέσης των πηγών στον τρισδιάστατο χώρο που προαναφέρθηκαν και αφορούν τις σχετικές στάθμες, χρόνους άφιξης ενός σήματος στα 2 αυτιά, της σκίασης του κεφαλιού, της ως προς

την γωνία φιλτραρίσματος του σήματος από το πτερύγιο του αυτιού και από τις ανακλάσεις στον άνω κορμό, μπορούν να γενικευθούν σε ένα πλαίσιο που συμπεριλαμβάνει όλους αυτούς τους καθοριστικούς μηχανισμούς και μπορούν να περιγράψουν την αμφιωτική ακρόαση.

Κατά τη γενική περίπτωση αμφιωτικής ακρόασης μιας ακουστικής πηγής, ο ήχος που φτάνει στα δύο αυτιά του ακροατή έχει σαν αποτέλεσμα την αντίληψη και τον προσδιορισμό της θέσης της πηγής σε κάποια γωνία τόσο στο οριζόντιο, όσο και στο κάθετο επίπεδο καθώς και την απόσταση που βρίσκεται αυτή (Σχήμα 2.2). Η ικανότητα αυτή προκύπτει από την αντιληπτική διαδικασία που αξιοποιεί τη διαφοροποίηση των σημάτων που φτάνουν στα δύο αυτιά και είναι χαρακτηριστική τόσο για κάθε γωνία στο κάθετο, όσο και στο οριζόντιο επίπεδο. Η δυνατότητα εντοπισμού, εξαρτάται από την μορφολογία του πτερυγίου, του κεφαλιού και του άνω κορμού του εκάστοτε ακροατή. Η ευκρίνεια προσδιορισμού στο οριζόντιο επίπεδο είναι της τάξης των 5^o ενώ στο κάθετο επίπεδο είναι αρκετά χειρότερη.

Για την ανάλυση και την περιγραφή της αντιληπτικής διαδικασίας μέσω της δυϊκής θεωρίας, αξιοποιούνται η Αμφιωτική Διαφορά Στάθμης (ILD) και η Αμφιωτική Διαφορά Χρόνου (ITD) που αναλύονται στην ενότητα 2.4.

Οι δύο αυτές παράμετροι όμως δεν αρκούν για τον προσδιορισμό της θέσης μιας πηγής. Πρέπει να ληφθεί υπόψιν και η διαμόρφωση που εισάγει το εξωτερικό αυτί μέσω του σχήματος του πτερυγίου. Το προσπίπτον σήμα στο κάθε αυτί διαφοροποιείται και φασματικά, αναλόγως με τη γωνία πρόσπτωσης, λόγω της μορφολογίας του πτερυγίου του αυτιού που είναι χαρακτηριστική για κάθε άνθρωπο. Ο τρόπος που το εξωτερικό αυτί συνδυάζεται (μοντελοποιείται) μαζί με τις αμφιωτικές παραμέτρους που αναφέρθηκαν, αναλύονται στην ενότητα 2.2



Σχήμα 2.2: Αντίληψη και εντοπισμός θέσης πηγής μέσω αμφιωτικής ακρόασης

2.2 Συνάρτηση Μεταφοράς Κεφαλιού - HRTF

Η συνάρτηση μεταφοράς του κεφαλιού (Head Related Trasnfer Function - HRTF) είναι μια συνάρτηση μεταφοράς, που για μια συγκεκριμένη γωνία πρόσπτωσης, περιγράφει τη μετάδοση του ήχου από ένα ελεύθερο πεδίο (free field) και την επίδραση του κεφαλιού, του λοβού και του θώρακα σε ένα σημείο στο κανάλι του αυτιού (Σχήμα 2.6) ενός ανθρώπου [23]. Στο πεδίο του χρόνου η HRTF λέγεται κρουστική απόκριση κεφαλιού (Head Related Impulse Response - HRIR). Η HRIR Παρουσιάζει με έναν εύχρηστο και συνοπτικό τρόπο το ILD και ITD μεταξύ των αυτιών καθώς και άλλες φασματικές πληροφορίες [24].

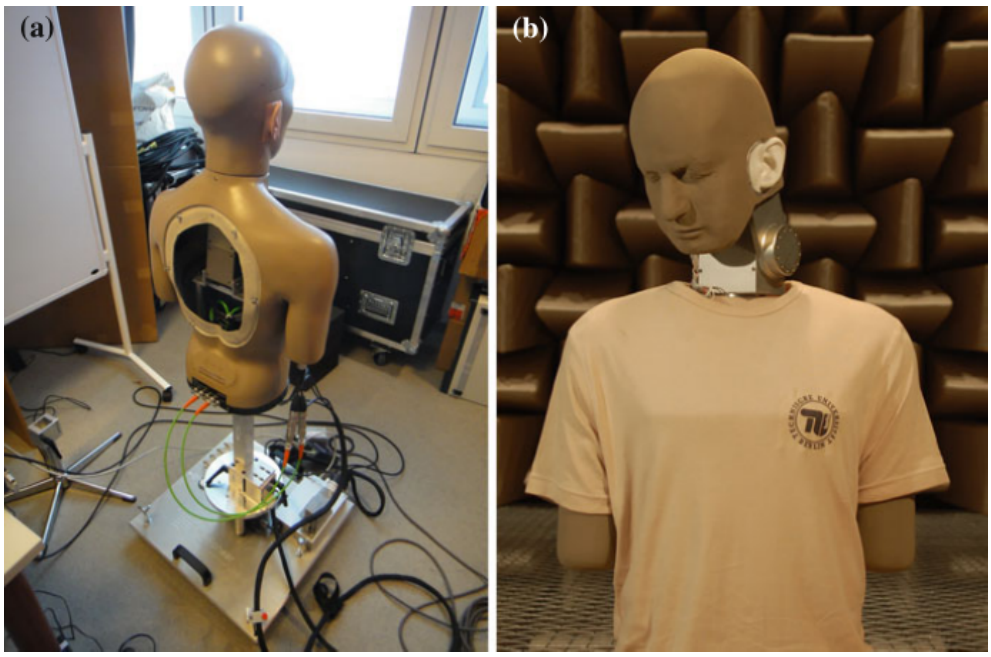
Στην [25], η συνολική ακουστική μεταφορά από μία ακουστική πηγή σε ελεύθερο χώρο έως το τύμπανο ενός ακροατή, χωρίζεται σε τρία κομμάτια ως εξής:

- Μεταφορά από ελεύθερο πεδίο στην φραγμένη είσοδο του ακουστικού καναλιού
- Μετατροπή της αντίστασης που σχετίζεται με την φραγή του ακουστικού καναλιού
- Μεταφορά μέσω του ακουστικού καναλιού

2.2.1 Μέτρηση HRTF

Ένα σετ HRTF αριστερού και δεξιού αυτιού εξαρτάται από την θέση της πηγής, την θέση του ακροατή, τη θέση του κεφαλιού και του θώρακα και άλλες παραμέτρους. Οι τυπικές διατάξεις μέτρησης, επιτρέπουν τη μεταβολή ενός ή δύο βαθμών ελευθερίας. Στις περισσότερες περιπτώσεις, μεταβάλλεται η γωνία της πηγής, σε σχέση με το κεφάλι και τον θώρακα που έχουν σταθερό προσανατολισμό. Για την αποτελεσματικότερη μέτρηση HRIR και BRIR (βλ. 2.3) έχουν κατασκευαστεί βοηθητικά μοτέρ που επιτρέπουν την περιστροφή ή την αλλαγή της κλίσης του κεφαλιού μέσω λογισμικού, με ακρίβεια μέχρι και 0.01^o. Συνήθως γίνονται

μετρήσεις στο διάστημα $\pm 90^o$, με τις 0 να είναι η πηγή ακριβώς μπροστά από τον δέκτη. Στο Σχήμα 2.3 φαίνονται δύο τυπικές διατάξεις μέτρησης HRTF.



Σχήμα 2.3: Τυπικά σενάρια μέτρησης HRTF: Τροποποιημένο ανδρείκελο KEMAR (a), FABIAN (b)

2.2.2 Υπολογισμός HRTF

Η επιλογή του σήματος διέγερσης που εκπέμπει η πηγή διαδραματίζει πολύ μεγάλο ρόλο στα αποτελέσματα. Η βέλτιστη διέγερση, εξαρτάται κυρίως από τον αλγόριθμο υπολογισμού της HRIR, αλλά ταυτόχρονα και από το σενάριο μέτρησης (στατικό ή δυναμικό, το περιβάλλον, το υλικό που χρησιμοποιείται κλπ.). Για την επίτευξη ενός υψηλού SNR, το σήμα διέγερσης πρέπει να έχει υψηλή ενέργεια, συγκριτικά με το σύστημα μέτρησης, σε όλη τη συχνοτική περιοχή ενδιαφέροντος. Συχνότερα χρησιμοποιούνται Maximum Length Sequences (MLS), τα οποία είναι περιοδικές ακολουθίες bit που δημιουργούνται από καταχωρητές μετατόπισης γραμμικής ανατροφοδότησης (linear-feedback shift registers) ή

απεριοδικά sweep.

Στη συνέχεια, το σήμα αναπαράγεται από το ηχείο-πηγή, και γίνεται συγχρονισμένα η καταγραφή της απόκρισης του αριστερού και δεξιού καναλιού. Η απόκριση συχνότητας προκύπτει από γραμμική αποσυνέλιξη, συνήθως στο πεδίο της συχνότητας. Ένα αποδεκτό SNR για τέτοιου είδους μετρήσεις είναι κάπου ανάμεσα στα 60-90 dB. Μετρήσεις HRTF, πλέον γίνονται και σε ανθρώπους, όχι μόνο σε ανδρείκελα, με ειδικά κατασκευασμένα ακουστικά το περίβλημα των οποίων τυπώνεται με 3Δ εκτυπωτές.

2.2.3 Βάσεις δεδομένων HRTF

Οι HRTF είναι ένα αναπόσπαστο κομμάτι της τεχνολογίας αμφιωτικής ακοής και ακρόασης. Από την άλλη, η μέτρηση των HRTF είναι ένα δύσκολο και ευαίσθητο εργαστηριακό αντικείμενο. Συνήθως απαιτείται ένας ανηχοϊκός θάλαμος για τη μέτρηση των HRTF, και επιπλέον αρκετός χρόνος από την πλευρά του χειριστή του πειράματος αλλά και του υποκειμένου μέτρησης, ώστε να ολοκληρωθούν οι μετρήσεις με ικανοποιητικό βαθμό χωρικής ανάλυσης. Απαιτούνται ένα ή περισσότερα ηχεία, ακουστικά in-ear, και λογισμικό αναπαραγωγής και καταγραφής ήχου.

Υπάρχουν αρκετά τέτοια συστήματα, αλλά είναι σαφές πως δεν είναι φορητά. Ως αποτέλεσμα, πολλοί οργανισμοί έχουν επιλέξει να παρέχουν τις μετρήσεις ως δημόσια διαθέσιμες βάσεις δεδομένων στην κοινότητα. Παρακάτω αναφέρονται κάποιες από αυτές. Εκτός αν αναφέρεται διαφορετικά, οι κρουστικές αποκρίσεις παρέχονται με συχνότητα δειγματοληψίας $F_s = 44.1kHz$.

- KEMAR - βάση δεδομένων HRTF από το MIT-Media-Lab [26]: Η πρώιμη αυτή βάση δεδομένων είναι ακόμα ιδιαίτερα δημοφιλής, χρησιμοποιώντας το Knowles-Electronics Mannequin for Acoustic Research (KEMAR), αναπαριστά μια εκτενή καταγραφή HRTF. Συνολικά, δειγματοληπτούνται 710 διαφορετικές θέσης, για ανύψωση από -40^o μέχρι $+90^o$ με χωρική ανάλυση 10^o στον κάθετο άξονα και περίπου 5^o

στον οριζόντιο, για απόσταση περίπου 1.4 m μεταξύ του ηχείου και του KEMAR.

- AUDIS - Ο κατάλογος AUDIS ανθρωπίνων HRTF: Στο γενικό πλαίσιο της E.E., το project Auditory Displays (AUDIS) [27], το οποίο βασίστηκε σε μεγάλο βαθμό στην αμφιωτική τεχνολογία και αξιόπιστες HRTF ανθρώπων, πραγματοποιήθηκε ένα ειδικό πρόγραμμα συλλογής HRTF. Εδώ οι συνθήκες μετρήσεων είναι: 2.4 m από το ηχείο, ανάλυση 10^o στον κάθετο άξονα από -10^o μέχρι $+90^o$, και ανάλυση 15^o στον οριζόντιο άξονα. Οι συνολικές μετρήσεις περιλαμβάνουν 122 κατευθύνσεις για 20 ανθρώπους. Τέθηκε επίσης ένα σύνολο *Golden Rules* για μετρήσεις HRTF.
- CIPIC - βάση δεδομένων HRTF του εργαστηρίου CIPIC [28]: Η βάση δεδομένων, περιέχει HRTF μετρημένες με υψηλή χωρική ανάλυση, για περισσότερους από 90 ανθρώπους, με 45 αυτών να είναι δημόσια διαθέσιμοι, συμπεριλαμβανομένου του KEMAR, με μεγάλο και μικρό λοβό. Η χωρική ανάλυση είναι 5^o στον κάθετο αλλά και στον οριζόντιο άξονα. Προκύπτουν έτσι 1250 σημεία στην ακουστική σφαίρα 1m, όπου ήταν η απόσταση του ηχείου. Παρέχονται επίσης ανθρωπομετρικά χαρακτηριστικά για κάθε υποκείμενο μέτρησης, καθώς και βοηθητικές συναρτήσεις για το περιβάλλον MATLAB.
- LISTEN - η βάση δεδομένων HRTF IRCAM: Αναπτηγμένη σε πρότζεκτ της E.E., περιέχει μετρήσεις HRTF με για ανύψωση από -45^o μέχρι $+90^o$ με χωρική ανάλυση 5^o και περίπου 15^o ανάλυση στον οριζόντιο άξονα. Συνολικά μετριούνται 187 θέσεις. Παρέχονται οι μετρήσεις HRTF, προαιρετική διόρθωση diffuse-field και μορφολογικά δεδομένα. Τα δεδομένα που μετρήθηκαν για 50 ανθρώπους είναι διαθέσιμα online[1].
- ARI - Βάση δεδομένων του Acoustics-Research-Institute: Περιέχει HRTF υψηλής ανάλυσης για πάνω από 70 ανθρώπους. Μετρήθηκαν 1550

[1] http://recherche.ircam.fr/equipes/salles/listen/index.html

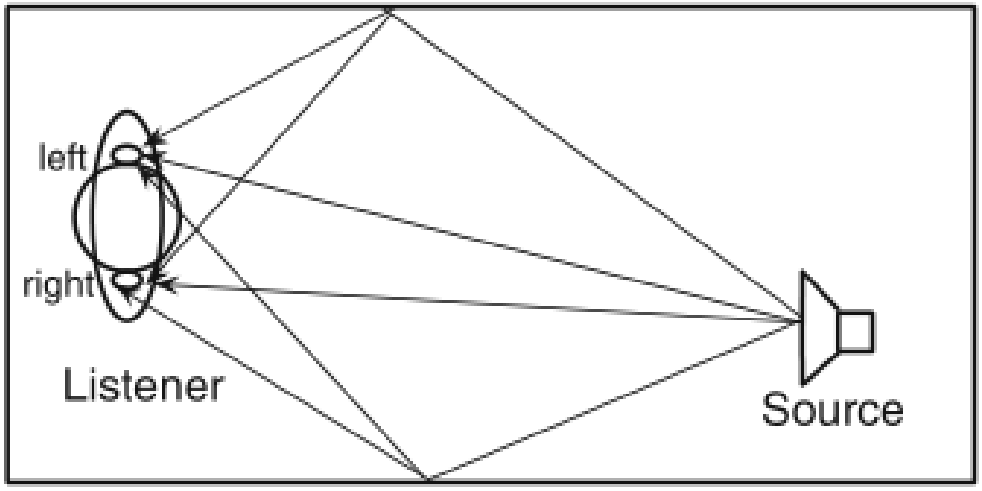
θέσεις για κάθε ακροατή, με 2.5^o ανάλυση στον οριζόντιο άξονα για γωνίες από $0^o - 359^o$, και ανυψώσεις από -30^o έως $+80^o$. Τα δεδομένα μπορούν να βρεθούν στον σύνδεσμο[2].

- FIU - Βάση δεδομένων HRTF του Florida-International-Univ. DSP-Lab: Μετρήθηκαν 15 διαφορετικά άτομα, για δώδεκα διαφορετικές γωνίες στον οριζόντιο άξονα και έξι στον κάθετο. Παρέχονται 3-Δ εικόνες των λοβών κάθε ατόμου, και ανθρωπομετρικά χαρακτηριστικά. Χρησιμοποιήθηκε συχνότητα δειγματοληψίας $F_s = 96kHz$ και είναι διαθέσιμη στον σύνδεσμο[3].

[2]http://www.kfs.oeaw.ac.at/content/view/608/606
[3]http://dsp.eng.fiu.edu/HRTFDB

2.3 Αμφιωτική Κρουστική Απόκριση Δωματίου - BRIR

Όταν ένας ήχος εκπέμπεται από μια πηγή σε έναν κλειστό χώρο, ένας ακροατής αρχικά θα λάβει τον άμεσο ήχο, ακολουθούμενο από ανακλάσεις από τους τοίχους ή αντικείμενα τοποθετημένα μέσα στο δωμάτιο, όπως φαίνεται στο Σχήμα 2.4.



Σχήμα 2.4: Ακροατής και ακουστική πηγή σε δωμάτιο με αντήχηση

Η ενέργεια του ανακλώμενου ήχου εξασθενεί, σύμφωνα με τα χαρακτηριστικά απορρόφησης των εκάστοτε επιφανειών στο δωμάτιο. Υποθέτοντας ότι η ακουστική του δωματίου, μοντελοποιείται ως ένα *γραμμικό*, *χρονικά-αμετάβλητο* (ΓΧΑ), σύστημα, η κρουστική απόκριση του δωματίου, (Room Impulse Response - RIR), παρέχει μια πλήρη περιγραφή των άμεσων και ανακλώμενων μονοπατιών από μια πηγή στον δέκτη. Ο χρόνος που χρειάζεται για να ελαττωθεί η ενέργεια του δωματίου κατά 60 dB, αφού η πηγή έχει σταματήσει να εκπέμπει καλείται χρόνος αντήχησης RT_{60} και είναι η παράμετρος που χρησιμοποιείται πιο συχνά για τον προσδιορισμό των ακουστικών ιδιοτήτων ενός δωματίου [29]. Σε ένα γενικό, πολυκαναλικό σενάριο, με μία πηγή και i δέκτες, το αντιχητικό σήμα $x_i(n)$, μπορεί να εκφραστεί σαν την συνέλιξη του ανηχοϊκού σήματος $s(n)$, με τις αντίστοιχες RIR, $h_i(n)$, ως εξής:

$$x_i(n) = \sum_{j=0}^{J_h-1} h_i(j)s(n-j) \tag{1}$$

Όπου n αναπαριστά τον δείκτη διακριτού χρόνου και J_h το μήκος της κρουστικής απόκρισης.

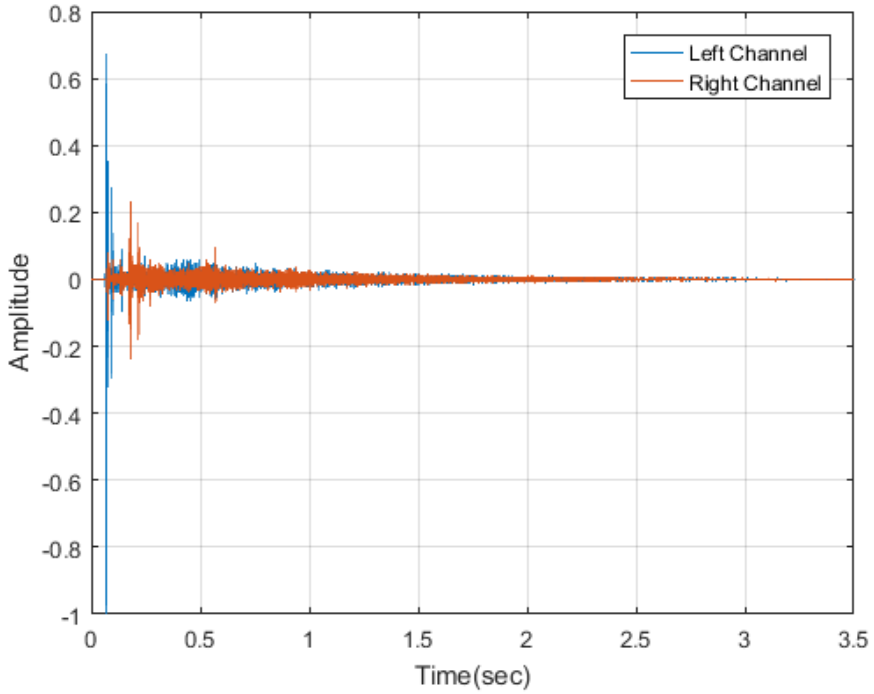
Σε ένα αμφιωτικό σενάριο (binaural) η απόκριση του δωματίου, συνδυάζεται με την κρουστική απόκριση του κεφαλιού (Head Related Impulse Response - HRIR), η οποία αποτελείται από δύο κανάλια, ένα για κάθε αυτί. Οι HRIR μετρούνται σε ανηχοϊκές συνθήκες. Συνεπώς, υποθέτοντας μια ιδεατή παντοκατευθυντική πηγή, μία αμφιωτική κρουστική απόκριση δωματίου, για το αριστερό αυτί-κανάλι, $h_L(n)$, μπορεί να εκφραστεί ως

$$\begin{aligned} h_L(n) = {} & g(r_s)(n-n_s) * h_{HRIR,L,\theta_d,\phi_d}(n) \\ & + \sum_{m=0}^{J_{h_m}-1} h_{m,L}(n) * h_{HRIR,L,\theta_m,\phi_m}(n) \end{aligned} \tag{2}$$

όπου $g(r_s)$ είναι η ελάτωση του κέρδους που εξαρτάται από την απόσταση πηγής-παρατηρητή r_s, $\delta(n)$ η συνάρτηση δέλτα του Kronecker, n_s η καθυστέρηση που εξαρτάται κυρίως από την απόσταση πηγής-παρατηρητή και τα φυσικά χαρακτηριστικά του μέσου διέλευσης. $h_{HRIR,L,\theta_m,\phi_m}(n)$ είναι η αριστερή HRIR για τον άμεσο ήχο, που αντιστοιχεί σε θ_d και ϕ_d, δηλαδή την οριζόντια και κάθετη γωνία μεταξύ του παρατηρητή και της πηγής. Η τιμή $h_m(n)$ αναφέρεται στην m-οστή ανάκλαση. J_{h_m} ο συνολικός αριθμός ανακλάσεων. Τέλος θ_m and ϕ_m είναι οι οριζόντιες και κάθετες γωνίες μεταξύ του δέκτη και της m-οστής ανάκλασης. Αντίστοιχα υπολογίζεται και η BRIR, $h_R(n)$.

