ΠΟΛΥΚΑΝΑΛΙΚΗ ΑΚΟΥΣΤΙΚΗ ΑΝΑΛΥΣΗ ΣΕ ΕΞΥΠΝΑ ΠΕΡΙΒΑΛΛΟΝΤΑ

MULTICHANNEL ACOUSTIC SCENE ANALYSIS IN SMART ENVIRONMENTS

Κυρίτσης Κωνσταντίνος

Διπλωματική εργασία

Πανεπιστήμιο Θεσσαλίας
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ

Πολυκαναλική ακουστική ανάλυση σε έξυπνα περιβάλλοντα

Κυρίτσης Κωνσταντίνος

Περίληψη

Η ανάλυση της ακουστικής σκηνής στοχεύει στην επεξεργασία και την ερμηνεία της ηχητικής πληροφορίας που διαδόθηκε στο περιβάλλον και ηχογραφήθηκε από πολύ-μικροφωνικές συστοιχίες. Η ανίχνευση και ο εντοπισμός ηχητικής πηγής, η αναγνώριση και η ταξινόμηση ακουστικών γεγονότων, η ενίσχυση της φωνής καθώς και η αναγνώριση ομιλητών αντικατοπτρίζουν ορισμένα από τα βασικά προβλήματα που βρίσκονται υπό έρευνα και που ουσιαστικά χαρακτηρίζουν την ακουστική σκηνή.

Η παρούσα διπλωματική εργασία ασχολείται με τον εντοπισμό της ηχητικής πηγής όπως αυτή έχει ηχογραφηθεί σε ένα "έξυπνο" περιβάλλον. Για τον σκοπό αυτό επιστρατεύτηκαν αλγόριθμοι εντοπισμού, καθώς και βελτιστοποιήσεις αυτών, βασισμένοι στις σχετικές καθυστερήσεις μεταξύ διαφορετικών ζευγαριών μικροφώνων. Για την διεξαγωγή των πειραμάτων χρησιμοποιήθηκαν δεδομένα από πραγματικές συστοιχίες 40 μικροφώνων εγκατεστημένων στα δωμάτια ενός "έξυπνου" διαμερίσματος. Τα αποτελέσματα που παρουσιάζονται υποδεικνύουν την απόδοση των μεθοδολογιών και κατά πόσο καλά μπορούν να προσεγγίσουν την πραγματική θέση της πηγής στον δισδιάστατο και στον τρισδιάστατο χώρο.

Multichannel acoustic scene analysis in smart environments

Kyritsis Konstantinos

Abstract

Acoustic scene analysis aims to elaborate and interpret the acoustic information that has been transmitted to the environment and has been recorded by multi microphone arrays. Detection and localization of sound sources, detection and classification of acoustic events, enhancement of the speech signal as well as speaker recognition constitute some of the key issues that are being actively researched and essentially characterize the acoustic scene.

The present thesis deals with the localization of the sound source as has been recorded in a “smart” environment. For this reason, not only detection algorithms were utilized but also their optimizations, based on the relative delays among different pairs of microphones. Experiments were conducted on data from multiple microphone arrays with 40 microphones total, installed in the rooms of a “smart” appartment. The results presented in this thesis indicate the algorithms effectiveness and how accurately can approximate the actual position of the source in two-dimensional and three dimensional spaces.

Ευχαριστίες

Αρχικά, θέλω να ευχαριστήσω τον επιβλέποντα της διπλωματικής εργασίας μου, αναπληρωτή καθηγητή κ. Γεράσιμο Ποταμιάνο, για την εμπιστοσύνη και την καθοδήγηση του στο πρακτικό αλλά και θεωρητικό τμήμα της δουλείας μου. Επίσης, οφείλω τις ευχαριστίες μου στο δεύτερο μέλος της επιτροπής, λέκτορα κ. Αντώνιο Αργυρίου που λόγω των μαθημάτων του, μου έδωσε το έναυσμα να ασχοληθώ με το προγραμματιστικό περιβάλλον MATLAB με το οποίο σχεδιάστηκε το μεγαλύτερο τεχνικό κομμάτι της εργασίας. Πάνω από όλα, είμαι ευγνώμων στους γονείς μου, Νίκο και Σοφία για την αγάπη και την υπομονή τους όλα αυτά τα χρόνια. Χωρίς αυτούς τίποτα δεν θα ήταν εφικτό.

Στην Ιουλία και την παρέα από Βόλο.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

Στόχος της διπλωματικής είναι η μελέτη της ακουστικής σκηνής και πιο συγκεκριμένα το πρόβλημα του εντοπισμού της ηχητικής πηγής σε οικιακά "έξυπνα" περιβάλλοντα. Τα προαναφερθέντα περιβάλλοντα αποτελούν ουσιαστικά κατοικίες εξοπλισμένες με συστοιχίες μικροφώνων, κάμερες και αισθητήρες κίνησης, με σκοπό να παρέχουν ασφάλεια, καθοδήγηση και άμεση υποστήριξη στις κατηγορίες των ανθρώπων που την έχουν ανάγκη (π.χ. άτομα τρίτης ηλικίας) καθώς και αλληλεπίδραση με οικιακές συσκευές. Η μελέτη που θα παρουσιαστεί σκοπεύει στην αξιοποίηση του πολυκαναλικού δικτύου των μικροφώνων όπως αυτά έχουν κατανεμηθεί στα δωμάτια των κατοικιών. Θεωρώντας κάθε δωμάτιο που βρίσκεται υπό μελέτη σαν έναν δισδιάστατο ή τρισδιάστατο χώρο, πραγματοποιείται η προσπάθεια εκτίμησης ενός συνόλου συντεταγμένων βασισμένων στις παρατηρήσεις των συστοιχιών μικροφώνων, που θα αντικατοπτρίζουν τις θέσεις του ομιλητή μέσα στην κατοικία. Τα πραγματικά δεδομένα που χρησιμοποιήθηκαν κατά την πειραματική μελέτη για την αξιολόγηση των συστημάτων εντοπισμού θέσης που αναπτύχθηκαν στα πλαίσια της διπλωματικής, περιλαμβάνουν τις προκλήσεις των παρασκηνιακών θορύβων καθώς και της αντήχησης (reverberation) καταστώντας το έργο του εντοπισμού μία μη-τετριμμένη διαδικασία.