Έτσι, το αντηχητικό σήμα στο αριστερό και δεξί αυτί ενός ακροατή $x_L(n)$ και $x_R(n)$, περιγράφονται με τη συνέλιξη του ανηχοϊκού σήματος $s(n)$, με τις BRIR του αριστερού και δεξιού αυτιού αντίστοιχα, όπως φαίνεται στην Εξίσωση 3. Ένα παράδειγμα BRIR από ένα δωμάτιο με ιδιαίτερα υψηλή αντήχηση φαίνεται στο Σχήμα 2.5

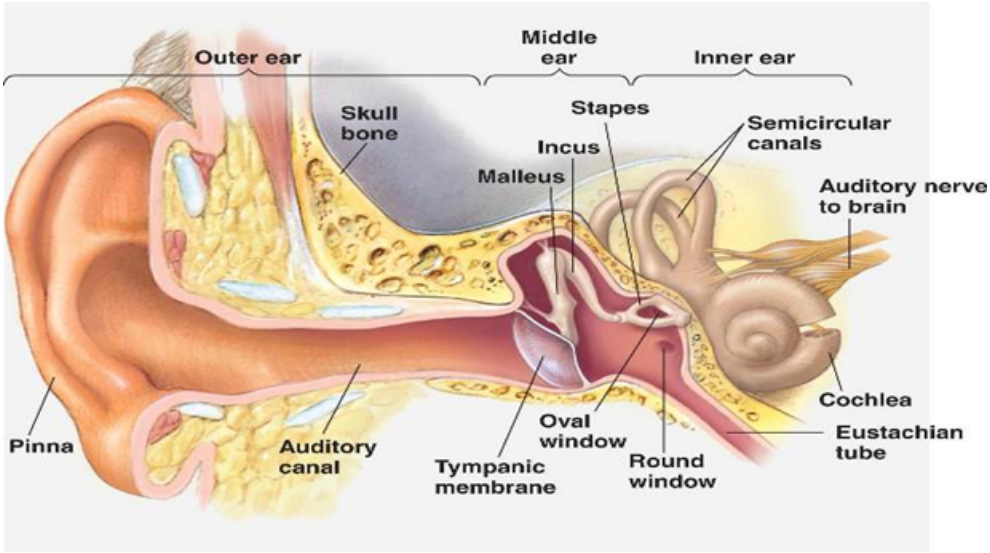
$$x_L(n) = \sum_{j=0}^{J_{h_L}-1} h_L(j)s(n-j)$$
$$x_R(n) = \sum_{j=0}^{J_{h_R}-1} h_R(j)s(n-j) \quad (3)$$



Σχήμα 2.5: BRIR από δωμάτιο με υψηλή αντίχηση

2.4 Αμφιωτικές Παράμετροι

Το ακουστικό σύστημα, μπορεί να παρομοιαστεί με έναν υπολογιστή πολλαπλών χρήσεων με δύο θύρες εισόδου. Οι θύρες είναι τα δύο αυτιά, στο ίδιο ύψος εκατέρωθεν ενός στερεού ελλειψοειδούς, το κεφάλι. Το κεφάλι λειτουργεί ως *φορέας μιας κεραίας*, που μπορεί να κινηθεί με έξι βαθμούς ελευθερίας σε σχέση με το σώμα, ενώ το ίδιο το σώμα μπορεί να προηγηθεί στον τρισδιάστατο χώρο, και να αλλάξει τον προσανατολισμό του σε σχέση με τη θέση αναφοράς [30]. Μία σύντομη ανατομική περιγραφή του ακουστικού συστήματος φαίνεται στο Σχήμα 2.6



Σχήμα 2.6: Ανατομική περιγραφή του ακουστικού συστήματος

Το ακουστικό σύστημα δέχεται εισόδους με τη μορφή ελαστικών δονήσεων και κυμάτων, από υγρά ή στερεά με τα οποία είναι σε μηχανική επαφή. Οι είσοδοι έρχονται είτε μέσω του αέρα από τα ακουστικά κανάλια (ear canals) είτε από την μεταβίβαση των οστών μέσω του κρανίου. Συνήθως το κρανίο αγνοείται όταν πρόκειται για ακρόαση σε αέρα, αφού διαφέρει κατά περίπου 60 dB σε σχέση με την μεταβίβαση λόγω αέρα, που αντιστοιχεί σε έναν λόγο ισχύος της τάξεως του 10^6.

Η ακοή επιτυγχάνεται και με ένα μόνο αυτί, αλλά η ακρόαση με δύο

λειτουργικά αυτιά, ή *αμφιωτική ακρόαση*, προσφέρει σημαντικά πλεονεκτήματα έναντι της μονοωτικής ακρόασης. Αυτό, συμβαίνει διότι η αμφιωτική ακρόαση προσφέρει επιπλέον πληροφορία, η οποία κωδικοποιείται στις διαφορές των σημάτων εισόδου στα δύο αυτιά.

Υποθέτοντας ότι αυτές οι διαφορές αναπαρίστανται με ένα γραμμικό, χρονικά αμετάβλητο σύστημα, προκύπτει το συμπέρασμα, ότι μπορούν να υπάρχουν μόνο δύο τέτοιες διαφορές. Η αμφιωτική διαφορά χρόνου άφιξης (Interaural Time Difference - ITD) και οι αμφιωτικές διαφορές έντασης (Interaural Level Difference - ILD). Και οι δύο εξαρτώνται από τη συχνότητα.

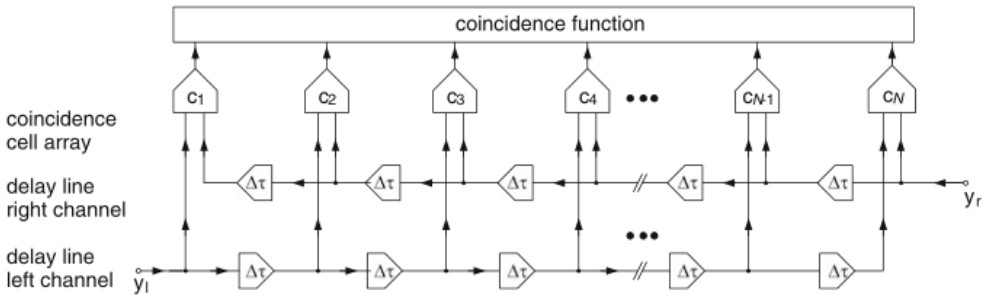
Τα υπάρχοντα μοντέλα, αξιοποιούν τις ακουστικές παραμέτρους, που διακρίνονται σε αμφιωτικές παραμέτρους, που είναι πιο εύρωστες και απαιτούν και τα δύο αυτιά για να αναλυθούν, και σε μονοωτικές παραμέτρους που χρειάζονται μόνο το ένα αυτί. Εδώ αναλύονται μόνο οι αμφιωτικές παράμετροι αφού αυτές χρησιμοποιήθηκαν για την προσέγγιση του προβλήματος.

2.4.1 Interaural Time Difference

Παρουσιάζεται πρώτα η ανάλυση του ITD, γιατί ιστορικά το μοντέλο Jefress [31] ήταν το πρώτο μοντέλο εντοπισμού, το 1948. Η βασική ιδέα αυτού του μοντέλου είναι ένας συνδυασμός, *γραμμών καθυστέρησης* και *κυττάρων σύμπτωσης* (coincidence cells). Με βάση το μοντέλο, υπάρχουν δύο ξεχωριστές, παράλληλες γραμμές καθυστέρησης σε κάθε αυτί. Τα σήματα διαδίδονται με αντίθετη κατεύθυνση σε κάθε γραμμή, όπως φαίνεται στο Σχήμα 2.7. Σε κάποιο σημείο, τα σήματα που ταξιδεύουν στις δύο γραμμές, συναντούνται σε ένα κύτταρο σύμπτωσης, το οποίο στέλνει το σήμα στο επόμενο επίπεδο. Λόγω της διαφοράς χρόνου άφιξης των δύο σημάτων στα αυτιά, αυτά θα ενεργοποιήσουν ένα πλευρικά μετατοπισμένο κύτταρο σύμπτωσης, το οποίο κατ' επέκταση αντιστοιχίζεται σε μια πλευρική γωνία άφιξης.

Οι Cherry και Sayers [32], εισήγαγαν τη χρήση της αμφιωτικής ετεροσυσχέτισης, ως μια μέθοδο για την εκτίμηση του ITD η οποία ορίζεται όπως φαίνεται στην Εξίσωση 4. Με την εσωτερική καθυστέρηση να ορίζεται με τ και τα αριστερά και δεξιά σήματα πίεσης, $y_l(t)$ και $y_r(t)$. Έχει αποδειχτεί πως αυτό είναι μια καλή προσέγγιση του μοντέλου Jeffress. Το ITD έχει παρατηρηθεί πως δεν παίζει ιδιαίτερο ρόλο στον εντοπισμό πηγών σε συχνότητες μεγαλύτερες των 1500Hz.

$$\psi_{y_{l,r}}(\tau) = \frac{\int_{t=-\infty}^{\infty} y_l(t) y_r(t+\tau) dt}{\sqrt{\int_{t=-\infty}^{\infty} y_l^2(t) dt \int_{t=-\infty}^{\infty} y_r^2(t) dt}} \tag{4}$$



Σχήμα 2.7: Μοντέλο σύμπτωσης όπως αρχικά προτάθηκε από τον Jeffress

2.4.2 Interaural Level Difference

Οι αμφιωτικές διαφορές έντασης συμβαίνουν λόγω φαινομένων επικάλυψης του κεφαλιού, όταν ένας ήχος φτάνει πλάγια στον δέκτη. Τυπικές τιμές είναι $\pm 30dB$ σε συχνότητες κοντά στα 5 kHz και γωνίες άφιξης στα $\pm 60^o$. Στις χαμηλές συχνότητες η επικάλυψη του κεφαλιού δεν παίζει ιδιαίτερο ρόλο, και διαφορές στο ILD σπανίζουν (κάτω από 1500 Hz). Η συχνότητα επικάλυψης ILD και ITD γίνεται φανερό πως είναι στα 1500 Hz.

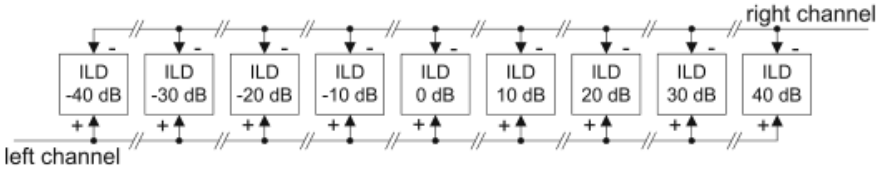
Το ILD, μπορεί να υπολογιστεί απευθείας από τα σήματα του αριστερού και δεξιού καναλιού όπως φαίνεται στην εξίσωση 5, πράγμα το οποίο τυπικά υπολογίζεται για διαφορετικές μπάντες συχνοτήτων, όπου

P_l και P_r οι ισχύες των σημάτων που φτάνουν στο αριστερό και δεξί αυτί.

$$\alpha = 10log_{10}(P_l) - 10log_{10}(P_r) \tag{5}$$

Οι Reed και Blum [33] εισήγαγαν έναν φυσιολογικό (physiological) αλγόριθμο για τον υπολογισμό του ILD, βασισμένο στην δραστηριότητα, $E(\alpha)$, μιας συστοιχίας *El cells*, όπως φαίνεται στο σχήμα 2.8 και περιγράφεται στην Εξίσωση 6. Η απόκριση κάθε El-cell είναι συντονισμένη για ένα συγκεκριμένο ILD, και ελατώνεται όσο πιο μακριά από αυτό είναι το ILD των σημάτων άφιξης.

$$E(\alpha) = \exp\left(10^{\alpha/ILD_{max}}\sqrt{P_l} - 10^{\alpha/ILD_{max}}\sqrt{P_r}\right)^2 \tag{6}$$



Σχήμα 2.8: Δομή El-cell

2.4.3 Εντοπισμός ηχητικών γεγονότων

Υπάρχουν πολλές μέθοδοι για τον υπολογισμό της θέσης ηχητικών πηγών από τις αμφιωτικές παραμέτρους. Μια μέθοδος που επιτυγχάνει αυτόν τον σκοπό, είναι η κατασκευή μια βάσης δεδομένων, που αντιστοιχίζει μετρηθείσες παραμέτρους ILD και ITD σε σφαιρικές συντεταγμένες.

Είναι ακόμα ασαφές πως το ακουστικό σύστημα συνδυάζει τις παραμέτρους για την εξαγωγή συμπερασμάτων για την θέση ακουστικών γεγονότων, συγκεκριμένα, αναπάντητα είναι ακόμα τα εξής ερωτήματα:

- Ο τρόπος που συνδυάζονται τα ILD και ITD.
- Η ολοκλήρωση της πληροφορίας ως προς τον χρόνο.
- Η ολοκλήρωση της πληροφορίας ως προς τη συχνότητα.
- Ο διαχωρισμός διαφορετικών, ταυτόχρονων πηγών.

- Η αντιμετώπιση των ανακλάσεων του δωματίου.

Ότι είναι γνωστό μέχρι στιγμής για τον τρόπο που το ακουστικό σύστημα συνδυάζει τις παραμέτρους έχει προκύψει πειραματικά, από τα αποκαλούμενα trading πειράματα, όπου το ILD υποδεικνύει μια κατεύθυνση, αλλά το ITD μια διαφορετική και ο δέκτης καλείται να κρίνει την πραγματική.

Πιστεύεται ότι το ακουστικό σύστημα, εφαρμόζει χρονικό ή/και συχνοτικό cue weighting, δηλαδή δίνει μεγαλύτερη βαρύτητα σε κάποιες παραμέτρους όταν πληρούνται κάποιες συνθήκες, αλλά ο ακριβής τρόπος που αυτό συμβαίνει είναι ακόμα υπό διερεύνηση. Υπάρχουν διαφορετικές απόψεις, με την μία, που είναι και η επικρατέστερη, να λέει πως το ακουστικό σύστημα δίνει έμφαση στο onset τμήμα του σήματος, ενώ η άλλη να ισχυρίζεται πως ολοκληρώνει την πληροφορία σε μεγαλύτερο χρονικό διάστημα. Τα πράγματα γίνονται σημαντικά δυσκολότερα, όταν δεν είναι σαφές πόσες πηγές υπάρχουν. Τότε οι παράμετροι εκτός από την ανάθεση βαρών σε αυτές, πρέπει να αντιστοιχηθούν και στη σωστή πηγή. Η μεγαλύτερη πρόκληση όμως παραμένει διερεύνηση της αντιμετώπισης των ανακλάσεων του δωματίου από το ακουστικό σύστημα.

2.5 Μοντέλο dietz2011

Οι Dietz et al. κατασκεύασαν ένα μοντέλο της ανθρώπινης ακοής, με στόχο τον εντοπισμό της κατεύθυνσης ταυτόχρονων ομιλητών στο [22]. Ο άνθρωπος έχει μια ιδιαίτερα εύρωστη ικανότητα να εντοπίζει ήχους σε αντίξοες συνθήκες (αντήχηση, έντονος θόρυβος κλπ) οπότε κρίθηκε αναγκαίο να δημιουργηθεί ένα μοντέλο που προσομοιάζει τα χαρακτηριστικά του ανθρώπινου ακουστικού συστήματος με στόχο να χρησιμοποιηθεί αυτό σε αυτόματα συστήματα προσέγγισης DOA.

Η δομή του μοντέλου που κατασκευάστηκε χωρίζεται σε τρία μέρη. Το πρώτο κομμάτι, που θα απασχολήσει αυτή την ενότητα, είναι η ακουστική επεξεργασία για την εξαγωγή των ακουστικών παραμέτρων. Στο δεύτερο, εξάγονται τα σημαντικά τμήματα αυτών των παραμέτρων και στο τρίτο υλοποιείται το σύστημα εντοπισμού.

2.5.1 Ακουστική επεξεργασία

Εδώ χρησιμοποιήθηκε το μοντέλο ακουστικής επεξεργασίας Interaural Phase Difference [34], και περιγράφεται παρακάτω.

1. Προσεγγίζεται η χαρακτηριστική μεταφοράς του μέσου αυτιού (βλ. 2.6), με ένα πρώτης τάξεως ζωνοπερατό φίλτρο $F_p = 500Hz, F_c = 2kHZ$.
2. Το ακουστικό ζωνοπερατό φιλτράρισμα στην basilar μεμβράνη μοντελοποιείται με μια γραμμική, 4ης τάξης, τράπεζα φίλτρων που αποτελούνται μόνο από πόλους (gammatone filter-bank). Το πλάτος κάθε φίλτρου ορίζεται ως το ισοδύναμο τετραγωνικό εύρως ζώνης (Equivalent Rectangular Bandwidth - ERB) των ακουστικών φίλτρων. Υλοποιούνται 23 μπάντες φίλτρων στο διάστημα των 200 Hz - 5 kHz με απόσταση 1 ERB. Σημειώνεται πως το ERB είναι ένα μέτρο που χρησιμοποιείται στην ψυχοακουστική και δίνει μια εκτίμηση του εύρους ζώνης των φίλτρων στην ανθρώπινη ακοή, χρησιμοποιώντας μη

ρεαλιστικά, αλλά βολικά μοντέλα τετραγωνικών ζωνοπερατών φίλτρων.

3. Η συμπίεση του κοχλία λογιστικοποιήθηκε με άμεση συμπίεση ισχύος 0.4, μετά το bandpass φιλτράρισμα.
4. Η διαδικασία της ηλεκτρομηχανικής μεταγωγής των εσωτερικών haircells, προσομοιώνεται με την ανόρθωση ημίσεος κύματος με διαδοχικά lowpass φίλτρα 5ης τάξης με $F_c = 770Hz$.
5. Οι αμφιωτικές χρονικές ανομοιότητες προκύπτουν από το ζωνοπερατό φιλτράρισμα με μιγαδικά gammatone φίλτρα 2ης τάξης.

Η μιγαδική έξοδος των φίλτρων (Εξίσωση 7) περιέχει τη διαχωρίσιμη πληροφορία του πλάτους $\alpha(t)$ και της φάσης του σήματος $\phi(t)$.

$$g(t) = \alpha(t)e^{i\phi(t)} \tag{7}$$

Από τα αντίστοιχα αριστερά-δεξιά ζεύγη εξόδων των φίλτρων, g_l και g_r, υπολογίζεται η αμφιωτική συνάρτηση μεταφοράς (Interaural Transfer Function - ITF) όπως φαίνεται στην Εξίσωση 8.

$$ITF(t) = g_l(t)\overline{g_r}(t) = \alpha_l(t)\alpha_r(t)e^{i(\phi_l(t)-\phi_r(t))} \tag{8}$$

Η ITF είναι μιγαδική και περιέχει και αυτή, την πληροφορία για τη φάση και για το πλάτος. Είναι άρα ιδανική για τη χρονική εξομάλυνση των αμφιωτικών συναρτήσεων. Η χρονικά εξομαλυμένη IPD εξάγεται στη συνέχεια από την ITF, αφού αυτή φιλτραριστεί από ένα χαμηλοπερατό φίλτρο, όπως φαίνεται στην Εξίσωση 9

$$IPD(t) = arg([ITF(t)]_{lp}) \tag{9}$$

Η IPD μπορεί να μετασχηματιστεί σε ITD διαιρώντας το IPD με την μέση άμεση συχνότητα του αριστερού και δεξιού σήματος. Ο τρόπος υπολογισμού της μέσης άμεσης συχνότητας φαίνεται στην Εξίσωση 10

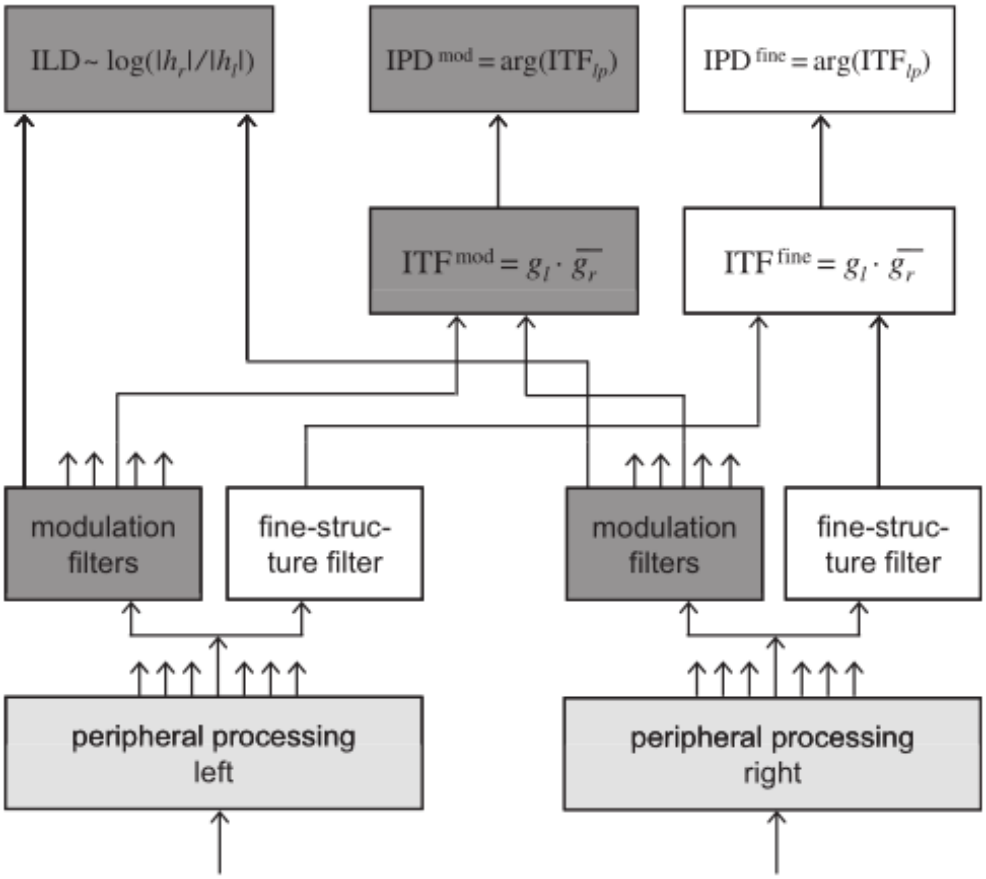
$$f_{inst} = \frac{1}{4\pi}(\frac{d\phi_l(t)}{dt} + \frac{d\phi_r(t)}{dt}) \tag{10}$$

Τα μιγαδικά φίλτρα επιτρέπουν τον υπολογισμό της ITF και κατ' επέκταση την IPD της λεπτής-δομής (fine structure) ή της περιβάλλουσας (envelope) του σήματος. Το πρώτο επιτυγχάνεται με το κεντράρισμα του φίλτρου στην ίδια συχνότητα με το προηγούμενο ακουστικό φίλτρο. Για την περιβάλλουσα, το φίλτρο κεντράρεται, στη συχνότητα διαμόρφωσης ενδιαφέροντος, κατά προτίμηση, στην έξοδο του ακουστικού φίλτρου υψηλής συχνότητας. Το φίλτρο λεπτής δομής έχει Q-value 3, ενώ τα φίλτρα διαμόρφωσης έχουν Q-value 8.

Το ILD εκφράζεται σε dB, και πολλαπλασιάζεται με τον παράγοντα συμπίεσης c, όπου $c \sim 0.4$,όπως φαίνεται στην εξίσωση 11 ώστε να κλιμακωθεί η εσωτερική αναπαράσταση στο προτώτυπο ILD, που συμβαίνει στα αυτιά, πριν τη συμπίεση της basilar μεμβράνης.

$$ILD(t) = 20clog_{10}(\frac{|h_r(t)|}{|h_l(t)|}) \tag{11}$$

Το πλεονέκτημα αυτού του μοντέλου είναι η υψηλή φασματική ανάλυση των αμφιωτικών παραμέτρων, που μπορούν να υπολογιστούν δείγμα προς δείγμα. Ακόμα ένα πλεονέκτημα είναι το επιπλέον gammatone φιλτράρισμα που ακολουθεί το μοντέλο ηλεκτρομηχανικής μεταγωγής. Στα κανάλια χαμηλής συχνότητας (περίπου στο 1.4 kHz) ο βασικός λόγος ύπαρξης αυτών των φίλτρων είναι ο διαχωρισμός της DC συνιστώσας από την χρονική λεπτή δομή, όπως απαιτείται από τον υπολογισμό της φάσης. Στα κανάλια υψηλών συχνοτήτων, μπορούν να εφαρμοστούν παράλληλα φίλτρα με τη μορφή μιας τράπεζας φίλτρων διαμόρφωσης ή να προσαρμοστούν στην θεμελιώδη συχνότητα ή στον τόνο ενός συγκεκριμένου ομιλητή. Συνοπτικά το μοντέλο παρουσιάζεται στο Σχήμα 2.9

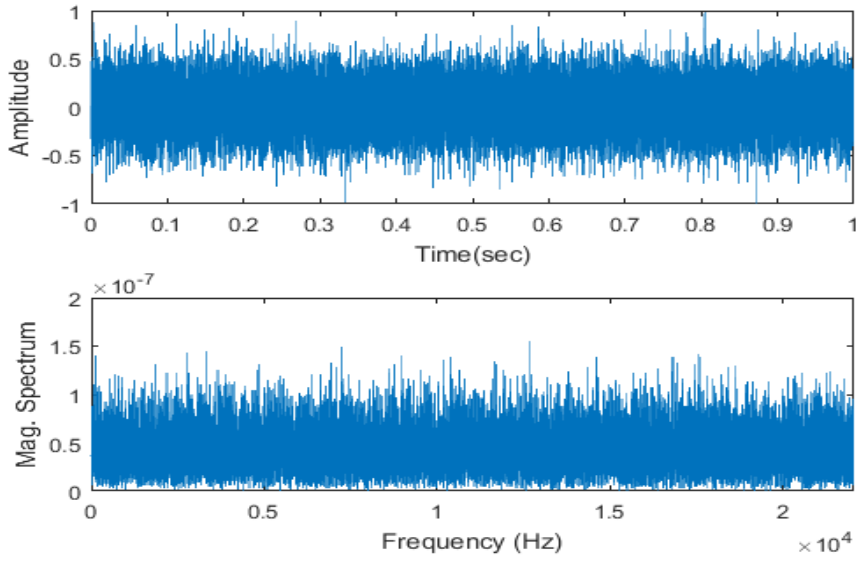


Σχήμα 2.9: Τα στάδια επεξεργασίας του ακουστικού μοντέλου. Η περιφερική επεξεργασία χωρίζει το σήμα εισόδου σε 23 ακουστικά φίλτρα ανά αυτί, και ακολουθείται από ανόρθωση ημίσεος κύματος, χαμηλοπερατό φιλτράρισμα και συμπίεση.

2.6 Λευκός Θόρυβος

Στην επεξεργασία σημάτων ο λευκός θόρυβος, είναι ένα τυχαίο σήμα, που έχει σταθερή ένταση σε όλες τις συχνότητες, δηλαδή σταθερή πυκνότητα φάσματος ισχύος [35]. Ο όρος χρησιμοποιείται σε διάφορους τομείς, όπως η φυσική, οι τηλεπικοινωνίες και η ακουστική. Στον διακριτό χρόνο, ο λευκός θόρυβος είναι ένα διακριτό σήμα, του οποίου τα δείγματα αντιμετωπίζονται ως μία σειρά ασυσχέτιστων τυχαίων μεταβλητών, με μέση τιμή μηδέν, και πεπερασμένη απόκλιση. Εφόσον ο λευκός θόρυβος ακολουθεί την κατανομή που περιγράφεται στην εξίσωση 12, αποκαλείται Additive White Gaussian Noise (AWGN). Αξίζει να σημειωθεί πως ο λευκός θόρυβος είναι μια καθαρά θεωρητική κατασκευή, από την άποψη ότι δεν υφίστανται σήματα με άπειρο εύρος ζώνης. Τυπικά ένα σήμα λευκού θορύβου έχει τη μορφή που φαίνεται στο Σχήμα 2.10

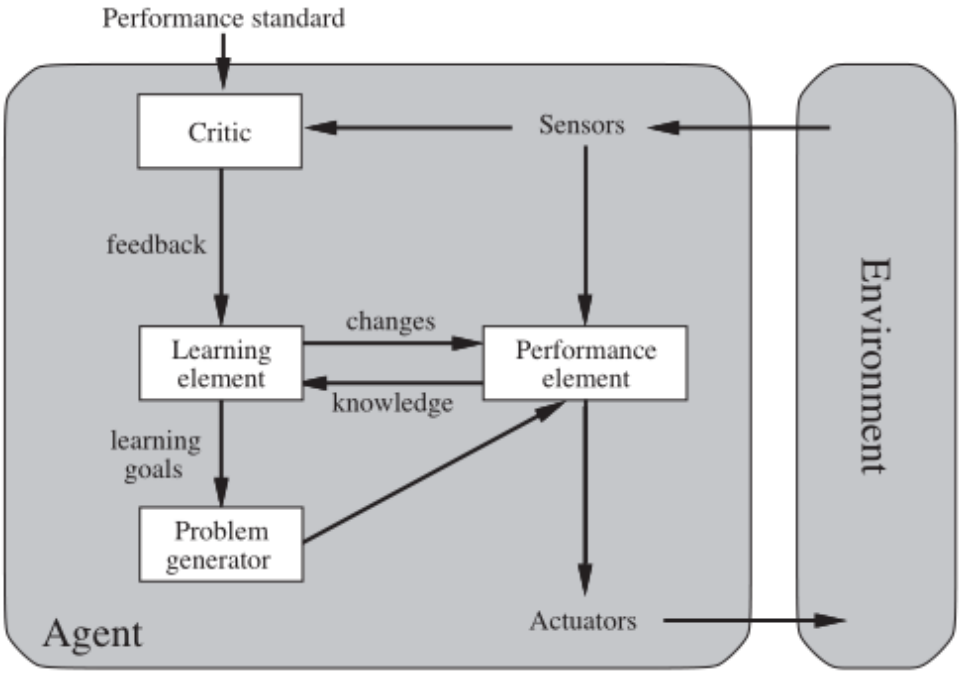
$$Z_n \sim N(0, W) \tag{12}$$



Σχήμα 2.10: Τυπική μορφή σήματος λευκού θορύβου, στον χρόνο (Πάνω) και στη συχνότητα (Κάτω)

2.7 Μηχανική Μάθηση

Η ιδέα πίσω από τη μάθηση είναι ότι οι αισθήσεις δεν πρέπει να χρησιμοποιούνται απλώς για άμεση δράση, αλλά και για τη βελτίωση της ικανότητας ενός πράκτορα να ενεργεί στο μέλλον. Πράκτορας είναι οτιδήποτε μπορεί να θεωρηθεί ότι αντιλαμβάνεται το περιβάλλον του μέσω αισθητήρων, και επενεργεί σε αυτό το περιβάλλον μέσω μηχανισμών δράσης (Σχήμα 2.11). Η μάθηση μπορεί να κυμαίνεται από την απλή απομνημόνευση των εμπειριών, μέχρι τη δημιουργία ολόκληρων επιστημονικών θεωριών. Εδώ περιγράφεται η επαγωγική μάθηση, κατά την οποία ένα σύστημα βελτιώνεται μέσω παρατηρήσεων [36].



Σχήμα 2.11: Αλληλεπίδραση πράκτορα με το περιβάλλον του

Ένας πράκτορας μπορεί να θεωρηθεί ότι περιλαμβάνει ένα στοιχείο εκτέλεσης, το οποίο αποφασίζει τι ενέργειες θα πραγματοποιήσει, και ένα στοιχείο μάθησης που τροποποιεί το στοιχείο εκτέλεσης έτσι ώστε να λαμβάνει καλύτερες αποφάσεις. Οι ερευνητές της μηχανικής μάθησης έχουν ανακαλύψει μια μεγάλη ποικιλία στοιχείων μάθησης. Για να γίνουν αυτά

κατανοητά, είναι χρήσιμο να αποσαφηνιστεί το με ποιον τρόπο η σχεδίασή τους επηρεάζεται από το περιβάλλον στο οποίο θα εφαρμοστούν. Η σχεδίαση ενός στοιχείου μάθησης επηρεάζεται από τρία κύρια ζητήματα:

- Ποιες συνιστώσες του στοιχείου εκτέλεσης πρέπει να μαθευτούν.
- Τι ανάδραση διατίθεται για τη μάθηση αυτών των συνιστωσών.
- Ποια αναπαράσταση χρησιμοποιείται για τις συνιστώσες.

Συνοπτικά **στοιχείο εκτέλεσης** είναι το στοιχείο του πράκτορα που είναι υπεύθυνο για εξωτερικές πράξεις και **στοιχείο μάθησης** το στοιχείο που είναι υπεύθυνο για την υλοποίηση βελτιώσεων στο σύστημα.

Ο *τύπος της ανάδρασης* που διατίθεται για τη μάθηση είναι συνήθως ο πιο σημαντικός παράγοντας για τον προσδιορισμό της φύσης του μαθησιακού προβλήματος. Διακρίνονται τρεις περιπτώσεις: **μη επιβλεπόμενη μάθηση** (unsupervised learning), **επιβλεπόμενη μάθηση** (supervised learning) και **ενισχυτική μάθηση** (reinforcement learning).

2.7.1 Μη επιβλεπόμενη μάθηση

Το πρόβλημα της μη επιβλεπόμενης μάθησης περιλαμβάνει τη μάθηση προτύπων εισόδου χωρίς να παρέχονται συγκεκριμένες τιμές εξόδου. Δύο βασικές μέθοδοι είναι η *ανάλυση σε κύριες συνιστώσες* (Principal Component Analysis - PCA) και η *ανάλυση συστάδων* (cluster analysis). Χρησιμοποιείται ευρέως στην εκτίμηση πυκνοτήτων πιθανοτήτων (pdf) στη στατιστική.

2.7.2 Ενισχυτική μάθηση

Η ενισχυτική μάθηση είναι ένα πεδίο της μηχανικής μάθησης που ασχολείται με τον τρόπο που πράκτορες λογισμικού πρέπει να πάρουν αποφάσεις και να τις εκτελέσουν σε ένα περιβάλλον ώστε να μεγιστοποιήσουν κάποια ανταμοιβή. Όπως και η μη επιβλεπόμενη μάθηση, δεν χρειάζεται συγκεκριμένες τιμές εξόδου. Υλοποιείται τυπικά με την μορφή αλυσίδων αποφάσεων Markov

2.7.3 Επιβλεπόμενη μάθηση

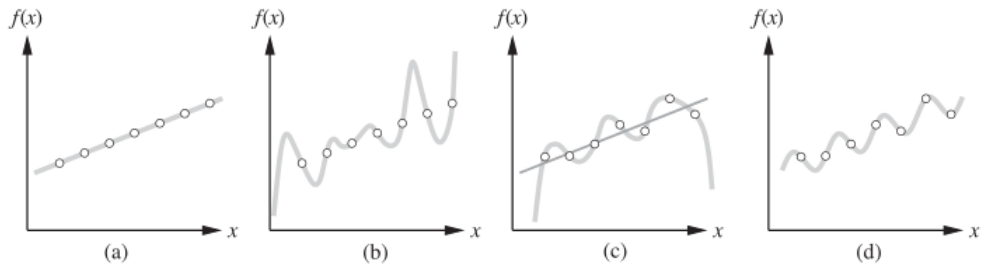
Η επιβλεπόμενη μάθηση είναι η διαδικασία της χαρτογράφησης μιας εισόδου σε μία έξοδο, με βάση *labeled* δεδομένων εισόδου, που αποτελούνται από παραδείγματα εισόδου-εξόδου. Με αυτόν τον τρόπο, εκτιμάται μια συνάρτηση με βάση αυτά τα ζεύγη παραδειγμάτων. Κάθε παράδειγμα, είναι ένα ζεύγος, που αποτελείται από ένα αντικείμενο εισόδου, τυπικά ένα διάνυσμα, και μια επιθυμητή τιμή εξόδου. Σε ένα ιδανικό σενάριο, ο αλγόριθμος θα μπορεί να εκτιμήσει σωστά την επιθυμητή έξοδο από περιπτώσεις που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση. Σε αυτή την κατηγορία αλγορίθμων μάθησης, εντάσσονται και τα Νευρωνικά Δίκτυα που θα απασχολήσουν αυτή την εργασία.

Πιο αυστηρά, ο στόχος της επιβλεπόμενης μάθησης είναι: Δοθέντος ενός **συνόλου εκπαίδευσης** N παραδειγμάτων ζευγών εισόδου-εξόδου:

$$(x_1, y_1), (x_2, y_2), ...(x_N, y_N)$$

όπου κάθε y_i δημιουργείται από μια άγνωστη συνάρτηση $y = f(x)$, να βρεθεί μια συνάρτηση h που προσεγγίζει την f.

Όταν η έξοδος y ανήκει σε ένα σύνολο πεπερασμένων τιμών, (λ.χ. ναι, όχι, ίσως) το πρόβλημα μάθησης λέγεται **ταξινόμηση** (classification). Όταν το y είναι ένας πραγματικός αριθμός, το πρόβλημα λέγεται **regression**.



Σχήμα 2.12: Παραδείγματα ζευγών $(x, f(x))$ και υποθέσεις διαφορετικών βαθμών πολυωνύμου (a) 1ου βαθμού, (b) 7ου βαθμού. (c) Ένα διαφορετικό dataset και μια γραμμική εκτίμηση, (d) ημιτονοειδής εκτίμηση

Στο Σχήμα 2.12 φαίνεται ένα σύνηθες παράδειγμα, η προσαρμογή

μιας καμπύλης σε ένα σύνολο ζευγών σημείων. Δεν είναι γνωστό ποια είναι η f αλλά προσεγγίζεται με μια h επιλεγμένη από τον χώρο υποθέσεων $\mathcal{H}$. Στο παράδειγμα αυτό είναι ένα σύνολο πολυωνύμων.

Η επιβλεπόμενη μάθηση μπορεί να υλοποιηθεί επιλέγοντας την υπόθεση h^* που είναι πιθανότερη με βάση τα δεδομένα:

$$h^* = \underset{h \subset \mathcal{H}}{\operatorname{argmax}} P(h|data) \tag{13}$$

Η Εξίσωση 13 με βάση τον κανόνα του Bayes, μετασχηματίζεται στην Εξίσωση 14 και μπορεί κανείς να ισχυριστεί πως η εκ των προτέρων πιθανότητα $P(h)$ είναι μεγάλη για ένα πολυώνυμο μικρού βαθμού και χαμηλότερη για ένα πολυώνυμο μεγάλου βαθμού. Γενικότερα επιτρέπονται περίπλοκες ή *παράξενες* προσεγγίσεις μόνο όταν κρίνεται ότι το απαιτούν τα δεδομένα, αλλά αποφεύγεται με τον να τους ανατίθεται χαμηλή πιθανότητα. Γενικά, αξίζει να σημειωθεί πως υπάρχει πάντα ένα tradeoff μεταξύ της εκφραστικότητας ενός χώρου υποθέσεων, και της πολυπλοκότητας της εύρεσης μιας καλής υπόθεσης μέσα σε αυτό το χώρο.

$$h^* = \underset{h \subset \mathcal{H}}{\operatorname{argmax}} P(data|h)P(h) \tag{14}$$

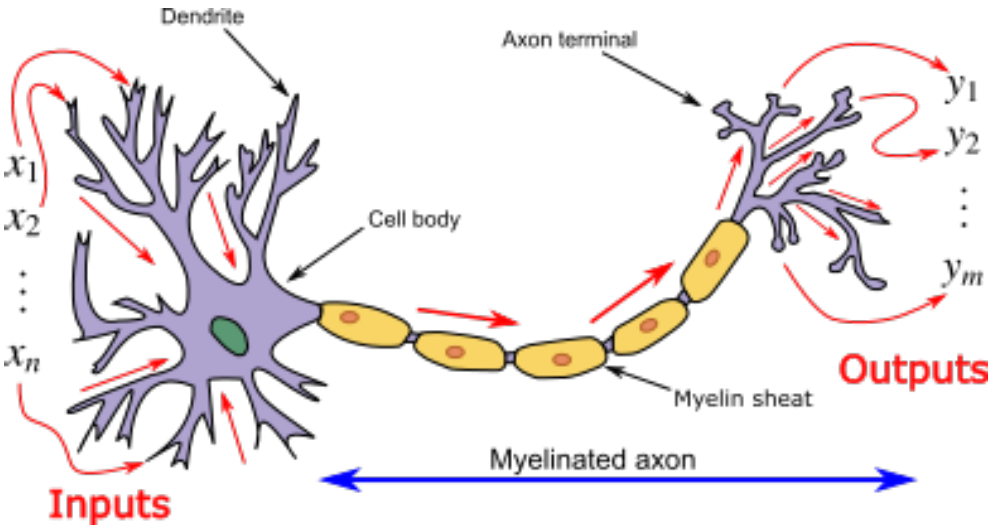
2.8 Νευρωνικά Δίκτυα

Νευρωνικά Δίκτυα, ή με τον πιο δόκιμο όρο, Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANN), είναι ένα σύνολο από τεχνητούς νευρώνες ή κόμβους [37]. Ουσιαστικά είναι υπολογιστικά συστήματα εμπνευσμένα από βιολογικά νευρωνικά δίκτυα που έχουν μελετηθεί σε εγκεφάλους διαφόρων ζώων [38]. Τα συστήματα αυτά 'μαθαίνουν' να εκτελούν συγκεκριμένο έργο λαμβάνοντας υπόψιν τους παραδείγματα, συνήθως χωρίς να είναι προγραμματισμένα με συγκεκριμένους κανόνες που αφορούν το εν λόγω έργο. Εκτελούν το έργο που τους ανατίθεται, χωρίς να έχουν πρότερη γνώση επί αυτού.

2.8.1 Νευρώνες

Ένα ANN βασίζεται σε μια συλλογή νευρώνων οι οποίοι μοντελοποιούν τους νευρώνες ενός βιολογικού εγκεφάλου. Κάθε σύνδεση, όπως οι συνάψεις στον εγκέφαλο, μπορεί να μεταδόσει σήματα στους άλλους νευρώνες. Ένας τεχνητός νευρώνας δέχεται το σήμα, το επεξεργάζεται και στη συνέχεια μεταβιβάζει το επεξεργασμένο σήμα στους νευρώνες με τους οποίους συνδέεται.

Αυστηρότερα, ένας νευρώνας είναι μια *συνάρτηση*, και αποτελεί το δομικό στοιχείο ενός ANN. Δέχεται μία ή περισσότερες εισόδους, (που αντιπροσωπεύουν τα διεγερτικά μετασυναπτικά δυναμικά και τα ανασταλτικά μετασυναπτικά δυναμικά στους δενδρίτες των νευρώνων) τις προσθέτει και μεταδίδει μια έξοδο, (ή ενεργοποίηση που αντιπροσωπεύει το δυναμικό δράσης ενός βιολογικού νευρώνα που μεταδίδεται κατά μήκος του άξονά του). Η αντιστοίχηση φαίνεται στο Σχήμα 2.13. Συνήθως σε κάθε είσοδο ανατίθεται ένα βάρος και το άθροισμά τους δίνεται ως είσοδος σε μια μη γραμμική συνάρτηση που αποκαλείται ***συνάρτηση ενεργοποίησης*** *ή συνάρτηση μεταφοράς* του νευρώνα.



Σχήμα 2.13: Νευρώνας και μυελινωμένος άξονας, με τη ροή του σήματος από τις εισόδους στους δενδρίτες στις εξόδους στον άξονα.

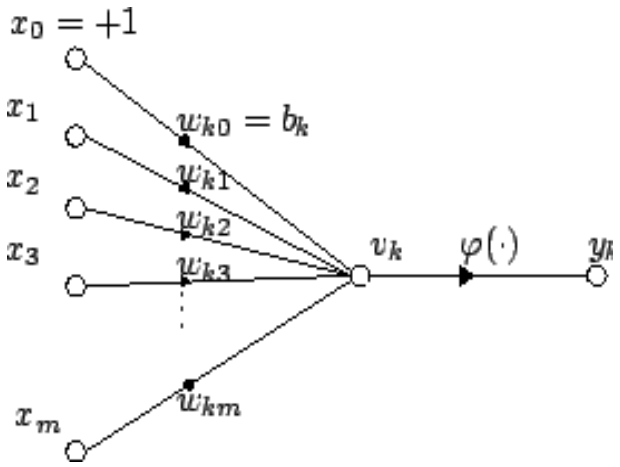
Βασική Δομή Νευρώνα

Για έναν τεχνητό νευρώνα, έστω m + *1* είσοδοι με σήματα από x_0 μέχρι x_m. Η είσοδος σε έναν νευρώνα k φαίνεται στην εξίσωση 15, όπου με φ συμβολίζεται η συνάρτηση μεταφοράς ενώ η δομή ενός νευρώνα φαίνεται στο Σχήμα 2.14.

$$y_k = \phi(\sum_{j=0}^{m} w_{kj} x_j) \tag{15}$$

Συναρτήσεις ενεργοποίησης

- **Βηματική Συνάρτηση**: Η έξοδος αυτής της συνάρτησης είναι δυαδική και εξαρτάται από το αν η είσοδος ξεπερνά ή είναι ίση με ένα ορισμένο κατώφλι θ. Εκτελεί μια διαίρεση του χώρου εισόδων με ένα υπερεπίπεδο. Είναι ιδιαίτερα χρήσιμη όταν εφαρμόζεται στο τελευ-



Σχήμα 2.14: Μαθηματικό μοντέλο ενός νευρώνα με bias στο $x_0 = 1$

ταίο επίπεδο ενός ANN, που κάνει δυαδική συσταδοποίηση.

$$y = \begin{cases} 1 & \text{if } x \geq \theta \\ 0 & \text{if } x < \theta \end{cases} \tag{16}$$

- **Σιγμοειδής:** Μια μη-γραμμική συνάρτηση με παράγωγο που υπολογίζεται εύκολα, και κατ' επέκταση είναι υπολογιστικά απλή. Ένα παράδειγμα είναι η γνωστή logistic function:

$$y = \frac{e^x}{e^x + 1} \tag{17}$$

- **Συνάρτηση Ανόρθωσης:** Οι συναρτήσεις αυτές είναι γνωστές και ως συναρτήσεις ράμπας, και είναι το υπολογιστικό ανάλογο ενός ανορθωτή ημίσεος κύματος. Λέγονται επίσης Rectified Linear Units ή ReLU και ορίζονται ως εξής:

$$y = x^+ = \max(0, x) \tag{18}$$

2.8.2 Οργάνωση

Οι νευρώνες τυπικά είναι οργανωμένοι σε πολλά επίπεδα, ιδιαίτερα στην περίπτωση του deep learning. Το επίπεδο που δέχεται τα δεδομένα είναι το *επίπεδο εισόδου*, ενώ το επίπεδο που δίνει το τελικό αποτέλε-

σμα είναι το *επίπεδο εξόδου*. Μεταξύ τους μπορεί να υπάρχουν μηδέν ή περισσότερα *κρυφά επίπεδα*. Τα επίπεδα μπορεί να είναι συνδεδεμένα με πολλούς τρόπους. Για παράδειγμα μπορεί να είναι *πλήρως διασυνδεδεμένα*, με κάθε νευρώνα σε ένα επίπεδο να συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου.