1.1 Συναφείς εργασίες

Οι συστοιχίες μικροφώνων έχουν χρησιμοποιηθεί σε πολλές εφαρμογές, συμπεριλαμβανομένου τηλεδιάσκεψης [31], αναγνώριση φωνής, εύρεση της θέσης του κυρίαρχου ομιλητή σε ένα αμφιθέατρο [6]. Η εκτίμηση της κατεύθυνσης άφιξης (direction of arrival-DOA) των ηχητικών σημάτων χρησιμοποιώντας ένα σύνολο από χωρικά διαχωρισμένα μικρόφωνα έχει πολλές πρακτικές εφαρμογές στην καθημερινή ζωή. Για παράδειγμα η εκτίμηση των DOA μπορεί να χρησιμοποιηθεί για να στέφει τις κάμερες προς τον ενεργό ομιλητή σε κάποιο συνεδριακό χώρο [16].

Επιπρόσθετα, οι συστοιχίες μικροφώνων έχουν την δυνατότητα να αντικαταστήσουν τα παραδοσιακά μικρόφωνα (γνωστά ως ψείρες) ομιλητών καθώς και τα σταθερά μικρόφωνα γραφείου εξαιτίας των πολλαπλών πλεονεκτημάτων απέναντι σε συστήματα ενός μικροφώνου. Αρχικά, σε αντίθεση με ένα κατευθυνόμενο αισθητήρα, μια συστοιχία μπορεί να προσφέρει ένα υψηλής ποιότητας σήμα από την επιθυμητή περιοχή ενδιαφέροντος ενώ παράλληλα να ελαχιστοποιεί τον θόρυβο από δευτερεύοντες ομιλητές και περιβαλλοντολογικό θόρυβο. Στην περίπτωση πολλαπλών ή κινητών πηγών, μια συστοιχία δεν επιβαρύνει τον ομιλητή στην διαδικασία του να κρατάει στα χέρια του ή να φοράει στο κεφάλι του (Εικόνα 1.1) κάποιο μικρόφωνο όπως επίσης και δεν απαιτεί την περιστροφή

(α΄) συστοιχία μικροφώνων σε τοίχο

(β΄) μικρόφωνο κεφαλής

Σχήμα 1.1: Διάφορα είδη αισθητήρων

του αισθητήρα έτσι ώστε να αλλάξει η κατεύθυνση λήψης.

Η θεμελιώδης απαίτηση για κάθε εφαρμογή που χρησιμοποιεί συστοιχίες είναι η ικανότητα του να προσδιορίζει τις σχετικές χρονικές καθυστερήσεις (time delays) μεταξύ αφίξεων των σημάτων στις διαφορετικές τοποθεσίες των μικροφώνων. Η ακρίβεια και η ευρωστία αυτών των εκτιμήσεων παίζει σημαντικό ρόλο στο σύστημα του εντοπισμού της θέσης. Η ποιότητα του υλικού (hardware) που χρησιμοποιήθηκε, ο αριθμός των μικροφώνων καθώς και η συχνότητα δειγματοληψίας επηρεάζουν σε σημαντικό βαθμό τις προαναφερθέντες εκτιμήσεις. Επίσης, στην περίπτωση της παρακολούθησης κινητών πηγών (tracking) πρέπει οι εκτιμήσεις των χρονικών καθυστερήσεων να ανανεώνονται περιοδικά.

Σύμφωνα με την βιβλιογραφία, η μέθοδος της γενικευμένης ετεροσυσχέτισης (GCC ή generalized cross correlation) που προτάθηκε από τους Knapp και Carter το 1976 [17], είναι η πιο διαδεδομένη τεχνική για την εκτίμηση των χρονικών καθυστερήσεων. Οι παραπάνω εκτιμήσεις εξασφαλίζονται ως την χρονική καθυστέρηση (lag) που μεγιστοποιεί την συσχέτιση μεταξύ των σημάτων που έφτασαν στους αισθητήρες. Μετά τους Knapp και Carter έχουν προταθεί πολλές ιδέες για την καταπολέμηση της αντήχησης (reverberation), όπως:[20],[4],[8],[28],[21],[9],[30]. Παρόλα αυτά, η αντήχηση παραμένει πρόβλημα και στην περίπτωση δωματίων με υψηλές συνθήκες αντήχησης, όλες οι γνωστές μεθοδολογίες αποτυχαίνουν. Υπάρχουν δύο προσεγγίσεις του προβλήματος της αντήχησης. Η πρώτη είναι η "τυφλή" εκτίμηση της κρουστικής απόκρισης από την πηγή ως προς τα δύο μικρόφωνα. Όσο καλύτερη είναι η εκτίμηση της απόκρισης τόσο καλύτερη είναι οι εκτιμήσεις των καθυστερήσεων μεταξύ των μικροφώνων. Η δεύτερη προσέγγιση είναι να χρησιμοποιηθούν παραπάνω από δύο μικρόφωνα και να γίνει εκμετάλλευση της επιπλέον πολυκαναλικής πληροφορίας [10].

1.2 Υποθέσεις και περιορισμοί

Καθ όλη την έκταση της διπλωματικής έχουν θεωρηθεί οι ακόλουθες προϋποθέσεις υπό τις οποίες εκτιμάται η θέση της πηγής του ήχου:

- Δυνατότητα εύρεσης μίας απείρως μικρής μοναδικής πηγής στον χώρο (single source estimation).

- Η πηγή εκπέμπει την ηχητική πληροφορία με σφαιρικό τρόπο. Οι πολύπλοκοι μέθοδοι εκπομπής του ανθρώπινου κεφαλιού δεν τέθηκαν υπ όψιν.
- Μη ύπαρξη επιπρόσθετων πηγών θορύβου.
- Πλήρως συγχρονισμένα μεταξύ τους μικρόφωνα.
- Γνώση της ακριβούς γεωμετρίας του χώρου που βρίσκεται υπό μελέτη καθώς και της τοποθεσίας των αισθητήρων και της εκάστοτε μεταξύ τους αποστάσεων.
- Το μέσο διάδοσης είναι ομογενές. Δηλαδή, οι αλλαγές στην ταχύτητα διάδοσης του ήχου εξαιτίας των αλλαγών της ατμοσφαιρικής πίεσης και θερμοκρασίας έχουν παραληφθεί. Η ταχύτητα μετάδοσης του ήχου στον αέρα έχει θεωρηθεί σταθερή ίση με $c = 330$ m/s.
- Το μέσο διάδοσης δεν έχει απώλειες και άρα δεν απορροφάει ενέργεια από τα μεταδιδόμενα σήματα.

Όπως αναφέρθηκε στις παραπάνω προϋποθέσεις, η ηχητική πηγή, είτε αυτή είναι ένας άνθρωπος είτε ένα μηχανικό μέσο, δεν έχει την ιδεατή σφαιρική μετάδοση. Σε ρεαλιστικά περιβάλλοντα η μετάδοση είναι κατευθυνόμενη και υπάρχει χωρική εξασθένιση. Τα μικρόφωνα που είναι τοποθετημένα μπροστά στον ομιλητή θα δεχτούν σήματα μεγαλύτερης ισχύος από τα πλαϊνά ή τα μικρόφωνα που βρίσκονται από πίσω του. Για λόγους απλότητας μοντελοποιήσαμε τις ηχητικές πηγές σαν σημεία. Αξίζει να σημειωθεί πως η εύρεση της τέλειας λύσης δεν είναι εφικτή εξαιτίας των παρακάτω παραγόντων:

- Ανακρίβειες στην γεωμετρία των μικροφώνων και της πηγής.
- Παρουσία πηγών θορύβου στον χώρο.
- Αριθμητικά λάθη στις στρογγυλοποιήσεις.
- Μη ακριβείς καθυστερήσεις μετάδοσης.
- Η τοποθέτηση των συστοιχιών μπορεί να μην είναι η βέλτιστη.
- Αντήχηση του σήματος πάνω σε διάφορες επιφάνειες.

1.3 Οργάνωση της διπλωματικής

Η παρούσα διπλωματική οργανώνεται ως εξής. Στο κεφάλαιο 2, γίνεται η μοντελοποίηση του προβλήματος και περιγράφεται το μαθηματικό υπόβαθρο. Επιπρόσθετα, παρουσιάζονται οι αλγόριθμοι γενικευμένης ετεροσυσχέτισης με τον μετασχηματισμό φάσης (Generalized Cross Correlation PHase Transform GCC-PHAT) και κατευθυντήριας δύναμης απόκρισης με τον μετασχηματισμό φάσης (Steered Response Power PHAse Transform SRP-PHAT) που χρησιμοποιήθηκαν για την εκτίμηση των κατευθύνσεων άφιξης (DOA). Στο κεφάλαιο 3 ακολουθεί η περιγραφή της βάσης δεδομένων "DIRHA" [29] και η γεωμετρία των χώρων στους οποίους έχουν πραγματοποιηθεί οι ηχογραφήσεις. Ακολούθως, στο κεφάλαιο 4 γίνεται συζήτηση πάνω στα πειράματα που πραγματοποιήθηκαν στην βάση δεδομένων, παρουσιάζονται τα αποτελέσματα που προέκυψαν από την εφαρμογή των εκάστοτε μεθοδολογιών καθώς και η σύγκριση των αποτελεσμάτων. Το κεφάλαιο 5 κλείνει την διπλωματική με συμπεράσματα καθώς και μελλοντικές ερευνητικές κατευθύνσεις της εργασίας.

Κεφάλαιο 2

Μοντελοποίηση του προβλήματος

2.1 Κοντινό και μακρινό πεδίο

Στην περίπτωση που η απόσταση από την ακουστική πηγή στην συστοιχία μικροφώνων είναι πολύ μεγαλύτερη από το μέγεθος της συστοιχίας τότε η κατάσταση αναφέρεται ως μακρινού πεδίου (far field). Με αυτόν τον τρόπο, τα ηχητικά σήματα εμφανίζονται ως επίπεδα (planar) όταν φτάσουν στην συστοιχία. Στην αντίθετη περίπτωση, όπου η απόσταση είναι μικρότερη ή περίπου ίση με το μέγεθος της συστοιχίας, η κατάσταση αναφέρεται ως κοντινού πεδίου (near field).

2.2 Μαθηματικό μοντέλο σήματος

Έστω ότι έχουμε $L+1$ σήματα μικροφώνων $x_l(n), l = 0, 1, 2, ..., L$. Χωρίς να βλάψουμε την γενίκευση, μπορούμε να υποθέσουμε πως το σήμα είναι στην ίδια φάση με το μικρόφωνο 0. Επομένως μπορούμε να θεωρήσουμε το παρακάτω μοντέλο διάδοσης:

$$\begin{bmatrix} x_0(n) \\ x_1(n) \\ x_2(n) \\ \vdots \\ x_L(n) \end{bmatrix} = \begin{bmatrix} a_0 & 0 & 0 & \cdots & 0 \\ 0 & a_1 & 0 & \cdots & 0 \\ 0 & 0 & a_2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & a_L \end{bmatrix} \times \begin{bmatrix} s(n-t) \\ s(n-t-\tau) \\ s(n-t-f_2(\tau)) \\ \vdots \\ s(n-t-f_L(\tau)) \end{bmatrix} + \begin{bmatrix} w_0(n) \\ w_1(n) \\ w_2(n) \\ \vdots \\ w_L(n) \end{bmatrix} \tag{2.1}$$