Υπερπαράμετροι

Μια υπερπαράμετρος είναι μια σταθερή παράμετρος της οποίας η τιμή ορίζεται πριν την εκκίνηση της διαδικασίας εκπαίδευσης. Τέτοιες παράμετροι είναι ο ρυθμός μάθησης (learning rate - LR), ή ο αριθμός των κρυφών επιπέδων.

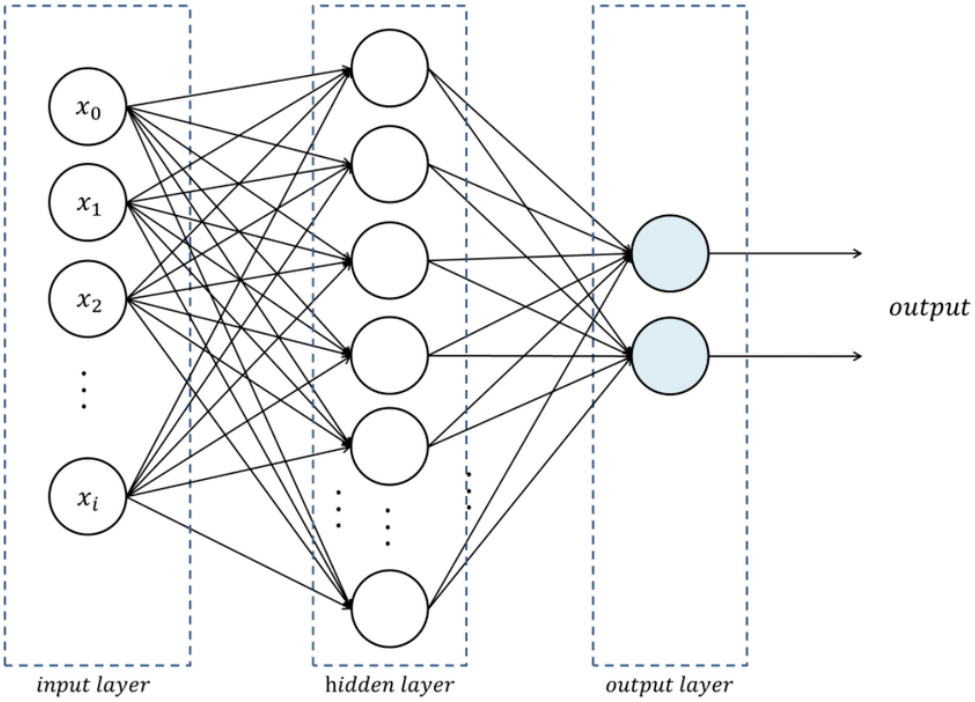
Backpropagation

Το backpropagation είναι μια μέθοδος ρύθμισης των βαρών των συνάψεων για την διόρθωση των σφαλμάτων κατά την εκπαίδευση. Το σφάλμα πρακτικά 'μοιράζεται' σε όλες τις συνάψεις. Υπολογίζονται οι παράγωγοι των συναρτήσεων ενεργοποίησης ανάλογα με τα τωρινά βάρη, και ενημερώνονται αναλόγως με την μέθοδο βελτιστοποίησης που έχει επιλεχθεί. Σε αυτή την εργασία όλα τα μοντέλα χρησιμοποιούν τον Adam optimizer.

Αρχιτεκτονικές

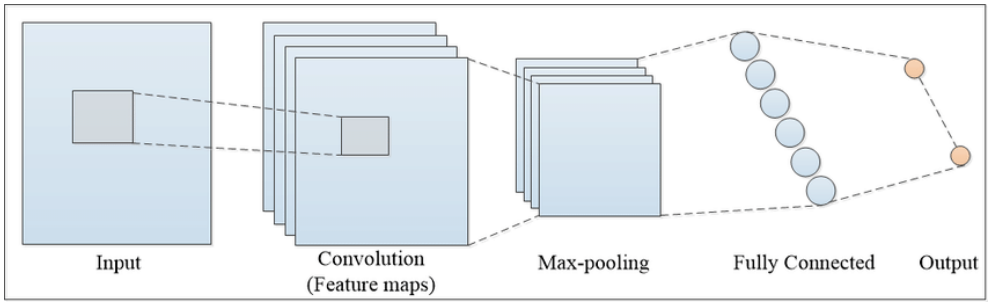
Μία από τις αρχιτεκτονικές που χρησιμοποιείται σε αυτή την εργασία είναι ο πολυεπίπεδος Perceptron (Multilayered Perceptron - MLP). Η αρχιτεκτονική αυτή αποτελείται από τουλάχιστον τρία επίπεδα, ένα επίπεδο εισόδου, ένα επίπεδο εξόδου και τουλάχιστον ένα κρυφό επίπεδο, όλα πλήρως διασυνδεδεμένα μεταξύ τους. Η απλούστερη δομή (από την άποψη του αριθμού των επιπέδων) ενός MLP φαίνεται στο Σχήμα 2.15.

Η δεύτερη αρχιτεκτονική που χρησιμοποιείται είναι αυτή των συνελικτικών δικτύων (Convolutional Neural Networks - CNNs). Η δομή αυτή



Σχήμα 2.15: Η δομή του απλούστερου MLP.

είναι αρκετά πιο σύγχρονη από αυτή του MLP, και χρησιμοποιείται συχνά για την ανάλυση εικόνων και προβλήματα που αφορούν την υπολογιστική όραση. Αυστηρά, συνελικτικό δίκτυο είναι κάθε ANN που χρησιμοποιεί την πράξη της συνέλιξης. Στην πράξη, τα CNN, σε αντίθεση με τους MLP αντί να εκτολούν ένα γινόμενο πινάκων, χρησιμοποιούν συνέλιξη. Κάθε συνελικτικό επίπεδο, αποτελείται από έναν αριθμό φίλτρων που ορίζεται προτού ξεκινήσει η εκπαίδευση, όπως και οι διαστάσεις αυτών. Κάθε συνελικτικός νευρώνας επεξεργάζεται μόνο τα δεδομένα που είναι σε μια περιορισμένη υποπεριοχή της εξόδου του προηγούμενου επιπέδου. Μια τυπική δομή ενός CNN φαίνεται στο Σχήμα 2.16. Οι λειτουργία των επιπέδων Pooling εξηγείται σε επόμενη ενότητα.



Σχήμα 2.16: Η δομή ενός απλού CNN.

2.8.3 Συναρτήσεις Απώλειας

Κάθε αλγόριθμος βελτιστοποίησης χρησιμοποιεί μια συνάρτηση για την *αξιολόγηση* μιας πιθανής λύσης (π.χ. το σύνολο των βαρών στις συνάψεις ενός NN). Η συνάρτηση αυτή καλείται αντικειμενική συνάρτηση, και η βέλτιστη λύση είναι αυτή που ελαχιστοποιεί την αντικειμενική συνάρτηση σε προβλήματα ελαχιστοποίησης (ή το αντίθετο σε προβλήματα μεγιστοποίησης) [39].

Μέσο Απόλυτο Σφάλμα

Το MAE, είναι ένα μέτρο της διαφοράς μεταξύ δύο συνεχών μεταβλητών. Υποθέτοντας μεταβλητές X και Υ, οι οποίες είναι η πραγματική και η εκτιμώμενη τιμή αντίστοιχα, τότε για n παρατηρήσεις το MAE υπολογίζεται από την Εξίσωση 19. Το MAE έχει τις ίδιες μονάδες με τα δεδομένα που παρατηρούνται και πεδίο ορισμού $D = [0, +\infty)$.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{19}$$

Μέσο Τετραγωνικό Σφάλμα

Το MSE ενός εκτιμητή, υπολογίζει τον μέσο όρο των τετραγώνων των σφαλμάτων, δηλαδή τον μέσο όρο των τετραγώνων των διαφορών της εκτιμώμενης τιμής από την πραγματική. Το MSE, είναι η δεύτερη ροπή

(ως προς το 0) του σφάλματος, οπότε περιέχει τόσο την τυπική απόκλιση, δηλαδή το πόσο 'απλωμένες' είναι οι εκτιμήσεις, αλλά και το *bias*, δηλαδή το πόσο μακριά είναι η μέση εκτιμώμενη από την μέση πραγματική τιμή. Όπως η διακύμανση, έτσι και το MSE, έχει τις μονάδες των δεδομένων παρατήρησης υψωμένες στο τετράγωνο. Χρησιμοποιείται, όταν είναι ιδιαίτερα ανεπιθύμητα μεγάλα σφάλματα στον εκτιμητή, έχει πεδίο ορισμού το $D = [0, +\infty)$, και υπολογίζεται όπως φαίνεται στην εξίσωση 20.

$$MSE = \frac{\sum_{i=1}^{n} |y_i - x_i|^2}{n} \tag{20}$$

Ρίζα Μέσου Τετραγωνικού Σφάλματος

Το RMSE, αντιπροσωπεύει την τετραγωνική ρίζα της δεύτερης ροπής των διαφορών μεταξύ των εκτιμώμενων τιμών και των πραγματικών. Όπως και τα προαναφερθέντα μέτρα, είναι θετικά ορισμένο, με πεδίο ορισμού το $D = [0, +\infty)$. Συνδυάζει τα πλεονεκτήματα του MSE και του MAE, από την άποψη ότι 'τιμωρεί' περισσότερο τις μεγάλες αποκλίσεις από τις πραγματικές τιμές, αλλά είναι ταυτόχρονα πιο εύκολα ερμηνεύσιμο, όπως το MAE, αφού είναι και αυτό σε γραμμική κλίμακα. Υπολογίζεται όπως φαίνεται στην Εξίσωση 21

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} |y_i - x_i|^2}{n}} \tag{21}$$

ΚΕΦΑΛΑΙΟ 3

Υλοποίηση

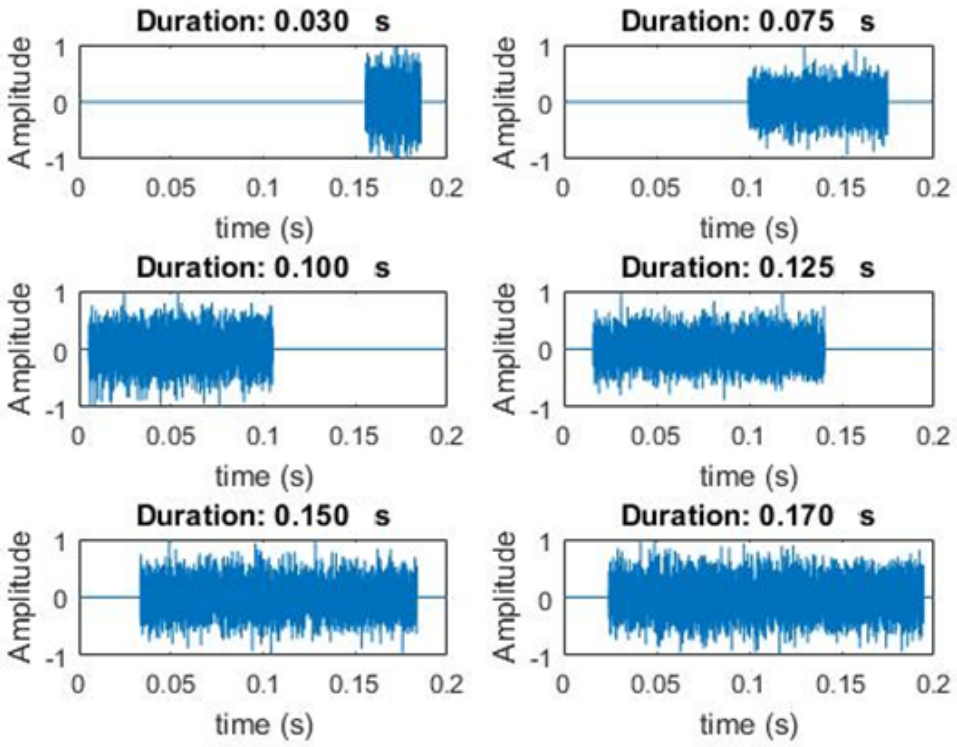
Στο κεφάλαιο 2 αναλύθηκε με λεπτομέρεια το θεωρητικό υπόβαθρο στο οποίο βασίστηκε η εργασία. Εδώ περιγράφεται η προσέγγιση που επιλέχθηκε για την επίλυση του προβλήματος της εκτίμησης DOA μέσω αμφιωτικών παραμέτρων. Αρχικά προσεγγίζεται η δημιουργία των σημάτων θορύβου που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων, στη συνέχεια η διαδικασία κατασκευής των αμφιωτικών σημάτων μέσω BRIR, καθώς και η εξαγωγή των διωτικών παραμέτρων με το μοντέλο dietz2011. Έπειτα περιγράφεται η καινούρια μέθοδος συμπίεσης των αμφιωτικών παραμέτρων που υλοποιήθηκε για τους σκοπούς αυτής της εργασίας, καθώς και η επιπλέον προεπεξεργασία των δεδομένων προτού αυτά χρησιμοποιηθούν για την εκπαίδευση των διάφορων μοντέλων. Αναλύονται επίσης οι διάφορες αρχιτεκτονικές που χρησιμοποιήθηκαν, καθώς και οι μετρικές που είναι απαραίτητες για την αξιολόγηση των μοντέλων. Τα μοντέλα δοκιμάστηκαν για δειγματοληψία $44.1\ kHz$.

3.1 Σήματα εισόδου

Τα πηγαία σήματα που χρησιμοποιήθηκαν για την εκπαίδευση των NN ήταν 'εκρήξεις' (bursts) λευκού θορύβου, ο οποίος έχει χαρακτηριστικά που περιγράφονται στην υποενότητα 2.6. Συνολικά η διάρκεια κάθε σήματος ήταν $200ms$. Στην περίπτωση της δειγματοληψίας με $f_s = 44.1kHz$ αυτό αντιστοιχεί σε 8820 δείγματα. Οι εκρήξεις είχαν μεταβλητή διάρκεια $3 - 170ms$ με τυχαίο τρόπο. Τυχαίο ήταν επίσης το σημείο εκκίνησης της κάθε έκρηξης στο συνολικό σήμα. Δημιουργήθηκαν συνολικά 29 τέτοια σήματα και η γενική περιγραφή κάθε σήματος φαίνεται στην εξίσωση 22, όπου $Z_n \sim N(0, P)$ δηλαδή μια κανονική κατανομή με μέση τιμή μηδέν και διακύμανση-ισχύ P.

$$Sig_in(n) = \begin{cases} Z_n, & n_{start} < n < n_{end} \\ 0, & \text{αλλού} \end{cases} \tag{22}$$

Μερικά από τα σήματα εισόδου φαίνονται στο Σχήμα 3.1. Μετά την δημιουργία τους κανονικοποιούνται στο κλειστό διάστημα $amplitude = [-1, 1]$, ώστε τελικά να αποθηκευτούν σε μορφή wav για μελλοντική χρήση. Η διάρκεια των σημάτων καθορίστηκε στα $200ms$, προκειμένου να διατηρηθεί σχετικά μικρό το διάνυσμα εισόδου στο NN, αλλά και να προλάβει να επέλθει η σταθερή κατάσταση μετά τη μεταβατική κατάσταση (transient) που δημιουργείται στην έναρξη και την παύση του σήματος, αφού κρίθηκε πως αυτό έχει ιδιαίτερη σημασία στον τρόπο που αντιλαμβάνεται ο άνθρωπος τον ήχο, και άρα να τροφοδοτηθεί το NN, με δεδομένα που έχουν υψηλή συγκέντρωση χρήσιμης πληροφορίας.



Σχήμα 3.1: Σήματα λευκού θορύβου που χρησιμοποιήθηκαν ως είσοδος στο μοντέλο.

3.2 Δημιουργία αμφιωτικών σημάτων

Τα σήματα που περιγράφονται στην υποενότητα 3.1 στη συνέχεια συνελίσονται με BRIRs, για διαφορετικές γωνίες, που προέρχονται από τρία δωμάτια: Το TU Berlin Auditorium 3 [40], το Spirit [41], και το Calypso [42]. Σε κάθε περίπτωση, το ανδρείκελο απείχε 1m από την πηγή, ενώ έγιναν μετρήσεις με ανάλυση 1^o στο οριζόντιο επίπεδο, από -90^o μέχρι $+90^o$ χρησιμοποιώντας το ανδρείκελο KEMAR και ηχεία Genelec, τοποθετημένα σε ανύψωση 0^o.

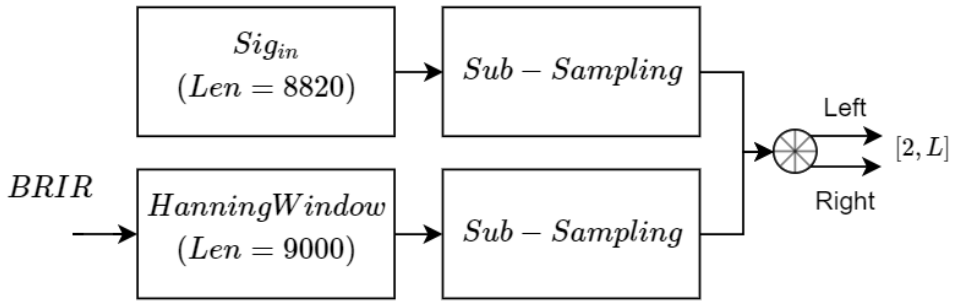
Στην εργασία αυτή χρησιμοποιήθηκε ανάλυση στο οριζόντιο επίπεδο 2^o, με αποτέλεσμα να προκύψουν 91 κρουστικές για κάθε δωμάτιο. Χρησιμοποιήθηκαν και τα τρία δωμάτια, και άρα μετά από τη συνέλιξη κάθε ενός από τα 29 σήματα εισόδου, με μία από τις $3 * 91 = 273$ διαφορετικές κρουστικές, δημιουργούνται συνολικά $29 * 273 = 7917$ διαφορετικά σήματα εισόδου.

Αναλυτικότερα, αφού φορτωθεί το εκάστοτε dataset, και από αυτό διαβαστεί η BRIR, η οποία εμφανώς αποτελείται από δύο κανάλια διότι είναι binaural, εφαρμόζεται σε αυτή ένα παράθυρο Half Hanning μήκους 9000 δειγμάτων, αφού ρυθμιστεί η ζητούμενη δειγματοληψία. Τα δύο κανάλια της BRIR συνελίσονται με το mono σήμα εισόδου και προκύπτει το αμφιωτικό σήμα από το οποίο θα εξαχθούν στη συνέχεια οι αμφιωτικές παράμετροι. Έτσι, το αποτέλεσμα της συνέλιξης είναι ένα αμφιωτικό σήμα διαστάσεων $[2, L]$, όπου $L = M + N - 1 = 17819$, με δεδομένο ότι η BRIR έχει μήκος $M = 9000$ δείγματα και το mono σήμα μήκος $N = 8820$ δείγματα.

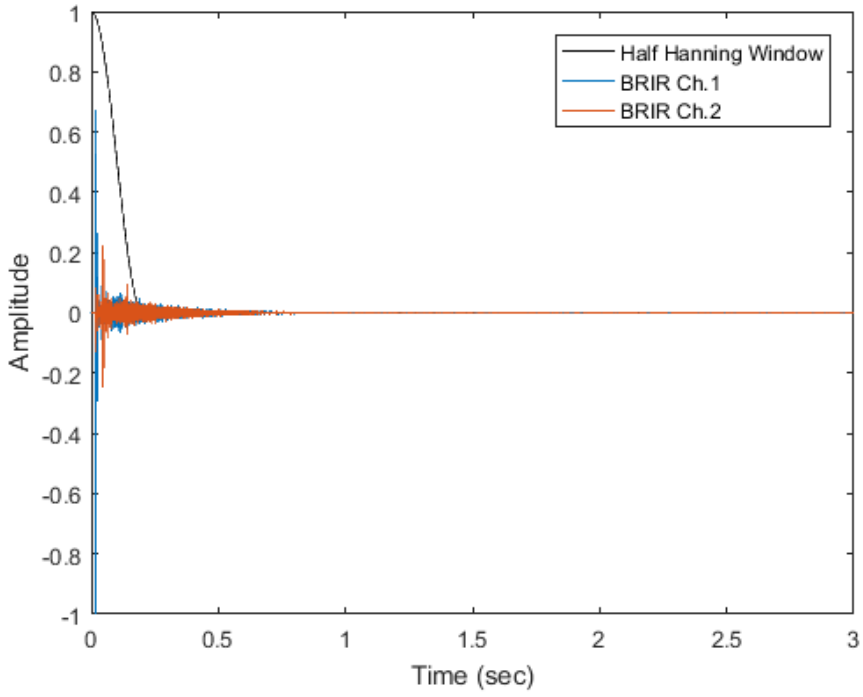
Η παραθύρωση έγινε λόγω περιορισμών στους υπολογιστικούς πόρους, αλλά επιλέχθηκε τέτοιο μέγεθος παραθύρου ώστε να διατηρούνται τα σημαντικότερα χαρακτηριστικά των BRIR.

Μια κρουστική μαζί με το παράθυρο που εφαρμόζεται σε αυτή και η BRIR μετά την εφαρμογή του, φαίνονται στο Σχήμα 3.3. Συνοπτικά η

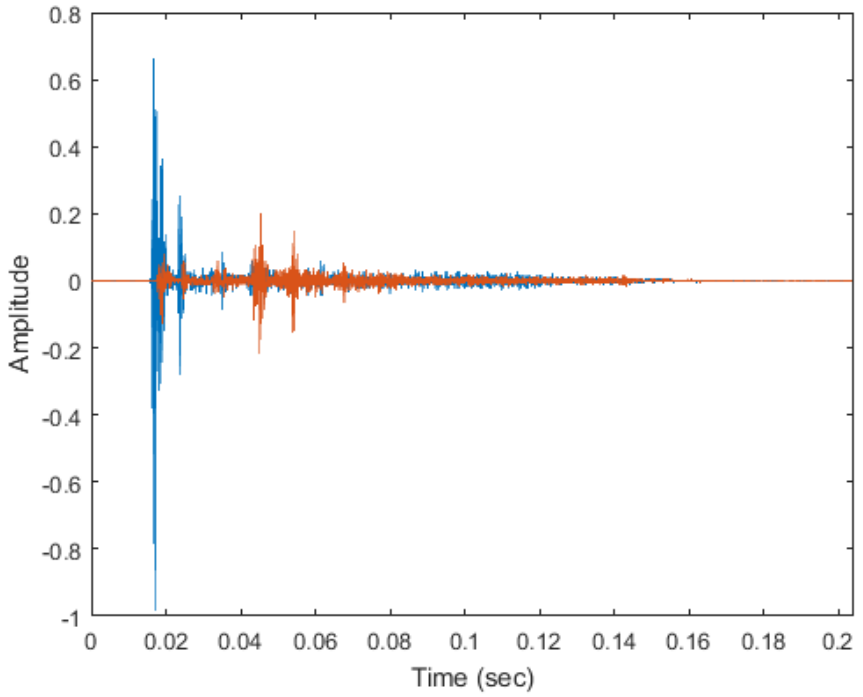
διαδικασία της δημιουργίας των αμφιωτικών σημάτων φαίνεται στο Σχήμα 3.2.