όπου $a_l, l = 0, 1, 2, ..., L$ είναι οι παράγοντες εξασθένησης (attenuation), εξαιτίας των φαινομένων διάδοσης στο μέσο, t είναι ο χρόνος διάδοσης από την άγνωστη πηγή $s(n)$ στο μικρόφωνο 0, $w_l(n)$ είναι το σήμα του προσθετικού θορύβου στο $l-$οστό μικρόφωνο, τ η σχετική καθυστέρηση μεταξύ των μικροφώνων 0 και 1, και $f_l(\tau)$ η σχετική καθυστέρηση μεταξύ των μικροφώνων 0 και l. Η συνάρτηση f_l εξαρτάται από το τ αλλά επίσης και από την γεωμετρία της συστοιχίας. Για παράδειγμα, στην περίπτωση μακρινού πεδίου (far field) για μία γραμμική συστοιχία όπου τα μικρόφωνα έχουν τις ίδιες αποστάσεις μεταξύ τους έχουμε:

$$f_l(\tau) = l\tau \tag{2.2}$$

και για μία μη-γραμμική έχουμε:

$$f_l(\tau) = \frac{\sum_{i=0}^{l-1} d_i}{d_0}\tau \tag{2.3}$$

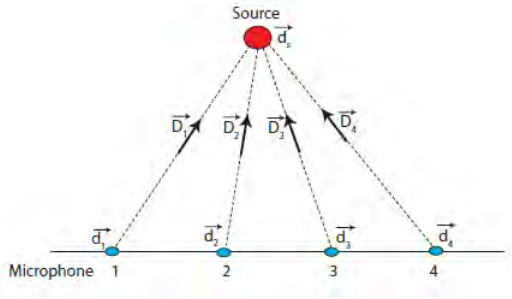
όπου d_i είναι η απόσταση μεταξύ των μικροφώνων i και $i+1, i = 0, 1, 2, ..., L-1$. Στην περίπτωση κοντινού πεδίου (near field), η f_l εξαρτάται και από την θέση της πηγής. Γενικότερα το τ δεν είναι γνωστό αλλά η γεωμετρία της συστοιχίας είναι, επομένως η ακριβής σχέση της σχετικής καθυστέρησης μεταξύ των μικροφώνων 0 και l είναι καλώς ορισμένη. Επιπρόσθετα μπορούμε να υποθέσουμε ότι τα σήματα $s(n)$ και $w_l(n), l = 0, 1, 2, ..., L$ είναι ασυσχέτιστα, έχουν μηδενική μέση τιμή (zero mean) και ακολουθούν στάσιμες Γκαουσιανές κατανομές.

2.3 Κατεύθυνση άφιξης

Στο σενάριο κοντινού πεδίου για μία συστοιχία με M-μικρόφωνα, υπάρχουν M διαφορετικές κατευθύνσεις άφιξης. Κάθε μια από αυτές είναι το άμεσο μονοπάτι από την ηχητική πηγή στο μικρόφωνο. Μαθηματικώς, μπορούν να ορισθούν όπως παρακάτω:

$$\vec{D_m} = \frac{\vec{d_m} - \vec{d_s}}{|\vec{d_m} - \vec{d_s}|} \tag{2.4}$$

για $m = 1, 2, ..., M$



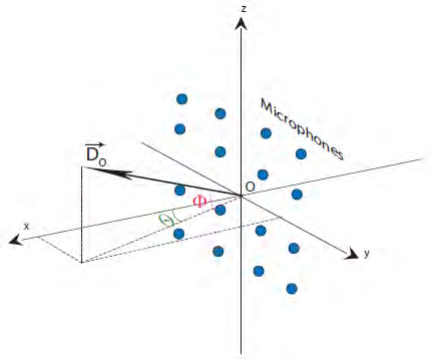
Σχήμα 2.1: DOA's τεσσάρων μικροφώνων σε συστοιχία στην περίπτωση κοντινού πεδίου από [13].

Στην περίπτωση μακρινού πεδίου, όλα τα μικρόφωνα τις συστοιχίας έχουν την ίδια κατεύθυνση άφιξης (DOA), η οποία επιλέγεται ως το μονοπάτι από την πηγή στην συστοιχία. Με O υποδηλώνουμε την θέση της συστοιχίας στο σύστημα συντεταγμένων του χώρου και μπορούμε να το εκφράσουμε ως εξής,

$$\vec{D_O} = \frac{\vec{d_O} - \vec{d_s}}{|\vec{d_O} - \vec{d_s}|} \tag{2.5}$$

Ο προσανατολισμός της κατεύθυνσης άφιξης (DOA) μπορεί να οριστεί από μια αζιμούθια γωνία (azimuth angle) θ και από μία γωνία ανύψωσης (elevation angle) ϕ:

$$\vec{D_O} = \begin{bmatrix} \cos\phi \ \sin\theta \\ \cos\phi \ \cos\theta \\ \sin\theta \end{bmatrix} \tag{2.6}$$

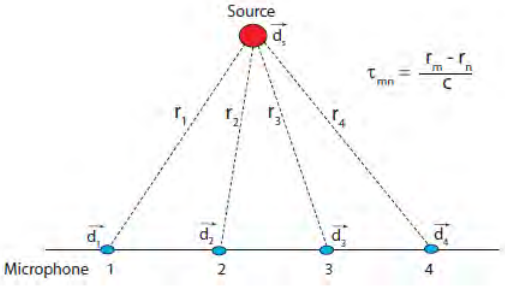


Σχήμα 2.2: DOA στην περίπτωση μακρινού πεδίου με την αζιμούθια γωνία θ και την γωνία ανύψωσης ϕ από [13]

Στο πρόβλημα του εντοπισμού ομιλητή, η απόσταση από την ηχητική πηγή στην συστοιχία (range) δεν μπορεί να καθοριστεί επομένως η κατεύθυνση αφίξεως είναι η μοναδική χωρική πληροφορία για την πηγή.