Σχήμα 3.2: Δημιουργία αμφιωτικών σημάτων. Το block sub-sampling χρησιμοποιείται μόνο στην περίπτωση που τα σήματα έχουν διαφορετική δειγματοληψία από την επιθυμητή $f_s = 44.1kHz$



(α′)



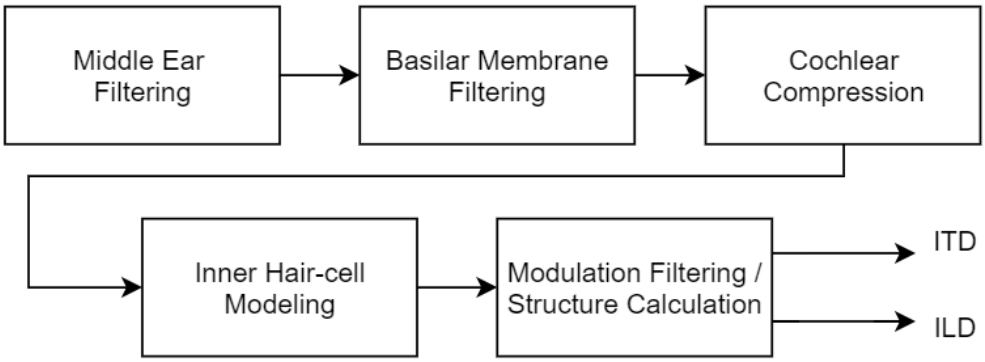
(β′)

Σχήμα 3.3: (α'): BRIR και το αντίστοιχο Half Hanning παράθυρο, (β'): BRIR μετά την παραθύρωση. Τα σχήματα είναι για $f_s = 44.1kHz$

3.3 Εξαγωγή αμφιωτικών παραμέτρων

Σε αυτή την ενότητα, αναλύεται η εφαρμογή του μοντέλου του Dietz το θεωρητικό υπόβαθρο του οποίου περιγράφεται στην υποενότητα 2.5. Η συνάρτηση είναι υλοποιημένη σε MATLAB, στο Auditory Modeling Toolbox. Η συνάρτηση χρειάζεται ως ορίσματα το binaural σήμα και τη συχνότητα δειγματοληψίας, ενώ στην έξοδό της δίνονται οι αμφιωτικές παράμετροι.

Συνοπτικά η διαδικασία που ακολουθείται για τον υπολογισμό των παραμέτρων, φαίνεται στο Σχήμα 3.4 και τα επιμέρους στοιχεία της αναλύονται στη συνέχεια.



Σχήμα 3.4: Συνοπτική περιγραφή της υλοποίησης του μοντέλου Dietz για το ακουστικό σύστημα.

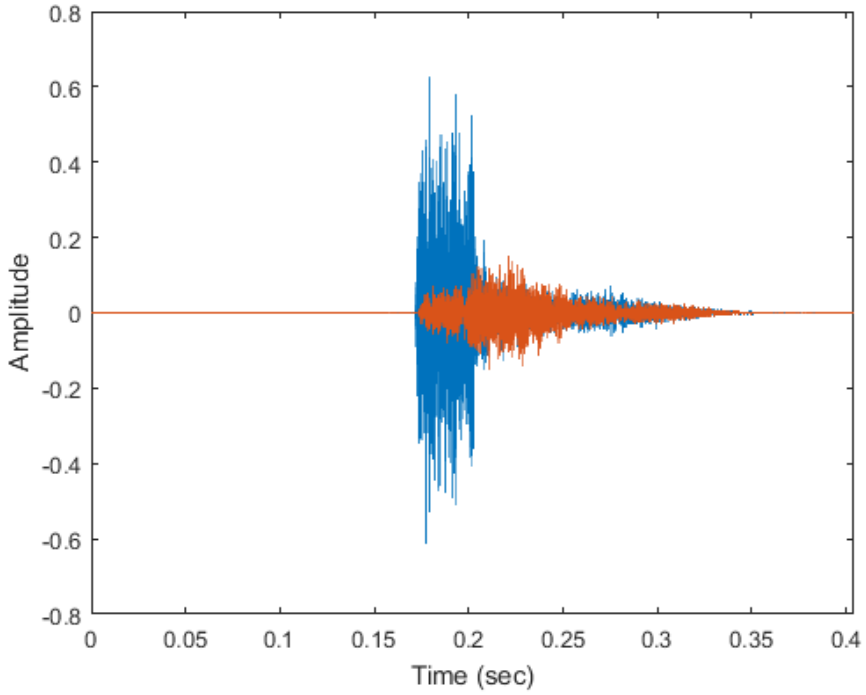
Αξίζει εδώ να σημειωθεί πως η υλοποίηση του μοντέλου, πιστώνεται στους:

- Tobias Peters (tobias@medi.physik.uni-oldenburg.de)
- Mathias Dietz (mathias.dietz@uni-oldenburg.de)
- Martin Klein-Hennig (martin.klein.hennig@uni-oldenburg.de)

Παρακάτω, μελετάται η λειτουργία του μοντέλου, με βάση ένα από τα binaural σήματα εισόδου που χρησιμοποιούνται, που αντιστοιχεί σε γωνία άφιξης $+80^o$ από το δωμάτιο Spirit, όπως φαίνεται στο Σχήμα ??.

3.3.1 Φιλτράρισμα Μέσου Αυτιού

Αρχικά, το σήμα περνά από ένα bandpass butterworth φίλτρο με $f_{c1} = 500Hz$ και $f_{c2} = 2kHz$ το οποίο φαίνεται στο Σχήμα 3.6, ενώ το

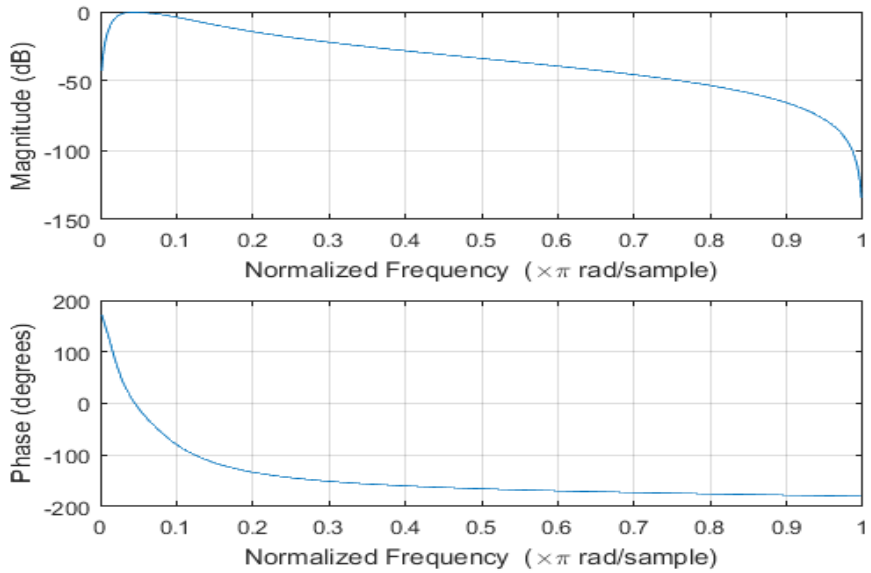


Σχήμα 3.5: Binaural σήμα για γωνία άφιξης $+80^o$, στο δωμάτιο Spirit.

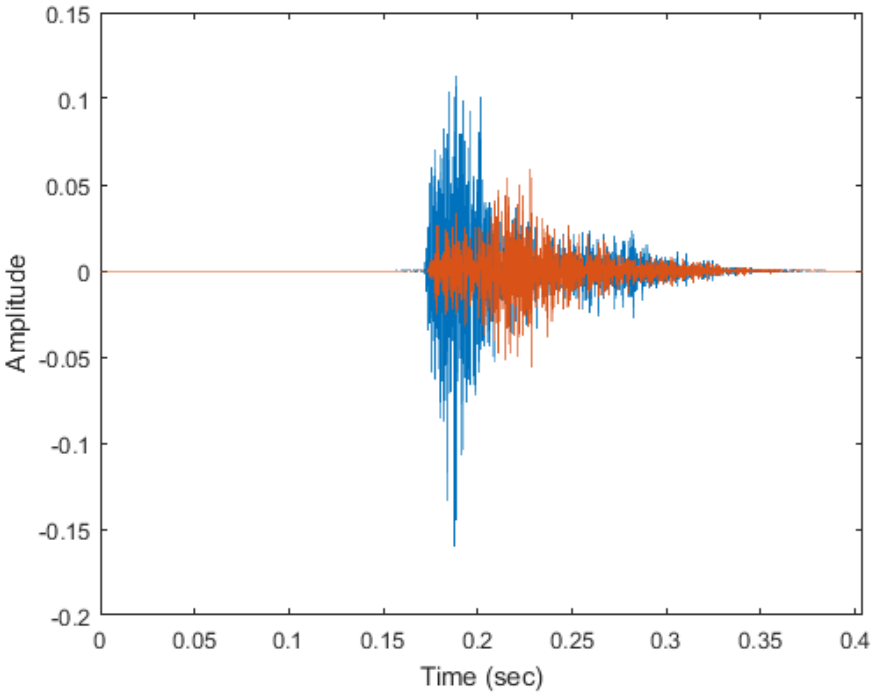
τελικό αποτέλεσμα της επεξεργασίας φαίνεται στο Σχήμα 3.7. Όπως είναι αναμενόμενο, όντας ουσιαστικά σήμα λευκού θορύβου, η μορφή του αλλάζει αισθητά στο πεδίο του χρόνου, αφού μειώνεται σημαντικά το υψίσυχνο περιεχόμενο.

3.3.2 Προσομοίωση Εσωτερικού Αυτιού

Σε αυτό το σημείο, για την προσομοίωση της basilar μεμβράνης, υπολογίζονται 23 gammatone φίλτρα με μιγαδικούς συντελεστές για τον υπολογισμό της ITF (Εξίσωση 8). Το πλήθος των φίλτρων προκύπτει από το γεγονός ότι μεταξύ τους απέχουν 1 ERB, το οποίο αντιστοιχίζεται σε περίπου 217 Hz, και καλύπτουν το διάστημα 200 - 5000 Hz. Η κρουστική απόκριση ενός gammatone φίλτρου, περιγράφεται στην Εξίσωση 23. Ουσιαστικά είναι ένα φίλτρο που προκύπτει από τον πολλαπλασιασμό μιας κατανομής γάμμα, και ενός ημιτονοειδούς τόνου. Οι αποκρίσεις συχνότη-



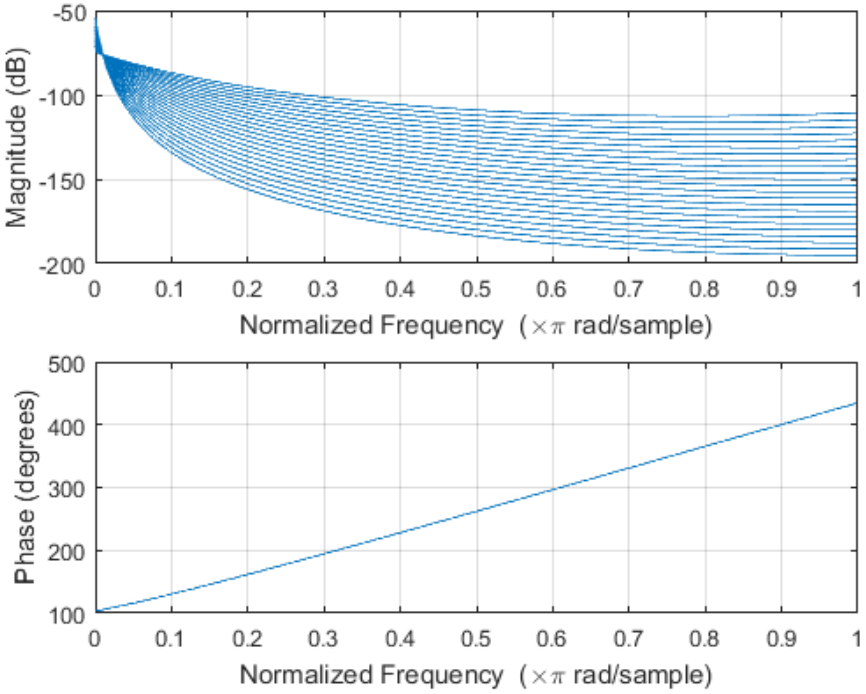
Σχήμα 3.6: IIR φίλτρο τύπου butterworth που προσομοιώνει το φιλτράρισμα του μέσου αυτιού.



Σχήμα 3.7: Αποτέλεσμα της επεξεργασίας του μέσου αυτιού.

τας των 23 φίλτρων παρουσιάζονται στο Σχήμα 3.8. Το αποτέλεσμα του φιλτραρίσματος, είναι μιγαδικό, και αποτελείται από 23 συχνοτικές μπάντες.

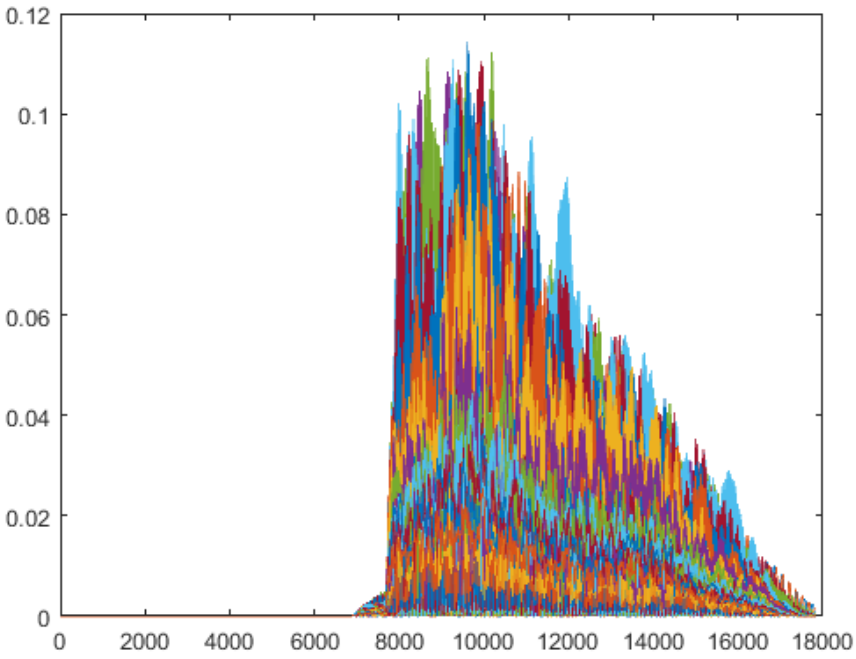
$$g(t) = \alpha t^{n-1} \cos(2\pi f_c t) e^{-2\pi\beta t} \tag{23}$$



Σχήμα 3.8: Αποκρίσεις συχνότητας της τράπεζας φίλτρων gammatone.

Στη συνέχεια για τη μοντελοποίηση της συμπίεσης του κοχλία χρησιμοποιείται η Εξίσωση 24 δείγμα προς δείγμα. Ακολούθως, τα εσωτερικά hair-cells του αυτιού μοντελοποιούνται όπως έχει ήδη αναφερθεί με μια ανόρθωση ημίσεος κύματος και στη συνέχεια ένα lowpass φίλτρο, με αποτέλεσμα την εξαγωγή μια 'περιβάλλουσας'. Στο σημείο αυτό, το ένα εκ των δύο καναλιών του σήματος, έχει τη μορφή που φαίνεται στο Σχήμα 3.9.

$$y(n) = sign(x(n)) * |x(n)|^c \tag{24}$$



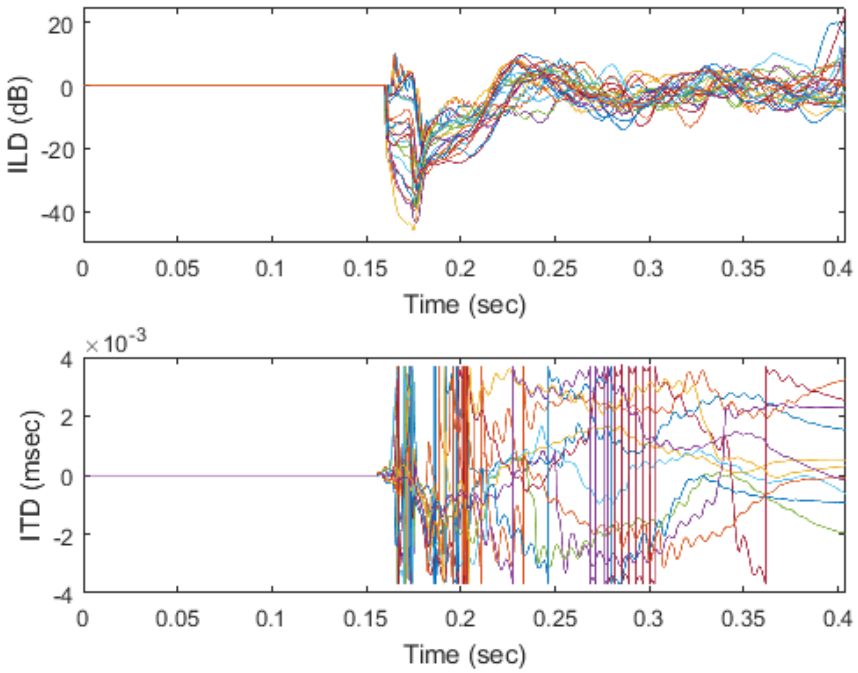
Σχήμα 3.9: Έξοδος του μοντέλου μετά τη μοντελοποίηση του εσωτερικού αυτιού.

3.3.3 Τράπεζα Φίλτρων Διαμόρφωσης

Το επόμενο στάδιο της επεξεργασίας περιέχει ακόμα μια τράπεζα φίλτρων, που αποτελείται από τρία gammatone φίλτρα 2ης τάξης, για τον διαχωρισμό στις δομές 'fine' και 'envelope' που περιέχουν πληροφορία χαμηλών (κάτω από 1.4 kHz) και υψηλών συχνοτήτων αντίστοιχα και ένα lowpass με $f_c = 30Hz$ για τον υπολογισμό του ILD. Κατ' επέκταση, η fine δομή έχει 12 διαστάσεις, που κάθε μια αντιστοιχεί σε συχνοτικές μπάντες από $200-1400Hz$ ενώ η δομή envelope 11 διαστάσεις που αντιστοιχούν στις εναπομείνασες συχνότητες. Τα προαναφερθέντα φίλτρα, εφαρμόζονται σε κάθε μία από τις συχνοτικές μπάντες της εξόδου του προηγούμενου σταδίου. Οι έξοδοι σε αυτό το σημείο, λόγω των μιγαδικών συντελεστών των gammatone φίλτρων είναι πάλι μιγαδικές.

3.3.4 Αμφιωτικός επεξεργαστής

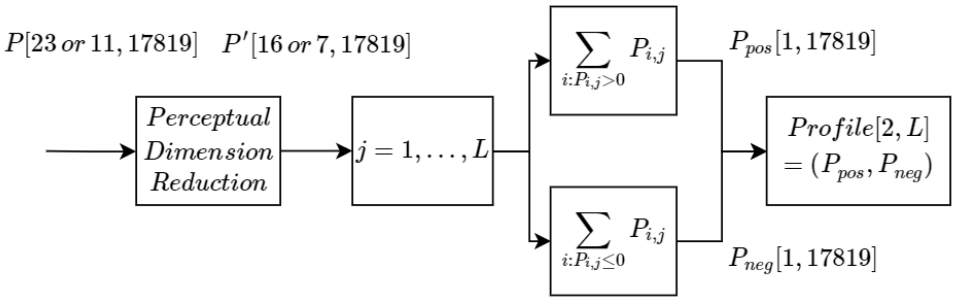
Το τελευταίο στάδιο της επεξεργασίας, υλοποιείται από τον binaural processor, ο οποίος εφαρμόζει τις εξισώσεις που έχουν περιγραφεί αναλυτικά στο κεφάλαιο 2.5 στις δομές fine και envelope. Τα τελικά αποτελέσματα, για τις αμφιωτικές παραμέτρους που αφορούν αυτή την εργασία, παρουσιάζονται στο Σχήμα 3.10.



Σχήμα 3.10: Τελικές έξοδοι του μοντέλου: (Πάνω) ILD για 23 συχνοτικές μπάντες, (Κάτω) ITD για την δομή envelope - 11 μπάντες.

3.4 Συμπίεση Δεδομένων

Σε αυτή την ενότητα, αναλύονται τα βήματα που ακολουθήθηκαν για την μείωση των δεδομένων των αμφιωτικών παραμέτρων, καθώς και τα κίνητρα πίσω από αυτή. Έγινε προσπάθεια για την επίτευξη της μέγιστης ευελιξίας, λόγω των δεδομένων, ως προς τις δομές των NN που μπορούσαν να χρησιμοποιηθούν, αφού μεγάλα παραδείγματα εισόδου, εισάγουν περιορισμούς ως προς το πλήθος των κρυφών επιπέδων που μπορούν να χρησιμοποιηθούν, αλλά και το πλήθος των νευρώνων σε κάθε ένα από αυτά. Η συμπίεση συμβαίνει σε δύο στάδια. Στο πρώτο, τα δεδομένα συμπιέζονται με βάση με βάση ψυχοακουστικά μοντέλα ως προς την αντίληψη του ήχου, που υποδεικνύουν τις σημαντικές μπάντες κάθε παραμέτρου, ενώ στο δεύτερο εφαρμόζεται ο αλγόριθμος συμπίεσης που σχεδιάστηκε, για την εξαγωγή των 'προφίλ' των παραμέτρων. Συνοπτικά η συμπίεση περιγράφεται στο Σχήμα 3.11, όπου με P σημειώνεται η παράμετρος ενδιαφέροντος, N_{dim} οι αρχικές διαστάσεις της, και με N'_{dim} οι μειωμένες διαστάσεις. Τα στοιχεία του διαγράμματος αναλύονται περαιτέρω στις επόμενες υποενότητες.



Σχήμα 3.11: Προτεινόμενη μέθοδος συμπίεσης για τις αμφιωτικές παραμέτρους.