2.4 Η μεθοδολογία Γενικευμένης Ετεροσυσχέτισης (GCC)

Το GCC (Generalized Cross-Correlation) είναι η πιο δημοφιλής μεθοδολογία για την εκτίμηση των καθυστερήσεων άφιξης (TDOA) μεταξύ ζευγαριών μικροφώνων. Με αυτόν τον τρόπο, χρησιμοποιώντας περισσότερα ζεύγη μπορούν να προκύψουν πολλαπλά TDOA και έτσι να εκτιμηθεί η θέση της ακουστικής πηγής. Ας πάρουμε για παράδειγμα την παρακάτω συστοιχία.



Σχήμα 2.3: TDOA μεταξύ δύο μικροφώνων από [13]

Αν η απόσταση από το μικρόφωνο m στην πηγή είναι $r_m (m = 1, 2, 3, 4)$, τότε η χρονική καθυστέρηση (traveling time) του σήματος από την πηγή στο μικρόφωνο είναι,

$$\tau_m = \frac{r_m}{c} \tag{2.7}$$

Τότε η καθυστέρηση άφιξης μεταξύ 2 μικροφώνων m και n μπορεί να ορισθεί ως,

$$\tau_{mn} = \tau_m - \tau_n = \frac{r_m - r_n}{c} \tag{2.8}$$

Από την παραπάνω σχέση των TDOA με τις αποστάσεις μεταξύ της πηγής και των μικροφώνων r_m, μπορεί να γίνει η εκτίμηση της θέσης της πηγής χρησιμοποιώντας διάφο-

ρες τεχνικές όπως: γραμμική τομή (linear intersection), σφαιρική παρεμβολή (spherical interpolation), κτλ [23],[25],[19].

2.5 Συνέπεια του GCC

Από την εξίσωση του μοντέλου διάδοσης 2.1 μπορούμε να πάρουμε τις εξισώσεις ενός ζεύγους μικροφώνων k και l:

$$x_k(n) = a_k s(n - t - f_k(\tau)) + w_k(n) \tag{2.9}$$

και

$$x_l(n) = a_l s(n - t - f_l(\tau)) + w_l(n) \tag{2.10}$$

Η συσχέτιση αυτών των δύο σημάτων μικροφώνων θα μεγιστοποιηθεί σύμφωνα με την χρονική καθυστέρηση (time lag) που απαιτείται έτσι ώστε τα δύο ολισθημένα σήματα να συμπέσουν (aligned). Η συσχέτιση των $x_k(n)$ και $x_l(n)$ ορίζεται ως,

$$c_{kl}(\tau) = \int_{-\infty}^{+\infty} x_k(t) x_l(t + \tau) dt \tag{2.11}$$

Παίρνοντας τον μετασχηματισμό Fourier της παραπάνω συσχέτισης έχει ως αποτέλεσμα το cross power spectrum,

$$C_{kl}(\omega) = \int_{-\infty}^{+\infty} c_{kl}(\tau) e^{j\omega\tau} d\tau \tag{2.12}$$

Εφαρμόζοντας τις ιδιότητες συνέλιξης του μετασχηματισμού Fourier στην (2.11) όταν την αντικαθιστούμε στην (2.12) παίρνουμε,

$$C_{kl}(\omega) = X_k(\omega) X_l^*(\omega) \tag{2.13}$$

όπου $X_i(\omega)$ είναι ο μετασχηματισμός Fourier του σήματος $x_i(t)$ και το σύμβολο $'*'$ προσδιορίζει το μιγαδικό συζυγές. Παίρνοντας τον αντίστροφο μετασχηματισμό Fourier στην (2.13) μας δίνει την συνάρτηση συσχέτισης των σημάτων των μικροφώνων δεδομένου των μετασχηματισμών τους:

$$c_{kl}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X_k(\omega) X_l^*(\omega) e^{j\omega\tau} d\omega \tag{2.14}$$

Ο GCC των σημάτων $x_k(t)$ και $x_l(t)$ είναι η συσχέτιση των δύο φιλτραρισμένων εκδοχών τους. Δηλώνοντας τους μετασχηματισμούς Fourier ως $W_k(\omega)$ και $W_l(\omega)$ έχουμε το GCC τους $R_{kl}(\tau)$ που ορίζεται ως,

$$R_{kl}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (W_k(\omega) X_k(\omega))(W_l(\omega) X_l(\omega))^* e^{j\omega\tau} d\omega \tag{2.15}$$

Επίσης ορίζουμε την συνάρτηση βαρών $\Psi_{kl}(\omega)$ ως

$$\Psi_{kl}(\omega) = W_k(\omega) W_l^*(\omega) \tag{2.16}$$

Αντικαθιστώντας την (2.15) στην (2.16), έχουμε:

$$R_{kl}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Psi_{kl}(\omega) X_k(\omega) X_l^*(\omega) e^{j\omega\tau} d\omega \tag{2.17}$$

Το TDOA μεταξύ των δύο μικροφώνων k και l είναι η καθυστέρηση τ που μεγιστοποιεί το GCC $R_{kl}(\tau)$ στο πραγματικό πεδίο τιμών που περιορίζεται απο την απόσταση μεταξύ των μικροφώνων:

$$\hat{\tau}_{kl} = \underset{\tau}{\text{argmax}}\, R_{kl}(\tau) \tag{2.18}$$

Στην πραγματικότητα το $R_{kl}(\tau)$ έχει πολλά τοπικά μέγιστα, όπου το καθιστά δύσκολο να βρεθεί το γενικό μέγιστο (global maximum). Επίσης, η επιλογή της συνάρτησης βάρους $\Psi_{kl}(\omega)$ μπορεί να επηρεάσει την απόδοση της μεθοδολογίας GCC.