3.4.1 Κίνητρο

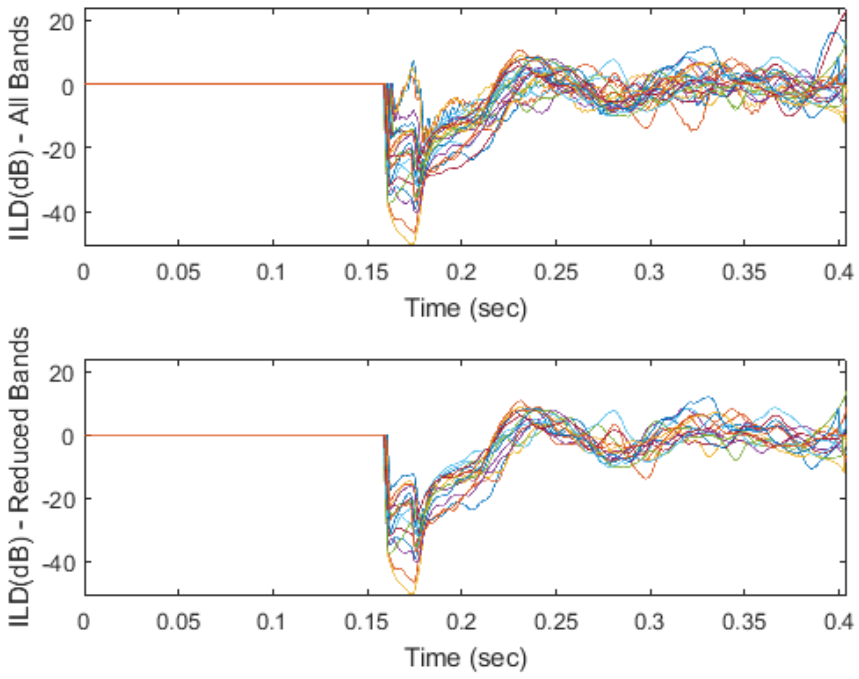
Για να γίνει σαφές το σκεπτικό πίσω από την επιλογή για τη συμπίεση των δεδομένων, είναι απαραίτητο να γίνει σαφές το μέγεθος των δεδομένων που το ακουστικό μοντέλο δίνει ως έξοδο. Όπως έχει αναφερθεί, το ILD έχει σε αυτό το σημείο 23 διαστάσεις και το ITD 11, κάθε μία από τις οποίες αντιστοιχεί σε μια μπάντα συχνοτήτων. Κάθε διάσταση έχει τον ίδιο αριθμό δειγμάτων με όλες τις υπόλοιπες, οπότε μπορούμε να πούμε ότι έχουμε συνολικά $23 + 11 = 34$ διαστάσεις δεδομένων. Το ακουστικό μοντέλο εκτελεί τις πράξεις της συνέλιξης, χωρίς να αυξάνει το πλήθος των δειγμάτων, οπότε κάθε διάσταση, όπως έχει αναφερθεί στην ενότητα 3.2, έχει 17819, σημεία. Συνεπώς, προκύπτουν $34 * 17819 = 605846$ δείγματα. Γίνεται αμέσως αντιληπτό, πως το πλήθος των δειγμάτων, καθιστά το διάνυσμα απαγορευτικό για χρήση στην εκπαίδευση ενός νευρωνικού δικτύου, τόσο από άποψη του χρόνου που θα απαιτούνταν για την εκπαίδευση ενός τέτοιου μοντέλου, όσο και από την άποψη των περιορισμένων υπολογιστικών πόρων που είναι διαθέσιμοι.

3.4.2 Αντιληπτική Συμπίεση

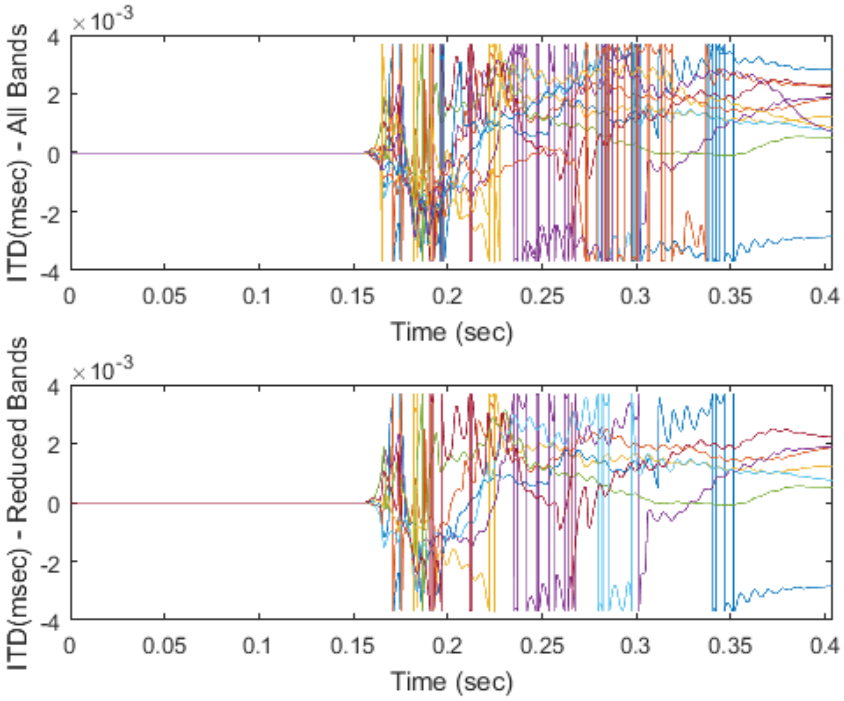
Όπως έχει αναφερθεί στην ενότητα 2.4, αναλόγως με τη συχνότητα, η κάθε παράμετρος αποκτά διαφορετική βαρύτητα στον εντοπισμό ακουστικών πηγών. Με το ILD, να παίζει μεγαλύτερο ρόλο στις συχνότητες που είναι μεγαλύτερες από $1500Hz$, ενώ το ITD το αντίθετο. Με βάση τις εξόδους του μοντέλου, και τη γνώση ότι κάθε μία από τις διαστάσεις των παραμέτρων αντιστοιχεί σε πλάτος 1 ERB, είναι αρκετά εύκολο να γίνει η αντιστοίχηση διαστάσεων-συχνοτικών μπαντών. Σε αυτή την εργασία, το crossover frequency, αντί για $1500Hz$ τέθηκε κοντά στα $2000Hz$, ώστε να υπάρχει μεγαλύτερη ισορροπία στο πλήθος των σημείων του ITD και του ILD. Σημειώνεται και εδώ πως χρησιμοποιείται η envelope δομή του μοντέλου. Πιο συγκεκριμένα, από το ILD διατηρούνται οι διαστάσεις $8 - 23$, ενώ από το ITD οι $1 - 7$, όπως φαίνεται στην Εξίσωση 25. Τα αποτελέ-

σματα της αντιληπτικής συμπίεσης παρουσιάζονται στα Σχήματα 3.12 και 3.13.

$$\begin{aligned} ILD_{new} &= ILD[8:23,:] \\ ITD_{new} &= ITD[1:7,:] \end{aligned} \tag{25}$$



Σχήμα 3.12: Σύγκριση πριν και μετά τη συμπίεση, της παραμέτρου ILD.



Σχήμα 3.13: Σύγκριση πριν και μετά τη συμπίεση, της παραμέτρου ITD.

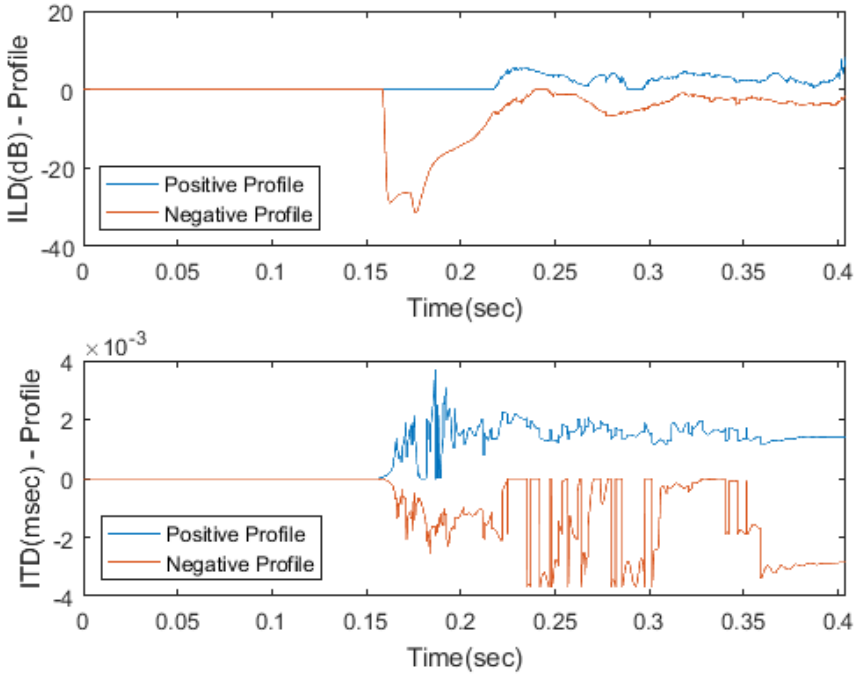
3.4.3 Αλγοριθμική Συμπίεση

Από τις παραμέτρους που προκύπτουν από την αντιληπτική συμπίεση, ο στόχος είναι ο υπολογισμός δύο χαρακτηριστικών καμπυλών, 'προφίλ', με σημαντικά μειωμένο μέγεθος δεδομένων, μία για τις θετικές και μία για τις αρνητικές τιμές κάθε παραμέτρου. Ο αλγόριθμος προσπελαύνει τις παραμέτρους ως προς το L, δηλαδή τη 'μεγάλη' διάσταση, και υπολογίζει δύο διαφορετικά αθροίσματα, ένα για τις θετικές τιμές κάθε διάστασης και ένα για τις αρνητικές, για το δείγμα i, και διαιρεί κάθε ένα από τα αθροίσματα με το πλήθος των στοιχείων που ανατέθηκαν σε αυτό, όπως φαίνεται στις Εξισώσεις 26 και 27. Στη συνέχεια, τα δύο διανύσματα, συνδυάζονται στο τελικό προφίλ της παραμέτρου $Profile[2, L]$, όπως φαίνεται στην Εξίσωση 28. Τυπικά αποτελέσματα του αλγορίθμου φαίνονται στο Σχήμα 3.14. Ο αλγόριθμος είναι απλός στην υλοποίηση και γρήγορος στην εκτέλεση, οπότε ενδείκνυται για την επεξεργασία ιδιαίτερα μεγάλων dataset. Ο λόγος συμπίεσης του αλγορίθμου δίνεται στην Εξίσωση 29 και στη συγκεκριμένη εφαρμογή είναι $CR = 88.24\%$. Τα αποτελέσματα των μοντέλων που εκπαιδεύτηκαν με τα δεδομένα που έχουν επεξεργαστεί από τον προτεινόμενο αλγόριθμο είναι με διαφορά καλύτερα σε σχέση με άλλες μεθόδους επεξεργασίας και θεωρείται πως αυτό συμβαίνει διότι διατηρούνται τα σημαντικότερα χαρακτηριστικά των αμφιωτικών σημάτων, που πιστεύεται πως είναι οι κορυφές που προκύπτουν στο onset / offset του burst.

$$P_{pos} = \frac{\sum_{i:P_{i,j}>0} P_{i,j}}{\sum_{i:P_{i,j}>0} 1} \tag{26}$$

$$P_{neg} = \frac{\sum_{i:P_{i,j}<0} P_{i,j}}{\sum_{i:P_{i,j}<0} 1} \tag{27}$$

$$Profile[2, L] = (P_{pos}, P_{neg}) \tag{28}$$



Σχήμα 3.14: Προφίλ αμφιωτικών παραμέτρων: (Πάνω): ILD, (Κάτω): ITD.

$$CR = \frac{4 * L}{N'_{dim} * L} = \frac{4}{N'_{dim}} \tag{29}$$

3.5 Προεπεξεργασία Δεδομένων

Το επόμενο στάδιο του συστήματος εκτίμησης DOA, είναι η προεπεξεργασία των συμπιεσμένων αμφιωτικών παραμέτρων, ώστε να μπορούν να χρησιμοποιηθούν για την εκπαίδευση ενός NN. Στη μηχανική μάθηση, η διαδικασία της προεπεξεργασίας, εκτιμάται ότι είναι τόσο σημαντική όσο και η διαδικασία της κατασκευής του μοντέλου. Παρόλα αυτά όμως, λόγω της πρόσφατης άνθησης του τομέα της μηχανικής μάθησης, δεν υπάρχει αρκετή εμπειρία πάνω σε αυτόν και κατ' επέκταση δεν είναι γνωστός ο καλύτερος τρόπος προεπεξεργασίας των δεδομένων, ανάλογα με τον τύπο τους. Αν και υπάρχουν μερικές τεχνικές οι οποίες είναι ευρέως αποδεκτό ότι παρέχουν καλά αποτελέσματα, επί το πλείστον, οι ερευνητές προσεγγίζουν το πρόβλημα μέσω trial and error.

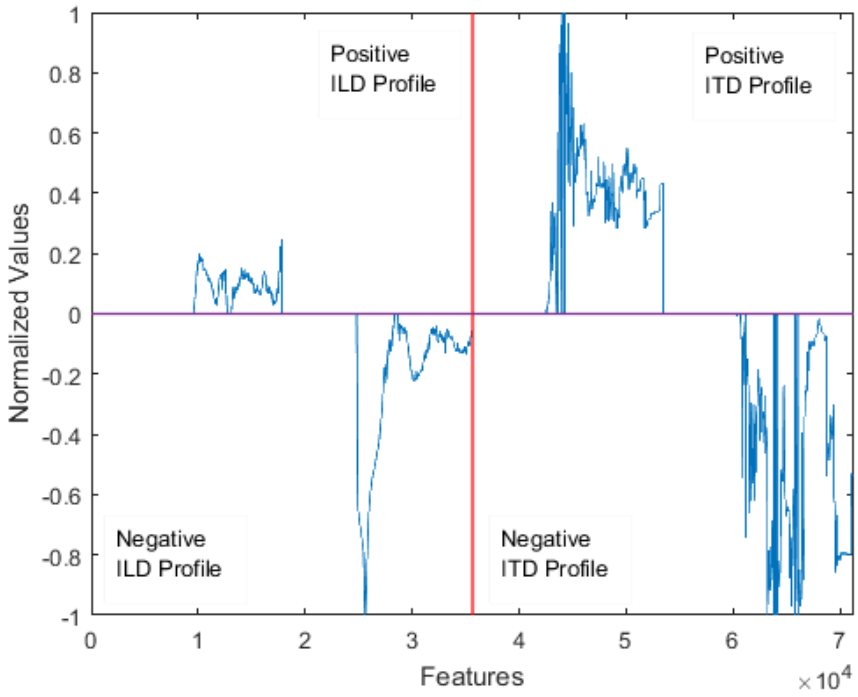
Μια από τις ευρέως γνωστές τεχνικές προεπεξεργασίας δεδομένων, είναι αυτή της κανονικοποίησης των τιμών των παραμέτρων, ώστε να αποφευχθεί ο κορεσμός του δικτύου όταν προκύπτουν πολύ μεγάλα βάρη στις συνάψεις. Με αυτόν τον τρόπο επιταχύνεται επίσης η διαδικασία της εκπαίδευσης του μοντέλου [43, 44]. Εδώ η κανονικοποίηση, έγινε στο κλειστό διάστημα $[-1, 1]$, και κρίθηκε πως είναι απαραίτητη λόγω της μεγάλης διαφοράς τάξεως μεγέθους μεταξύ των τιμών της παραμέτρου ILD, η οποία μετριέται σε dB, και της παραμέτρου ITD, η οποία μετριέται σε msec. Με αυτόν τον τρόπο υποδεικνύεται στο μοντέλο, ότι οι παράμετροι πρέπει να αντιμετωπιστούν με την ίδια βαρύτητα.

Τα προφίλ των παραμέτρων, με διαστάσεις $[2, L]$, μετασχηματίζονται σε διανύσματα γραμμής, με διαστάσεις $[1, 2L]$, όπως φαίνεται στην Εξίσωση 30. Τα δύο διανύσματα γραμμής διαστάσεων $[1, 2L]$, τελικά συνδυάζονται στο διάνυσμα που θα αποτελέσει την είσοδο του νευρωνικού, με ένα τυπικό παράδειγμα να φαίνεται στο Σχήμα 3.15. Σημειώνεται πως

τα δεδομένα μετά την προεπεξεργασία χάνουν τη φυσική σημασία τους.

$$\begin{aligned} Input\ Vector[1:L] &= Profile[1,L] \\ Input\ Vector[L+1:2L] &= Profile[2,L] \end{aligned} \tag{30}$$

Από τα 7917 διαφορετικά διανύσματα που υπολογίζονται, το 80% χρησιμοποιείται για την εκπαίδευση του νευρωνικού, το 10% χρησιμοποιείται για validation, και το υπόλοιπο 10% για testing. Κάθε διάνυσμα είναι μοναδικό, και αντιστοιχεί σε μία μετάθεση DOA-σήματος εισόδου-δωματίου. Η διαδικασία του διαχωρισμού σε train-validation-test data γίνεται με τυχαίο τρόπο, ώστε τα αποτελέσματα να είναι αξιόπιστα και γενικεύσιμα.



Σχήμα 3.15: Τυπικό παράδειγμα κανονικοποιημένου διανύσματος εισόδου στο NN από τις αμφιωτικές παραμέτρους για συγκεκριμένη γωνία άφιξης.

3.6 Αρχιτεκτονικές Νευρωνικών Δικτύων

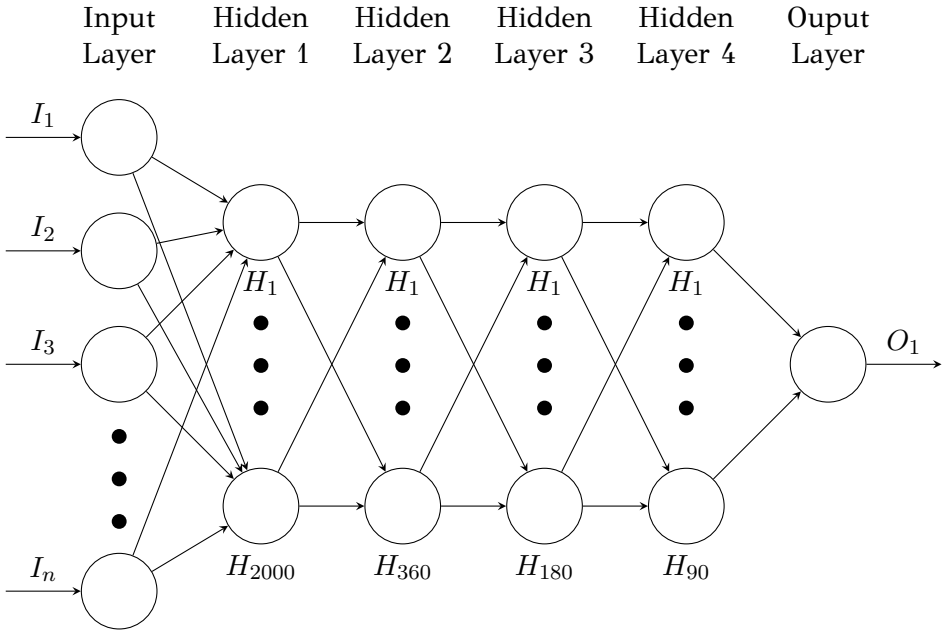
Ο όρος *'αρχιτεκτονική ANN'* αναφέρεται στην διάταξη των νευρώνων σε επίπεδα, τις συνδέσεις μεταξύ των επιπέδων, τις συναρτήσεις ενεργοποίησης και τις μεθόδους μάθησης [45]. Απλούστερα, γίνεται σαφές ότι αναφέρεται στο σύνολο της κατασκευής ενός NN. Το *μοντέλο* του NN, και η αρχιτεκτονική του, καθορίζουν τον τρόπο που η είσοδος παράγει με υπολογιστικό τρόπο μια έξοδο. Η βασική λειτουργία που πρέπει να ακολουθηθεί για την σωστή αντιμετώπιση ενός προβλήματος, είναι όπως φαίνεται, αυτή της επιλογής της σωστής αρχιτεκτονικής καθώς και των υπερπαραμέτρων που αναλύονται στην υποενότητα 2.8.2. Εξίσου σημαντική, είναι η επιλογή της αντικειμενικής συνάρτησης, η οποία συχνά αποκαλείται και συνάρτηση απώλειας (loss function), που είναι επιθυμητό να ελαχιστοποιηθεί (ή να μεγιστοποιηθεί αναλόγως με το πρόβλημα). Σε αυτή την εργασία, τα μοντέλα προσπαθούν να ελαχιστοποιήσουν το *Μέσο Τετραγωνικό Σφάλμα* (MSE), ενώ κατά την εκπαίδευσή τους παρακολουθείται το Μέσο Απόλυτο Σφάλμα (MAE) και η Ρίζα του Μέσου Τετραγωνικού Σφάλματος (RMSE). Δοκιμάστηκε η απόδοση αρκετών διαφορετικών μοντέλων, εδώ όμως αναλύονται τα καλύτερα εκ των δύο ευρύτερων κατηγοριών, των πλήρως διασυνδεδεμένων (Fully Connected), και των συνελικτικών (Convolutional). Εδώ σημειώνεται πως η προεπεξεργασία των δεδομένων, καθώς και τα ίδια τα μοντέλα έχουν υλοποιηθεί με τη βοήθεια της βιβλιοθήκης TensorFlow στη γλώσσα Python. Στο πακέτο αυτό προστέθηκαν μερικές custom συναρτήσεις για την λεπτομερέστερη παρακολούθηση του χρόνου εκτέλεσης.

3.6.1 Fully Connected

Το πλήρως διασυνδεδεμένο μοντέλο που έδωσε τα καλύτερα αποτελέσματα, απεικονίζεται στο Σχήμα 3.16, και αποτελείται από ένα επίπεδο εισόδου με 71276 νευρώνες, τέσσερα κρυφά επίπεδα με 2000, 360, 180 και

90 νευρώνες αντίστοιχα, καθώς και το επίπεδο εξόδου, που έχει έναν νευρώνα που δίνει την πρόβλεψη. Μετά το επίπεδο εισόδου, καθώς και μετά από το 1ο κρυφό επίπεδο, τοποθετείται ένα επίπεδο Dropout, το οποίο δεν αποτελείται από νευρώνες, αλλά ελέγχει τους νευρώνες του προηγούμενου επιπέδου. Όλα τα επίπεδα που περιέχουν νευρώνες, χρησιμοποιούν την ReLU (εξίσωση 18) ως συνάρτηση ενεργοποίησης.

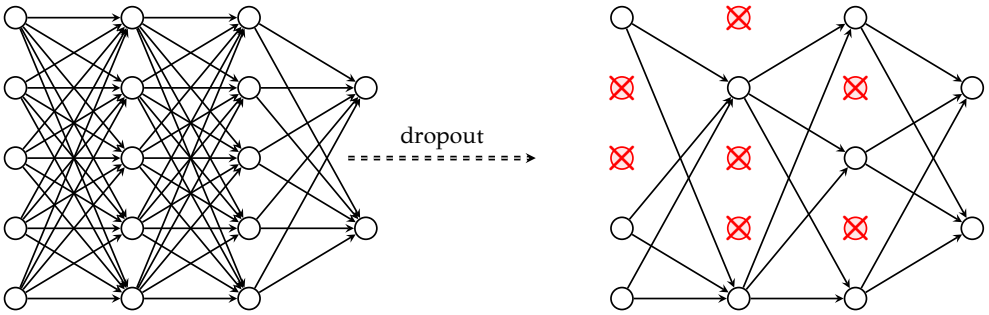
Αναλυτικότερα, η εκπαίδευση έγινε χρησιμοποιώντας, όπως έχει ήδη αναφερθεί, ως συνάρτηση απώλειας το MSE, και τον βελτιστοποιητή Adam (Adaptive Moment Estimation) με αρχικό ρυθμό μάθησης $Lr = 0.001$.



Σχήμα 3.16: Fully connected χωρίς τα επίπεδα Dropout.

Επίπεδο Dropout

Με απλά λόγια, η λειτουργία του επιπέδου Dropout, είναι η απενεργοποίηση, με τυχαίο τρόπο, νευρώνων του προγηούμενου επιπέδου. Το επίπεδο αυτό, έχει μία υπερπαράμετρο, το *Dropout Rate* (DR), το οποίο καθορίζει το πλήθος των νευρώνων που απενεργοποιούνται σε κάθε εποχή (ένα πέρασμα ολόκληρου του dataset από το νευρωνικό). Για παράδειγμα,



Σχήμα 3.17: Περιγραφή της λειτουργίας του Dropout Layer.

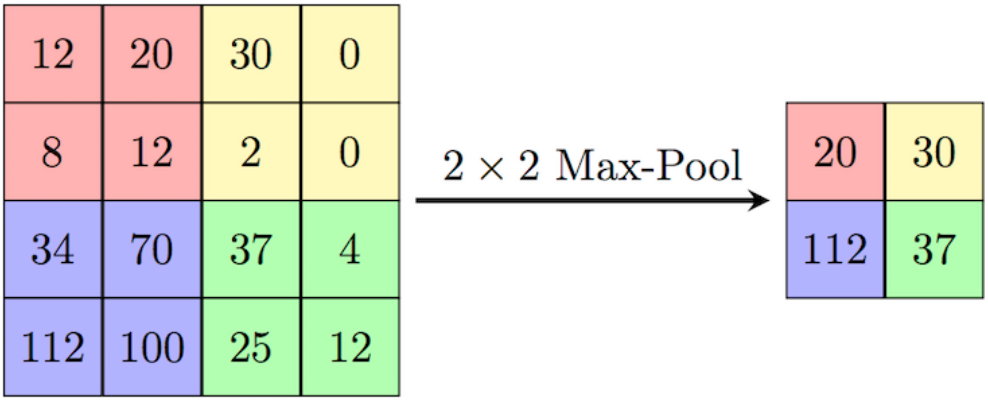
για $DR = 0.4$, με το προηγούμενο επίπεδο να έχει 2000 νευρώνες, σε κάθε εποχή απενεργοποιούνται οι $2000 * 0.4 = 800$ από αυτούς. Το αποτέλεσμα αυτής της διαδικασίας, είναι το νευρωνικό να μην βασίζεται αποκλειστικά σε μερικούς νευρώνες εκ του συνόλου για την τελική εκτίμηση που δίνει στην έξοδο μειώνοντας με αυτόν τον τρόπο σημαντικά το overfitting. Μια συνοπτική περιγραφή της λειτουργίας του επιπέδου, φαίνεται στο Σχήμα 3.17. Σε αυτό το σχήμα θεωρείται πως υπάρχει ένα Dropout Layer ενδιάμεσα από όλα τα επίπεδα.

3.6.2 Convolutional

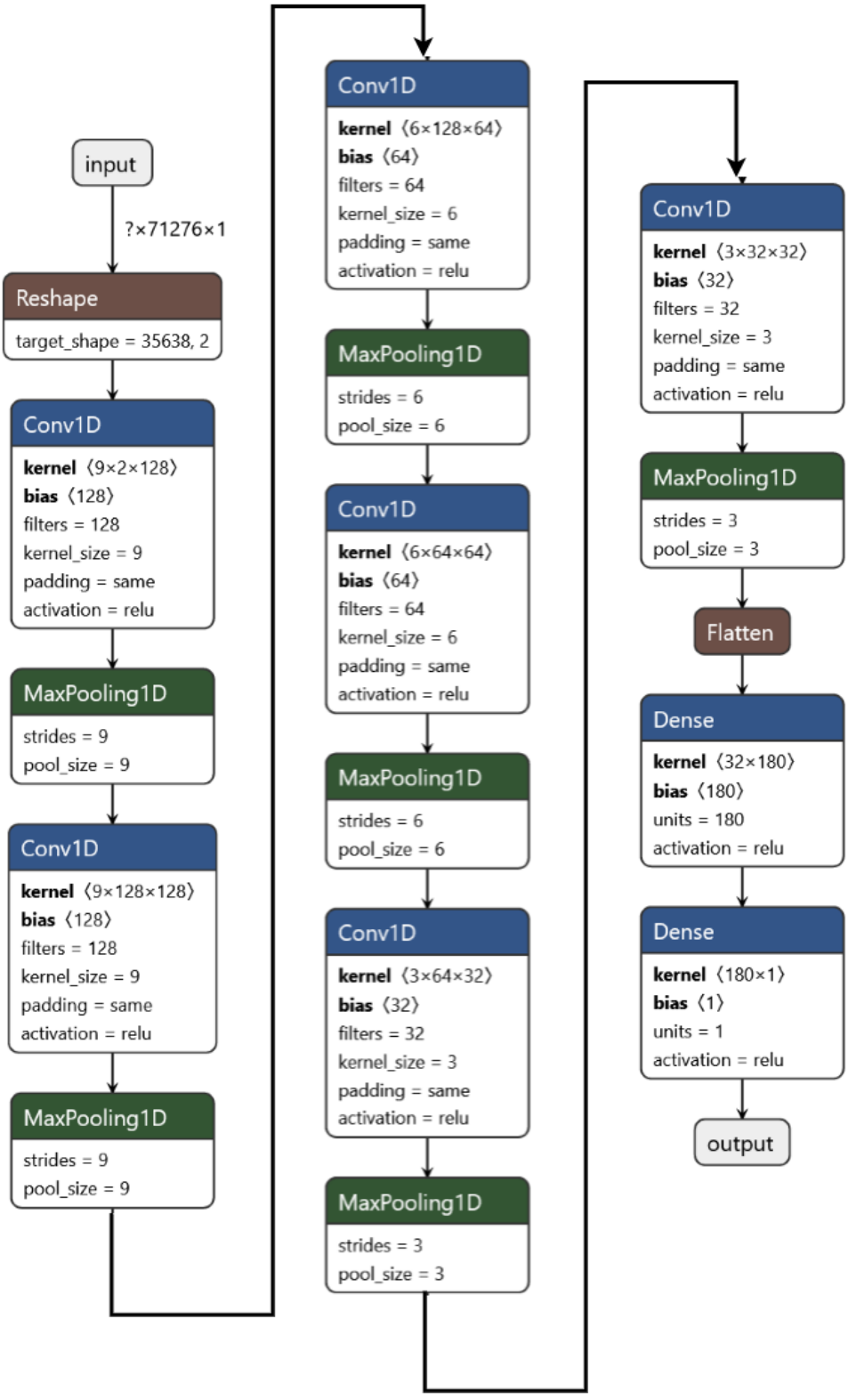
Η CNN αρχιτεκτονική που χρησιμοποιήθηκε σε αυτή την εργασία, αποτελείται από 6 1D-Convolutional επίπεδα, κάθε ένα από τα οποία ακολουθείται από ένα επίπεδο υποδειγματοληψίας, κοινώς γνωστό ως MaxPooling Layer. Πριν το επίπεδο εξόδου υπάρχει ένα πλήρως διασυνδεδεμένο επίπεδο. Όπως και στο προηγούμενο μοντέλο, έτσι και εδώ, όλα τα επίπεδα που περιέχουν νευρώνες, χρησιμοποιούν την ReLU (Εξίσωση 18) ως συνάρτηση ενεργοποίησης. Χρησιμοποιείται επίσης η τεχνική της μείωσης του ρυθμού μάθησης, όταν ο αλγόριθμος φτάνει σε κάποιο *plateau*, δηλαδή το σφάλμα μένει σταθερό για ένα ορισμένο διάστημα. Και σε αυτή την περίπτωση, η εκπαίδευση έγινε χρησιμοποιώντας, όπως έχει ήδη αναφερθεί, ως συνάρτηση απώλειας το MSE, και τον βελτιστοποιητή Adam (Adaptive Moment Estimation) με αρχικό ρυθμό μάθησης $Lr = 0.001$. Το

μοντέλο απεικονίζεται στο Σχήμα 3.19. Τα συνελικτικά επίπεδα, αποτελούνται ανά δύο από 128, 64 και 32 φίλτρα μήκους 3, ενώ το πλήρως διασυνδεδεμένο επίπεδο αποτελείται από 180 νευρώνες.

Η διαδικασία Max Pooling, που αναφέρεται σε αυτή την ενότητα, είναι μια διαδικασία διακριτοποίησης, βασισμένη στα δείγματα. Ο στόχος της είναι η υποδειγματοληψία της αναπαράστασης που δέχεται στην είσοδο, μειώνοντας τις διαστάσεις της. Πρακτικά, αν οριστεί μέγεθος pooling $pool\ size = 3$, τότε το διάνυσμα εισόδου, μήκους L, προσπελάζεται με βήμα $pool\ size$, και από κάθε διάστημα τριών δειγμάτων διατηρείται μόνο το μέγιστο. Το αποτέλεσμα είναι ένα διάνυσμα μήκους $\frac{L}{poolsize}$. Όταν αυτό ανάγεται στις δύο διαστάσεις, γίνεται ευκολότερα κατανοητό, όπως παρουσιάζεται στο Σχήμα 3.18.



Σχήμα 3.18: Αναγωγή του Max Pooling σε δύο διαστάσεις.



Σχήμα 3.19: Η αρχιτεκτονική του μοντέλου CNN για εκτίμηση DOA.

ΚΕΦΑΛΑΙΟ 4

Αποτελέσματα

Στις ενότητες 2 και 3, αναλύθηκε το θεωρητικό υπόβαθρο της παρούσας εργασίας και τα τμήματα που απαρτίζουν το σύστημα εκτίμησης γωνίας άφιξης αντίστοιχα. Σε αυτό το κεφάλαιο, παρουσιάζεται η πορεία της εκπαίδευσης των νευρωνικών δικτύων καθώς και τα αποτελέσματα-προβλέψεις των μοντέλων, σε σύγκριση με άλλα ήδη υπάρχοντα μοντέλα. Παρουσιάζονται επίσης συγκριτικά τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων, οι διαφορετικές αρχιτεκτονικές που χρησιμοποιήθηκαν καθώς και το μέγεθος των διανυσμάτων εισόδου, εφόσον αυτό είναι διαθέσιμο στην εκάστοτε εργασία.

4.1 Αποτελέσματα εκπαίδευσης

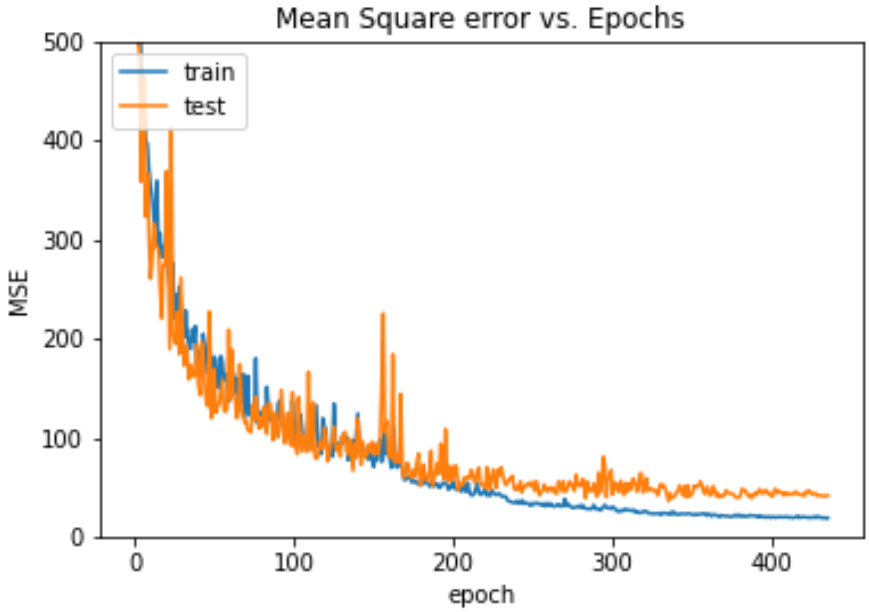
Σε αυτό το σημείο, παρουσιάζονται τα διαγράμματα που δείχνουν την πορεία εξέλιξης της τιμής των αντικειμενικών συναρτήσεων που χρησιμοποιούνται κατά την εκπαίδευση, οι οποίες είναι όπως έχει αναφερθεί το MSE, RMSE και MAE, για τα δύο διαφορετικά μοντέλα που εκπαιδεύτηκαν, το fully connected, και το convolutional.

4.1.1 Πλήρως διασυνδεδεμένη αρχιτεκτονική

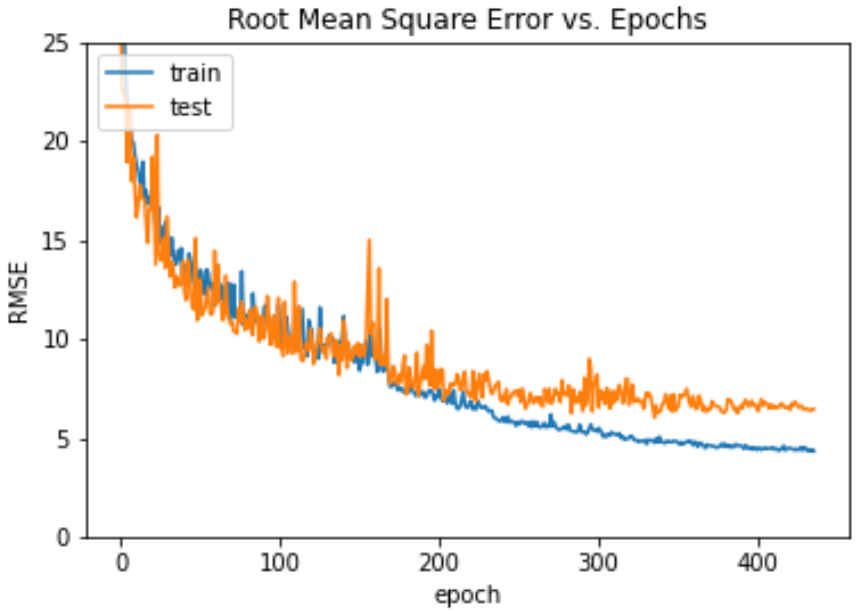
Κατά την διαδικασία της εκπαίδευσης, κάθε νευρώνας εκτελεί ένα σταθμισμένο άθροισμα, το οποίο τελικά δίνεται ως είσοδος στη συνάρτηση ενεργοποίησης, ώστε να παραχθεί η έξοδος του νευρώνα. Οι πράξεις αυτές εκτελούνται γρήγορα σε σχέση με τις πράξεις που είναι απαραίτητες στις άλλες αρχιτεκτονικές, (CNN, CRNN κλπ). Στα διαγράμματα 4.1, 4.2 και 4.3, φαίνονται με τη σειρά η εξέλιξη των τιμών του MSE, RMSE και MAE αντίστοιχα. Τα test data χρησιμοποιούνται κατά το πέρας της εκπαίδευσης, για την τελική αξιολόγηση του μοντέλου και παρουσιάζονται σε επόμενη υποενότητα. Τα διαγράμματα αυτά περιγράφουν ουσιαστικά την *ιστορία* του μοντέλου.

Το μοντέλο, ξεκινά με τελείως τυχαίες προβλέψεις, που αιτιολογούν και το εξαιρετικά μεγάλο σφάλμα κατά την αρχή της εκπαίδευσης, ενώ στη συνέχεια, *μαθαίνει*, να βγάζει συμπεράσματα από τα χαρακτηριστικά των διανυσμάτων εισόδου. Με τον όρο *epoch* στον άξονα X, υποδηλώνεται ένα **πλήρες** πέρασμα του dataset από τον αλγόριθμο. Το μοντέλο ξεκινά να συγκλίνει περίπου μετά από 350 εποχές, όπως φαίνεται από τη συνάρτηση απώλειας MSE, όπου το σφάλμα παραμένει σχετικά σταθερό. Σε εκείνο το σημείο αξιοποιείται η τεχνική Early Stopping, όπου όταν το σφάλμα δεν μειώνεται για έναν προκαθορισμένο αριθμό εποχών, τότε τερματίζεται η διαδικασία της εκπαίδευσης για την αποφυγή σπατάλης υπολογιστικών πόρων και προφανώς, χρόνου.

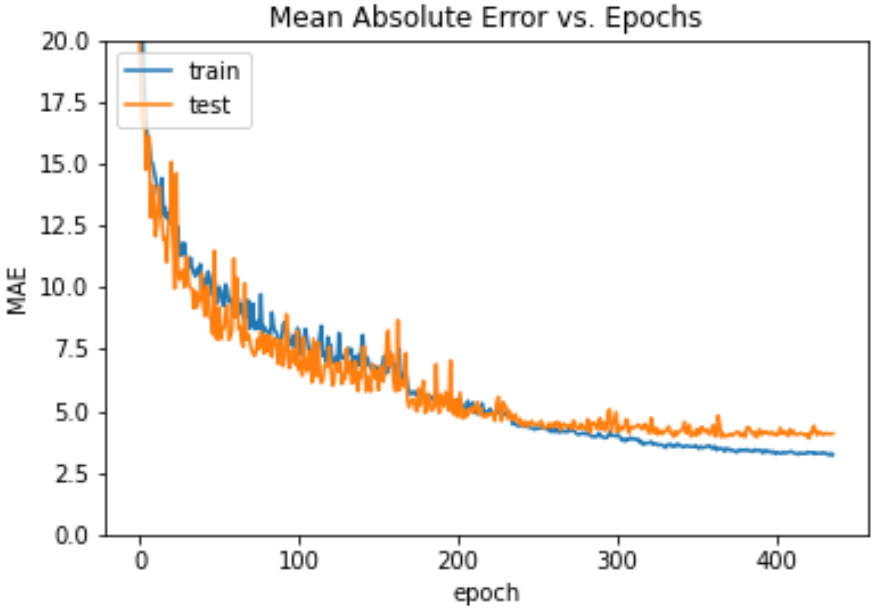
Γίνεται επίσης εμφανές, το γεγονός ότι οι καμπύλες που δείχνουν την απόδοση στα training και στα validation data απέχουν κάποια απόσταση μεταξύ τους. Αυτό είναι αναμενόμενο, και ο λόγος που η καμπύλη που αφορά τα δεδομένα εκπαίδευσης είναι πάντα κάτω από την καμπύλη του validation, είναι το γεγονός ότι τα νευρωνικά δίκτυα, και ιδιαίτερα οι fully connected αρχιτεκτονικές, έχουν την τάση να 'απομνημονεύουν' τα δεδομένα εισόδου, με αποτέλεσμα να έχουν χειρότερη απόδοση σε δεδομένα που 'βλέπουν' πρώτη φορά. Η περίπτωση στην οποία η απόσταση μεταξύ των δύο καμπυλών μεγαλώνει όσο περνάνε οι εποχές λέγεται ***overfitting*** και τα επίπεδα Dropout, έχουν χρησιμοποιηθεί για την αντιμετώπισή του.



Σχήμα 4.1: Ιστορικό τιμών του MSE για την Fully Connected αρχιτεκτονική.



Σχήμα 4.2: Ιστορικό τιμών του RMSE για την Fully Connected αρχιτεκτονική.

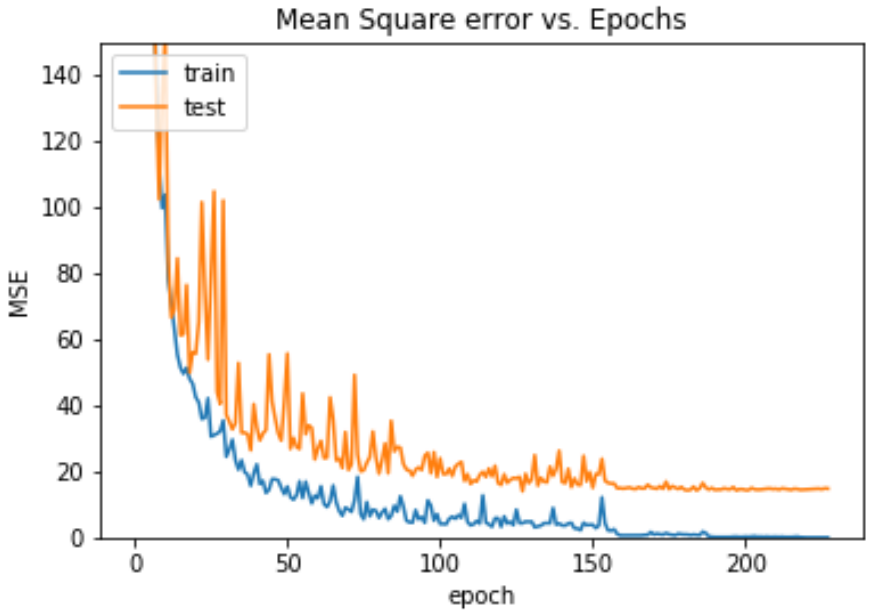


Σχήμα 4.3: Ιστορικό τιμών του MAE για την Fully Connected αρχιτεκτονική.

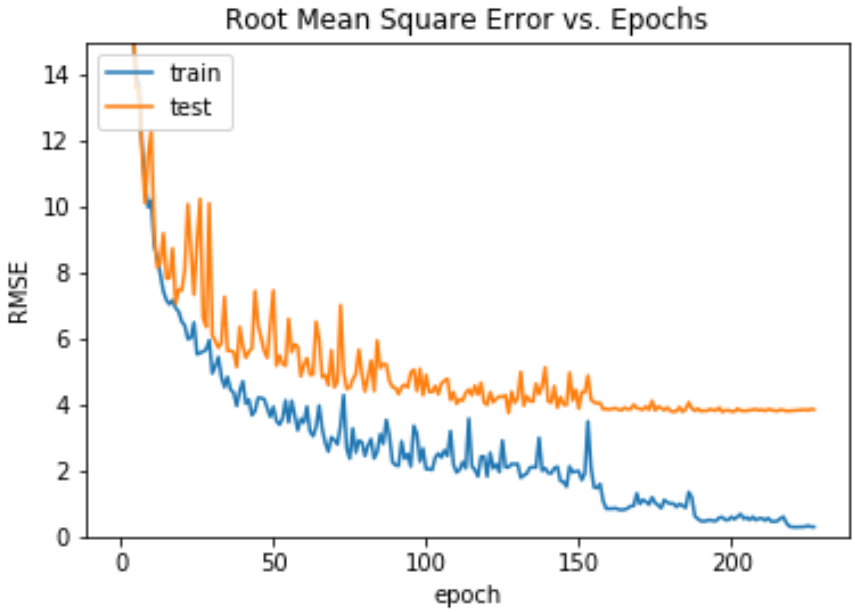
4.1.2 Συνελικτική αρχιτεκτονική

Τα αντίστοιχα αποτελέσματα για την convolutional δομή, παρουσιάζονται σε αυτή την υποενότητα, στα Σχήματα 4.4, 4.5 και 4.6 για τα MSE, RMSE και MAE αντίστοιχα. Αξίζει να σημειωθεί το γεγονός, πως το CNN, χρειάζεται τις μισές περίπου εποχές για να φτάσει στο τοπικό ελάχιστο της συνάρτησης, χωρίς αυτό να σημαίνει όμως ότι συγκλίνει γρηγορότερα από άποψη χρόνου. Η πράξη της συνέλιξης είναι ιδιαίτερα αργή από υπολογιστική άποψη, και αυτός είναι ο κυριότερος λόγος για αυτό. Παρατηρείται επίσης, πως οι καμπύλες έχουν μεγαλύτερη κλίση στην αρχή της εκπαίδευσης.