2.6 Ο μετασχηματισμός φάσης (PHat)

Στην βιβλιογραφία έχει δειχτεί πως ο μετασχηματισμός φάσης (PHAT) σαν συνάρτηση βάρους αποδίδει αξιόπιστα υπό ρεαλιστικές συνθήκες [24],[11], ακόμα και αν είναι υπό-βέλτιστος [7] σε σχέση με τον εκτιμητή μέγιστης πιθανοφάνειας (ML-maximum likelihood) που μελετήθηκε από [5],[17], σε συνθήκες χωρίς αντήχηση. Μπορούμε να ορίσουμε τον PHAT παρακάτω,

$$\Psi_{kl}(\omega) = \frac{1}{|X_k(\omega) X_l^*(\omega)|} \tag{2.19}$$

2.7 GCC-PHAT

Εφαρμόζοντας την (2.19) στην έκφραση που ορίσαμε για το GCC (2.17) έχουμε το GCC-PHAT για τα δύο μικρόφωνα k και l που ορίζεται ως εξής,

$$R_{kl}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{1}{|X_k(\omega) X_l^*(\omega)|} X_k(\omega) X_l^*(\omega) e^{j\omega\tau} d\omega \tag{2.20}$$

Σε μία συστοιχία M-μικροφώνων υπάρχουν $M(M-1)/2$ ζεύγη μικροφώνων. Χρησιμοποιώντας το GCC-PHAT σε οποιοδήποτε υποσύνολο Q των ζευγαριών για τον υπολογισμό των TDOA παίρνουμε Q TDOA εκτιμήσεις. Για κάθε σημείο $\vec{x}$ στον τρισδιάστατο ή δισδιάστατο χώρο του δωματίου όπου βρίσκεται η ηχητική πηγή, μπορούμε να βρούμε το πραγματικό TDOA από τα Q ζευγάρια μικροφώνων. Από τα TDOA's που εκτιμήθηκαν $\hat{\tau_Q}(\vec{x})$ και τα πραγματικά TDOA's $\tau_Q(\vec{x})$ μπορούμε να ορίσουμε το μέσο τετραγωνικό σφάλμα (RMS) ως εξής,

$$E_{RMS}(\vec{x}) = \sqrt{(\hat{\tau_Q}(\vec{x}) - \tau_Q(\vec{x}))^2} \tag{2.21}$$

Και η τελική εκτίμηση της θέσης της πηγής $\vec{x_s}$ είναι,

$$\vec{x_s} = \underset{\vec{x}}{\text{argmin}}\, E_{RMS}(\vec{x}) \tag{2.22}$$

2.8 Σχηματισμός δέσμης (Beamforming)

Όπως περιγράφηκε στην εξίσωση 2.1, το σήμα $x_m(t)$ που καταλήγει στο μικρόφωνο m είναι,

$$x_m(n) = a_m s(n - t - f_m(\tau)) + w_m(n) \tag{2.23}$$

Σε μία συστοιχία M-μικροφώνων η μέθοδος δημιουργίας δέσμης με καθυστέρηση-και-άθροιση (delay-and-sum) μπορεί να σχηματιστεί μέσω της καθυστέρησης στους αισθητήρες $x_m(n)$ με τις κατάλληλες κατευθυντήριες καθυστερήσεις (steering delays), δ_m με $m = 1, 2, ..., M$ με σκοπό να ευθυγραμμιστούν (allign) στον χρόνο, και στην συνέχεια αθροίζοντας όλα τα χρονικά ευθυγραμμισμένα σήματα μαζί. Μαθηματικώς μπορεί να οριστεί όπως παρακάτω,

$$y(t, \delta_1, \delta_2, ..., \delta_M) = \sum_{m=1}^{m=M} x_m(t - \delta_m) \tag{2.24}$$

Για να ευθυγραμμιστούν τα μικρόφωνα στον χρόνο, οι κατευθυντήριες καθυστερήσεις δ_m μπορούν να ορισθούν ως,

$$\delta_m = \tau_m - \tau_0 \tag{2.25}$$

όπου τ_m είναι η χρονική καθυστέρηση από την πηγή στο μικρόφωνο m, και το τ_0 έχει ορισθεί έτσι ώστε να είναι η ελάχιστη των χρονικών καθυστερήσεων $\tau_i, i = [1, 2, ..., M]$ με σκοπό το δ_m να είναι μη-αρνητικό και κατά συνέπεια το σύστημα να είναι αιτιατό (casual).

Επομένως, μπορούμε να εκφράσουμε την έξοδο του δημιουργού δέσμης με καθυστέρηση-και-άθροιση σε σχέση με το σήμα της πηγής, την κρουστική απόκριση του καναλιού και τον προσθετικό θόρυβο ως εξής,

$$y(t, \delta_1, \delta_2, ..., \delta_M) = s(t) \sum_{m=1}^{m=M} a_m + \sum_{m=1}^{m=M} w_m(t - \tau_m + \tau_0) \tag{2.26}$$

Όταν ένα προσαρμοστικό φίλτρο εφαρμοστεί στον δημιουργό δέσμης καθυστέρησης-και-άθροισης τότε έχει ως αποτέλεσμα τον δημιουργό δέσμης με φιλτράρισμα-και-άθροισμα (filter-and-sum). Στο πεδίο της συχνότητας, η έξοδος με την δεύτερη μέθοδο που αναφέρθηκε παραπάνω είναι,

$$Y(\omega, \delta_1, \delta_2, ..., \delta_M) = \sum_{m=1}^{m=M} G_m(\omega) X_m(\omega) e^{-j\omega\delta_m} \tag{2.27}$$

Όπου $X_m(\omega)$ είναι ο μετασχηματισμός Fourier του σήματος του μικροφώνου $x_m(t)$ και $G_m(\omega)$ είναι ο μετασχηματισμός Fourier του φίλτρου.