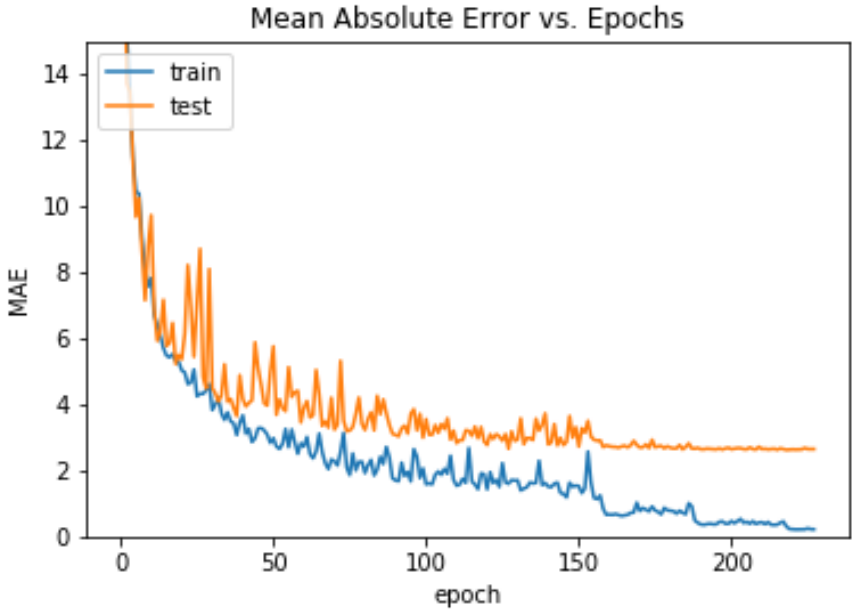
Τα CNN είναι πιο εύρωστα στο φαινόμενο του overfitting, οπότε δεν χρειάστηκε να προστεθούν επίπεδα Dropout για την εκπαίδευση του μοντέλου. Το CNN κάνει λάθος περίπου 3.5^o, στην πρόβλεψη της γωνίας άφιξης, δηλαδή είναι περίπου 50% καλύτερο από την fully connected αρχιτεκτονική που έχει μέσο σφάλμα σχεδόν 6.5^o.



Σχήμα 4.4: Ιστορικό τιμών του MSE για την συνελικτική αρχιτεκτονική.



Σχήμα 4.5: Ιστορικό τιμών του RMSE για την συνελικτική αρχιτεκτονική.



Σχήμα 4.6: Ιστορικό τιμών του MAE για την Fully Connected αρχιτεκτονική.

4.1.3 Χρόνοι σύγκλισης

Εδώ παρουσιάζονται με συνοπτικό τρόπο οι χρόνοι σύγκλισης των δύο μοντέλων στον Πίνακα 4.1. Σημειώνεται πως ο χρόνος που χρειάζονται

και τα δύο μοντέλα για να συγκλίνουν είναι σημαντικά μικρότερος από άλλα μοντέλα που έχουν κατασκευαστεί με παρόμοιο στόχο.

Model	Epochs	Time (mins)
Fully Connected	436	125.36
Convolutional	228	183.36

Πίνακας 4.1: Σύγκριση χρόνων σύκλισης των μοντέλων σε εποχές και λεπτά.

4.2 Αποτελέσματα εκτίμησης

Η απόδοση των μοντέλων αξιολογήθηκε ως προς την ικανότητά τους να εκτιμήσουν την DOA, σε δεδομένα που αντιμετωπίζουν πρώτη φορά. Αυτή ακριβώς είναι η αξία του διαχωρισμού σε train-validation-test δεδομένα, δηλαδή το ότι παρέχεται μια εγγύηση ως προς την αξιοπιστία των δεδομένων. Εκτός από τις συναρτήσεις που έχουν ήδη αναφερθεί (MSE, RMSE, MAE), χρησιμοποιούνται σε αυτό το στάδιο το *Accuracy* ως προς την ταξινόμηση, των μοντέλων σε τρεις κατηγορίες προβλέψεων. Μια DOA θεωρείται σωστά ταξινομημένη εφόσον η διαφορά της προβλεπόμενης γωνίας άφιξης από την πραγματική είναι μικρότερη ή ίση με 5, 10 και 15 μοίρες για τις μετρικές Acc5, Acc10 και Acc15 αντίστοιχα. Για κάθε περιθώριο σφάλματος, το classification accuracy υπολογίζεται όπως φαίνεται στην εξίσωση 31.

$$
\begin{aligned}
Accuracy(\%) &= \frac{100}{n}\sum_{i=1}^{n} d_i \\
\text{όπου } d_i &= \begin{cases} 1 & \text{αν } |y_i - \hat{y}_i| < err \\ 0 & \text{αλλού} \end{cases}
\end{aligned}
\tag{31}
$$

Τα αποτελέσματα των μοντέλων κατά το testing, φαίνονται στον πίνακα 4.2. Σημειώνεται πως τα 'Test Noise Signals' (TNS στον πίνακα) είναι τα burst θορύβου που δημιουργήθηκαν για τους σκοπούς αυτής της εργασίας. Στην περίπτωση του CNN, δοκιμάστηκε η απόδοσή του στην εκτίμηση της γωνίας άφιξης, όταν το σήμα εισόδου είναι κάτι τελείως διαφορετικό από αυτά που έχει δει στην εκπαίδευση, όπως φωνή ή μουσική. Το CNN φαίνεται πως γενικεύει εξαιρετικά σε νέα δεδομένα, επιτυγχάνοντας ένα ελάχιστο classification accuracy 79%, και μέγιστο 85%.

Είναι εμφανές πως το CNN, έχει μακράν καλύτερα αποτελέσματα από το αντίστοιχο Fully Connected μοντέλο και για αυτόν τον λόγο κρίθηκε πως είχε νόημα να δοκιμαστεί η απόδοσή του σε διαφορετικές συνθήκες.

Dataset	MSE	RMSE	MAE	Acc5(%)	Acc10(%)	Acc15(%)
Fully Connected Architecture						
TNS	40.5	6.4	4	89	97	98
Convolutional Architecture						
TNS	11.3	3.4	2.3	96	99	100
Voice	82.8	9.1	7.0	79	89	95
Bongo	212.4	14.6	10.5	80	90	96
Cello	123.0	11.1	9.1	88	95	98
Guitar	130.7	11.4	9.3	87	94	97
Xyloph.	212.4	14.6	10.5	80	90	96
CNN Mean	128.8	10.7	8.1	85	93	97

Πίνακας 4.2: Αποτελέσματα εκτίμισης γωνίας άφιξης για διαφορετικά σήματα διέγερσης.

Σημειώνεται πως οι BRIR που χρησιμοποιούνται σε αυτή την εργασία προκύπτουν από πραγματικά δωμάτια, ενώ στα περισσότερα άλλα συστήματα εντοπισμού γωνίας άφιξης, χρησιμοποιούνται συνήθως simulated δωμάτια.

4.3 Σύγκριση με άλλες μεθόδους

Η υποενότητα αυτή επικεντρώνεται στη σύγκριση της προτεινόμενης μεθόδου με άλλες δημοσιευμένες μεθόδους εκτίμησης DOA που χρησιμοποιούν τεχνικές μηχανικής μάθησης. Για το μοντέλο αυτής της εργασίας χρησιμοποιούνται οι μέσες τιμές κάθε μετρικής που προκύπτουν από εκτιμήσεις του CNN, όπως φαίνονται στον πίνακα 4.2, ενώ στα άλλα μοντέλα χρησιμοποιούνται τα καλύτερα αποτελέσματα που έχουν επιτευχθεί. Τονίζεται πως δεν χρησιμοποιούνται παντού οι ίδιες μετρικές, οπότε αναφέρονται όσες είναι διαθέσιμες. Σε αντίθετη περίπτωση χρησιμοποιούνται best και worst case αποτελέσματα στη θέση των Acc5 και Acc15 αντίστοιχα. Στον πίνακα 4.3 παρουσιάζονται οι συγκρίσεις μεταξύ των μοντέλων.

Model	Acc5(%)	Acc10(%)	Acc15(%)
Proposed Model	85	93	97
Intensity-CRNN (Simulated RIR) [19]	54.3	94.4	98.9
Intensity-CRNN (Real RIR) [19]	26.2	62.6	78.1
DoaNet [19]	59.3	-	95.4
CNN+masking [17]	65.8	-	87.0

Πίνακας 4.3: Σύγκριση DOA Accuracy της προτεινόμενης μεθόδου, με άλλες δημοσιευμένες μεθόδου.

Για τις μεθόδους που συγκρίθηκαν στον πίνακα 4.3, οι πίνακες 4.4 και 4.5 παρουσιάζουν τις διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων και τα datasets που χρησιμοποιήθηκαν για την εκπαίδευσή τους. Ελέγχοντας τους δύο πίνακες ταυτόχρονα, παρατηρείται ότι η προτεινόμενη μέθοδος εξαρτάται από δεδομένα που προέρχονται από πραγματικά δωμάτια, και σε συνδυασμό με την μέθοδο *profiling* που κατασκευάστηκε και αναλύθηκε στο κεφάλαιο 3.4, επιτυγχάνει εξαιρετικά αποτελέσματα ακόμα και με σχετικά μικρό dataset εκπαίδευσης. Ακόμα ένα πλεονέκτημα

της προτεινόμενης μεθόδου είναι η αξιοσημείωτη ταχύτητα σύγκλισης, ολοκληρώνοντας της διαδικασία της εκπαίδευσης σε 3.5 ώρες (για το μοντέλο CNN το οποίο απαιτεί και τον περισσότερο χρόνο).

Model	Architecture	R/C	NN Inputs
Proposed Model	1D-CNN	R	ILD+ITD Profiles
Intensity-CRNN	2D-CRNN	C	Acoustic Intensity Vectors
DoaNet	2D-CRNN	C	Magnitude + Phase Spectrograms
CNN+masking	CRNN	C	Magnitude Spectrograms

Πίνακας 4.4: Σύγκριση των μοντέλων ως προς την αρχιτεκτονική που χρησιμοποιήθηκε και τις εισόδους. Με R/C σημειώνεται αν το νευρωνικό ήταν τύπου regression ή classification.

Model	Signals	Real Rooms	Training Vectors
Proposed Model	Noise Burst	Yes	6334
Intensity-CRNN	Bref Corpus	No	127800
DoaNet	Real-life sound events	No	-
CNN+masking	TIMIT+ChiME3	No	24000

Πίνακας 4.5: Σύγκριση των μοντέλων ως προς τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευσή τους.

4.4 Σύγκριση αποτελεσμάτων με και χωρίς συμπίεση

Για την επιβεβαίωση της καλής λειτουργίας του αλγορίθμου συμπίεσης, συγκρίθηκαν τα αποτελέσματα με και χωρίς τη χρήση του. Εκπαιδεύτηκαν δηλαδή μοντέλα εκ νέου, χωρίς να συμπιεστούν τα δεδομένα. Χρησιμοποιώντας στη συνέχεια τις ίδιες αρχιτεκτονικές τα μοντέλα εκπαιδεύτηκαν με τα συμπιεσμένα δεδομένα. Παρατίθενται τα αποτελέσματα του παραπάνω πειράματος, το οποίο εκτελέστηκε για fully connected και για convolutional δομές στους Πίνακες 4.6 και 4.7.

Fully Connected architecture					
	MSE	MAE	Acc5(%)	Acc10(%)	Time(min)
Compressed	19.9	3.4	89	99	10.93
Uncompressed	49.2	5.3	83	96	41.78

Πίνακας 4.6: Σύγκριση accuracy και χρόνων στις fully connected αρχιτεκτονικές για συμπιεσμένα και ασυμπίεστα δεδομένα.

Convolutional architecture					
	MSE	MAE	Acc5(%)	Acc10(%)	Time(min)
Compressed	10.9	2.5	95	100	5.48
Uncompressed	6.4	2.0	98	100	19.76

Πίνακας 4.7: Σύγκριση accuracy και χρόνων στις convolutional αρχιτεκτονικές για συμπιεσμένα και ασυμπίεστα δεδομένα.

Αμέσως γίνεται εμφανές ότι και σε αυτή την περίπτωση οι συνελικτικές αρχιτεκτονικές δίνουν πολύ καλύτερα αποτελέσματα. Φαίνεται επίσης εδώ πως συγκλίνουν ταχύτερα από τις πλήρως διασυνδεδεμένες, όμως αυτό συμβαίνει διότι για να μπορέσει το μοντέλο να χωρέσει στη μνήμη περιορίστηκαν σημαντικά ο αριθμός των επιπέδων και των εκάστοτε φίλτρων σε αυτά.

Η προτεινόμενη μέθοδος συμπίεσης επιτυγχάνει ταχύτερη σύγκλιση και πολύ καλά αποτελέσματα, ακόμα και στην περίπτωση όμως που τα

ασυμπίεστα δεδομένα δίνουν καλύτερα αποτελέσματα στην περίπτωση του CNN, η διαφορά είναι εξαιρετικά μικρή. Επίσης λόγω του μικρού μεγέθους των δεδομένων εισόδου, ειδικά όταν αυτά είναι συμπιεσμένα, το μοντέλο εξάγει τα αποτελέσματα σε μόλις $7msec$, μετά το στάδιο εξαγωγής των αμφιωτικών παραμέτρων.

ΚΕΦΑΛΑΙΟ 5

Συμπεράσματα

Οι μέθοδοι για την εκτίμηση γωνίας άφιξης βασίζονται όλο και περισσότερο σε τεχνικές που χρησιμοποιούν μηχανική μάθηση, κυρίως νευρωνικά δίκτυα, που μέχρι στιγμής χρησιμοποιούν διαφορετικές προσεγγίσεις για τα δεδομένα εκπαίδευσης, τις αρχιτεκτονικές και μετρικές απόδοσης. Τέτοιες μέθοδοι συνήθως χρησιμοποιούν δεδομένα από φασματογραφήματα πλάτους ή/και φάσης.

Μια βασική απαίτηση όλων αυτών των μεθόδων είναι η αποτελεσματική χρήση των εξαχθέντων παραμέτρων από κατάλληλα σήματα, για εύρωστη και χαμηλής πολυπλοκότητας εκπαίδευση των νευρωνικών δικτύων. Τυπικά τέτοιες παράμετροι είναι οι αμφιωτικές παράμετροι ILD και ITD.

Σε αυτή την εργασία χρησιμοποιείται ένα fully connected και ένα convolutional νευρωνικό που χρησιμοποιούν μια νέα προσέγγιση για τη συμπίεση των εξαχθέντων παραμέτρων, που επιτρέπει την αποτελεσματική αναπαράστασή των ILD και ITD, για τη γρήγορη και αξιόπιστη εκπαίδευσή των μοντέλων λαμβάνοντας υπόψιν ταυτόχρονα το γεγονός ότι

οι υπολογιστικοί πόροι και τα dataset είναι συνήθως περιορισμένα. Κατά την λειτουργία σε πραγματικές συνθήκες, όπως παραμέτροοι οι οποίες εξάγονται από ακουστικά σήματα που παράγονται μέσα σε δωμάτια με αντήχηση, τότε το μέγεθος των παραμέτρων καθίσταται απαγορευτικό. Με κίνητρο αυτό το γεγονός, η εργασία αυτή εισάγει μια καινούρια μέθοδο προεπεξεργασίας των αμφιωτικών παραμέτρων, η οποία απλοποιεί τα αρχικά πολυδιάστατα δεδομένα σε μόνο δύο διαστάσεις, που αποτελούν μια συνοπτική, οπτική περιγραφή τους, τα προφίλ.

Κατά τη φάση του testing, χρησιμοποιήθηκαν δεδομένα που προέρχονται από πραγματικά δωμάτια με αντήχηση και τα μοντέλα επιτυγχάνουν αποτελέσματα με υψηλή ακρίβεια, από σήματα εισόδου που είναι μόλις $200msec$. Η ακρίβεια παραμένει υψηλή ακόμα και όταν τα σήματα είναι διαφορετικά από αυτά που χρησιμοποιήθηκαν κατά την εκπαίδευση των μοντέλων.

Οι μέχρι στιγμής δοκιμές έχουν δώσει αισιόδοξα αποτελέσματα, επιβεβαιώνοντας πως η προτεινόμενη μέθοδος προσέγγισης του προβλήματος της εκτίμησης DOA, λειτουργεί ακόμα και στην περίπτωση πολύ χαμηλότερης δειγματοληψίας, όπου τα δεδομένα είναι ακόμα πιο περιορισμένα. Τα αποτελέσματα σε αυτή την περίπτωση ορισμένες φορές ξεπερνούν αυτά που έχουν παρουσιαστεί σε αυτή την εργασία.

Ο έλεγχος της απόδοσης της μεθόδου στην περίπτωση που τα δεδομένα προέρχονται από ένα δωμάτιο που δεν έχει χρησιμοποιηθεί κατά την εκπαίδευση αποτελεί αντικείμενο μελλοντικής μελέτης, όπως επίσης και η τροφοδότηση του μοντέλου με περισσότερα δωμάτια, καθώς και διαφορετικές θέσης του ακροατή μέσα σε αυτά για την επίτευξη πιο εύρωστης λειτουργίας.

Κατάλογος σχημάτων

Κατάλογος πινάκων

Βιβλιογραφία

[1] M. Nicoletti, C. Wirtz, and W. Hemmert, The Technology of Binaural Listening, ch. Modeling Sound Localization with Cochlear Implants, pp. 309–331. Springer, 2013.

[2] J. Aronoff, Y. Yoon, D. Freed, A. Vermiglio, I. Pal, and S. Soli, "The use of interaural time and level difference cues by bilateral cochlear implant users," The Journal of the Acoustical Society of America, vol. 127, no. 3, pp. 87–92, 2010.

[3] V. Grimaldi, H. Lissek, G. Courtois, E. Georganti, and P. Estoppey, "Auditory externalization in hearing-impaired listeners with remote microphone systems for hearing aids," International Congress on Sound and Vibration, 2019.

[4] L. Tahmid, W. Eric, N. Tristan, and B. Alper, "Sound localization sensors for search and rescue biobots," IEEE Sensors Journal, vol. 16, no. 10, pp. 3444–3453, 2010.

[5] B.-W. Chen, C.-Y. Chen, and J.-F. Wang, "Smart homecare surveillance system: Behavior identification based on state-transition support vector machines and sound directivity pattern analysis," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 43, no. 6, pp. 1279–1289, 2013.

[6] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 34–43, 2012.

[7] J. Dibiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," pp. 157–180, 2001.

[8] D. Johnson and D. Dudgeon, "Array signal processing-concepts and techniques," 1993.

[9] S. Haykin, "Adaptive filter theory," 1991.

[10] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," EURASIP Joirnal on Advances in Signal Processing, p. 26503, 2006.

[11] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 4, pp. 320–327, 1976.

[12] R. S. and N. P., Artificial Intelligence: A Modern Approach. Pearson Educations, 2010.

[13] M. Abadi, "Large-scale machine learning on heterogeneous systems," 2015.

[14] G. Kamaris, S. Karlos, N. Fazakis, S. Terpinas, and J. Mourjopoulos, "Binaural auditory feature classification for stereo image evaluation in listening rooms," 2016.

[15] G. Kamaris, S. Karlos, S. Terpinas, D. Koutsaidis, and J. Mourjopoulos, "Audio system spatial image evaluation via binaural feature classification," 2017.

[16] G. Kamaris and J. Mourjopoulos, "Stereo image localization maps for loudspeaker reproduction in rooms," 2018.

[17] W. Zhang, Y. Zhou, and Y. Qian, "Robust doa estimation based on convolutional neural network and time-frequency masking," Interspeech, 2019.

[18] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation

for multiple sound sources using convolutional recurrent neural network," 2017.

[19] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," IEEE Journal of Selected Topics in Signal Processing, vol. 13, pp. 22–33, 2019.

[20] . Søndergaard and P. Majdak, "The auditory modeling toolbox," pp. 33–56, 2013.

[21] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[22] M. Dietz, S. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers," Speech Communication, pp. 592–605, May 2011.

[23] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," J. Audio Eng. Soc, vol. 43, no. 5, pp. 300–321, 1995.

[24] G. Enzner, C. Antweiler, and S. Spors, Trends in Acquisition of Individual Head-Related Transfer Functions, pp. 57–92. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[25] D. Hammersho/i and H. Mo/ller, "Sound transmission to and within the human ear canal," The Journal of the Acoustical Society of America, vol. 100, no. 1, pp. 408–427, 1996.

[26] W. Gardner and K. Martin, "Hrtf measurements of a kemar,"

J.Acoust.Soc.Am, vol. 97, pp. 3907–3908, 1995.

[27] J. Blauert, M. Brueggen, A. W. Bronkhorst, R. Drullman, G. Reynaud, L. Pellieux, W. Krebber, and R. Sottek, "The AUDIS catalog of human HRTFs," The Journal of the Acoustical Society of America, vol. 103, pp. 3082–3082, may 1998.

[28] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics, (New Paltz, NY), pp. 99–102, Mohonk Mountain House, Oct. 21-24, 2001.

[29] A. Tsilfidis, A. Westermann, J. M. Buchholz, E. Georganti, and J. Mourjopoulos, The Technology of Binaural Listening, ch. Binaural Dereverberation, pp. 359–396. Springer, 2013.

[30] A. Kohlrausch, J. Braasch, D. Kolossa, and J. Blauert, ch. An Introduction to Binaural Processing, pp. 1–32. Berlin, Heidelberg: Springer, 2013.

[31] L. Jeffress, "A place theory of sound localization," J. Comp. Physiol. Psychol, vol. 41, pp. 35–39, 1948.

[32] E. Cherry and B. Sayers, "Human 'cross-correlator' "-a technique for measuring certain parameters of speech perception," J. Acoust. Soc. Am, vol. 28, no. 5, pp. 889–895, 1956.

[33] M. Reed and J. Blum, "A model for the computation and encoding of azimuthal information by the lateral superior olive," J Acoust. Soc. Am, vol. 88, pp. 1442–1453, 1990.

[34] M. Dietz, S. D. Ewert, V. Hohmann, and B. Kollmeier, "Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences," Brain Research, vol. 1220, pp. 234–245, jul 2008.

[35] C. M. and R. Bruce, "Op amps for everyone," tech. rep., 2009.

[36] S. J. Russell and P. Norvig, "Artificial intelligence: A modern approach," 1994.

[37] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities.," Proceedings of the National Academy of Sciences, vol. 79, pp. 2554–2558, apr 1982.

[38] Y.-Y. Chen, Y.-H. Lin, C.-C. Kung, M.-H. Chung, and I.-H. Yen, "Design and implementation of cloud analytics-assisted smart power meters considering advanced artificial intelligence as edge analytics in demand-side management for smart homes," Sensors, vol. 19, p. 2047, may 2019.

[39] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. The MIT Press, 2017.

[40] H. Wierstorf and M. Geier, "Binaural room impulse responses recorded with kemar in a mid-size lecture hall [data set," 2016.

[41] H. Wierstorf and M. Geier, "Binaural room impulse responses recorded with kemar in a small meeting room [data set," 2016.

[42] H. Wierstorf, "Binaural room impulse responses of a 5.0 surround setup for different listening positions [data set," 2016.

[43] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," IEEE Transactions on Nuclear Science, vol. 44, no. 3, pp. 1464–1468, 1997.

[44] Y. Lecun, L. Bottou, G. Orr, and K. Müller, "Efficient backprop," Lecture Notes in Computer Science, pp. 9–48, 2012.

[45] S. A. Kalogirou, "Designing and modeling solar energy systems," in Solar Energy Engineering, pp. 583–699, Elsevier, 2014.

Πανεπιστήμιο Πατρών, Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών
Παναγιώτης Ζάχος του Σωτηρίου