2.9 Η μέθοδος κατευθυντήριας δύναμης απόκρισης

Γενικά, η κατευθυντήρια δύναμη απόκρισης (SRP) είναι η έξοδος του δημιουργού δέσμης με φιλτράρισμα-και-άθροισμα όταν τον κατευθύνουμε ως προς όλα τα σημεία $\vec{x}$ σε μία προκαθορισμένη περιοχή. Για κάθε σημείο $\vec{x}$ υπάρχει η συνάρτηση των κατευθυντήριων καθυστερήσεων που μπορεί να ορισθεί όπως παρακάτω στο πεδίο της συχνότητας,

$$P(\delta_1, ..., \delta_M) = \int_{-\infty}^{+\infty} Y(\omega, \delta_1, ..., \delta_M) Y^*(\omega, \delta_1, ..., \delta_M) d\omega \tag{2.28}$$

Αντικαθιστώντας την (2.27) στην (2.28) έχουμε,

$$P(\delta_1, ..., \delta_M) = \int_{-\infty}^{+\infty} (\sum_{k=1}^{k=M} G_k(\omega) X_k(\omega) e^{-j\omega\delta_k})(\sum_{l=1}^{l=M} G_l^*(\omega) X_l^*(\omega) e^{j\omega\delta_l}) d\omega \tag{2.29}$$

Μεταθέτοντας τους όρους από την παραπάνω εξίσωση, παίρνουμε,

$$P(\delta_1, ..., \delta_M) = \int_{-\infty}^{+\infty} \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} (G_k(\omega) G_l^*(\omega))(X_k(\omega) X_l^*(\omega)) e^{j\omega(\delta_l - \delta_k)} d\omega \tag{2.30}$$

Από την (2.25) είναι εύκολο να δούμε πως,

$$\delta_l - \delta_k = \tau_l - \tau_k \tag{2.31}$$

Αντικαθιστώντας την (2.31) στην (2.30) έχουμε,

$$P(\delta_1, ..., \delta_M) = \int_{-\infty}^{+\infty} \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} (G_k(\omega) G_l^*(\omega))(X_k(\omega) X_l^*(\omega)) e^{j\omega(\tau_l - \tau_k)} d\omega \tag{2.32}$$

Αξίζει να σημειωθεί πως το παραπάνω ολοκλήρωμα συγκλίνει, διότι στον πραγματικό κόσμο τα σήματα των μικροφώνων καθώς και τα φίλτρα έχουν πεπερασμένη ενέργεια. Επομένως, τα αθροίσματα μπορούν να αντιμεταθετιθούν με το ολοκλήρωμα εξωτερικά όπως παρακάτω,

$$P(\delta_1, ..., \delta_M) = \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} \int_{-\infty}^{+\infty} (G_k(\omega) G_l^*(\omega))(X_k(\omega) X_l^*(\omega)) e^{j\omega(\tau_l - \tau_k)} d\omega \tag{2.33}$$

Ορίζοντας την συνδυασμένη συνάρτηση βαρών,

$$\Psi_{kl}(\omega) = G_k(\omega) G_l^*(\omega) \tag{2.34}$$

Αντικαθιστώντας την (2.34) στην (2.33) και δεδομένου του ότι $\tau_l - \tau_k = \tau_{lk}$ παίρνουμε την έκφραση για την κατευθυντήρια δύναμη απόκρισης (SRP):

$$P(\delta_1, ..., \delta_M) = \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} \int_{-\infty}^{+\infty} \Psi_{kl}(\omega) X_{kl}(\omega) X_l^*(\omega) e^{j\omega\tau_{lk}} d\omega \tag{2.35}$$

Ας θυμηθούμε την έκφραση για την γενικευμένη ετεροσυσχέτιση από την (2.17):

$$R_{kl}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Psi_{kl}(\omega) X_k(\omega) X_l^*(\omega) e^{j\omega\tau} d\omega$$

Είναι εύκολο να δει κανείς πως οι εκφράσεις για την κατευθυντήρια δύναμη απόκρισης (SRP) και την γενικευμένη ετεροσυσχέτιση (GCC) είναι παρόμοιες, με την εξαίρεση πως το SRP αθροίζεται ως προς όλα τα ζεύγη των μικροφώνων και υπάρχει η σταθερά 2π. Επομένως, αυτό μας δίνει ένα τρόπο να υπολογίσουμε την κατευθυντήρια δύναμη απόκρισης μιας συστοιχίας μικροφώνων αθροίζοντας την γενικευμένη ετεροσυσχέτιση όλων των ζευγαριών μικροφώνων που βρίσκονται στην συστοιχία (όπου η σταθερά 2π αγνοείται γιατί είναι απλά βαθμωτός όρος).

2.10 Ο μετασχηματισμός φάσης στην κατευθυντήρια δύναμη απόκρισης

Παρόμοια με την ιδέα της μεθόδου γενικευμένης ετεροσυσχέτισης με τον μετασχηματισμό φάσης (GCC-PHAT), όταν εφαρμόζουμε στην κατευθυντήρια δύναμη απόκρισης την συνάρτηση βαρών του μετασχηματισμού φάσης (PHase transform) παίρνουμε την κατευθυντήρια δύναμη απόκρισης με μετασχηματισμό φάσης (SRP-PHAT). Το SRP-PHAT για κάθε σημείο $\vec{x}$ στο χώρο ορίζεται ως εξής,

$$P(\delta_1, ..., \delta_M) = \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} \int_{-\infty}^{+\infty} \frac{1}{|X_k(\omega) X_l^*(\omega)|} X_k(\omega) X_l^*(\omega) e^{j\omega\tau_{lk}} d\omega \tag{2.36}$$

2.11 Η μέθοδος SRP-PHAT για τον εντοπισμό πηγής

Δεδομένου του ότι η γενικευμένη ετεροσυσχέτιση μεταξύ των μικροφώνων k και l είναι η ίδια με την γενικευμένη ετεροσυσχέτιση μεταξύ των l και k, τα στοιχεία που αθροίζονται για να παραχθεί η παραπάνω κατευθυντήρια δύναμη απόκρισης με μετασχηματισμό φάσης σχηματίζουν ένα συμμετρικό πίνακα με σταθερή ενέργεια στους όρους της κυρίας διαγωνίου. Ως εκ τούτου, το μόνο κομμάτι που αλλάζει μαζί με το $\vec{x}$ είναι είτε το άνω τριγωνικό κομμάτι ή το κάτω τριγωνικό κομμάτι του πίνακα. Με άλλα λόγια, για ένα συγκεκριμένο σημείο $\vec{x}$ στον χώρο το κομμάτι της κατευθυντήριας δύναμης απόκρισης που αλλάζει στην εξίσωση 2.36 μπορεί να υπολογιστεί υπολογίζοντας την γενικευμένη ετεροσυσχέτιση όχι όλων των ζευγαριών μικροφώνων αλλά μόνο ενός υποσυνόλου Q των ζευγαριών, όπου $Q = [k, l], \forall k \in [1, ..., M-1], M \geq l > k$,

$$P^{'}(\delta_1, ..., \delta_M) = \sum_{k=1}^{k=M} \sum_{l=k+1}^{l=M} \int_{-\infty}^{+\infty} \frac{1}{|X_k(\omega)X_l^*(\omega)|} X_k(\omega)X_l^*(\omega)e^{j\omega\tau_{lk}}d\omega \tag{2.37}$$

Επομένως μπορούμε να κατευθύνουμε τον δημιουργό δέσμης (beamformer) σε όλα τα πιθανά σημεία στο χώρο για να βρούμε την θέση της πηγής. Τα σημεία που επιστρέφουν την μεγαλύτερη τιμή της κατευθυντήριας δύναμης με τον μετασχηματισμό φάσης (SRP-PHAT) είναι πιθανές θέσης της ηχητικής πηγής. Για μία μοναδική πηγή η εκτίμηση της θέσης της $\vec{x_s}$ είναι,

$$\vec{x_s} = \underset{\vec{x}}{\text{argmax}}\, P^{'}(\vec{x}) \tag{2.38}$$

Όπου $P^{'}(\vec{x})$ είναι η SRP-PHAT τιμή στο σημείο $\vec{x}$, που ορίζεται από την (2.37). Αξίζει να σημειωθεί πως ο υπολογισμός οποιοδήποτε σημείου του $P^{'}(\vec{x})$ ορίζεται ως η λειτουργική αξιολόγηση. Η υπόθεση μας είναι ότι η τιμή του SRP-PHAT θα κορυφωθεί στην πραγματική θέση της πηγής ακόμα και σε θορυβώδεις χώρους ή χώρους με υψηλή αντήχηση. Παρόλα αυτά, το πρόβλημα με την μέθοδο SRP-PHAT είναι το ακριβό υπολογιστικό κόστος λόγω της ύπαρξης πολλών τοπικών μεγίστων (local maxima) και της εξουθενωτικής αναζήτησης (exhaustive search) του πλέγματος (grid).

Κεφάλαιο 3

Η βάση δεδομένων DIRHA

Το πρόγραμμα DIRHA (Distant-speech Interaction for Robust Home Applications) ξεκίνησε τον Ιανουάριο του 2012 και απευθύνεται στο έργο της αναγνώρισης και κατανόησης απομακρυσμένης ομιλίας σε οικιακό περιβάλλον. Στα πλαίσια του προγράμματος, δημιουργήθηκε ένα πλήρως αυτοματοποιημένο διαμέρισμα στην πόλη Trento της Ιταλίας όπου και στεγάζονται, αναπτύσσονται και υποβάλλονται σε δοκιμές όλα τα πρωτότυπα συστήματα του προγράμματος (Εικόνα 3.1). Στο προαναφερθέν διαμέρισμα ο φωτισμός,η θέρμανση, οι πόρτες και τα παράθυρα ελέγχονται από μία κεντρική μονάδα. Η αλληλεπίδραση μεταξύ του χρήστη και του συστήματος πραγματοποιείται μέσω φωνητικών εντολών αξιοποιώντας ένα δίκτυο μικροφώνων τοποθετημένων στα διάφορα δωμάτια του διαμερίσματος. Παράλληλα με τις παραπάνω λειτουργίες, διεξάγεται επίσης έρευνα στα πεδία της πολυκαναλικής ακουστικής επεξεργασίας, της αναγνώρισης/επιβεβαίωσης ομιλητή καθώς και διαχείρισης φωνητικού διαλόγου.

Στο πεδίο της αναγνώρισης και κατανόησης ομιλίας και πιο συγκεκριμένα στην περίπτωση της απομακρυσμένης αλληλεπίδρασης από τα μικρόφωνα, η κατάλληλη επιλογή των δεδομένων που θα χρησιμοποιηθούν στην εκπαίδευση και την δοκιμή των αλγορίθμων επεξεργασίας, ενίσχυσης και αναγνώρισης της φωνής είναι αναγκαία. Το μέγεθος καθώς και η ικανότητα της βάσης δεδομένων να ταιριάζει με ρεαλιστικές συνθήκες χρήσης επηρεάζουν την ακρίβεια των στατιστικών μοντέλων που χρησιμοποιούνται στο βήμα της αναγνώρισης και τελικά στην απόδοση ολόκληρου του συστήματος. Από την άλλη πλευρά, η συλλογή των κατάλληλων δεδομένων για την κάλυψη όλων των ερευνητικών αναγκών που μπορεί να προκύψουν είναι ένα ακριβό και χρονικά απαιτητικό έργο. Επομένως, δεν είναι εφικτή η δημιουργία μίας βάσης η οποία θα είναι εκτενής και αρκετά αντιπροσωπευτική για να καλύψει οποιεσδήποτε θορυβώδης συνθήκες καθώς και συνθήκες αντήχησης σε οικιακά περιβάλλοντα.

Για τις ερευνητικές ανάγκες του προγράμματος DIRHA δημιουργήθηκαν δύο διαφορετικά σετ δεδομένων σε τέσσερις γλώσσες (Ελληνικά, Ιταλικά, Πορτογαλικά και Αυστριακά Γερμανικά). Τα πραγματικά δεδομένα (real data) όπου περιέχουν μια συλλογή υλικού από πραγματικές απομακρυσμένες ομιλίες σε οικιακούς χώρους καθώς και τα προσομοιωμένα (simulated data) δεδομένα τα οποία δημιουργήθηκαν χρησιμοποιώντας μια τεχνική[18],[15] που ανακατασκευάζει σε ρεαλιστικό βαθμό τις παρατηρήσεις ακουστικών σκηνών σε οικιακά περιβάλλοντα. Στο κεφάλαιο που ακολουθεί θα περιγραφούν οι τεχνικές πληροφορίες της βάσης καθώς και ο τρόπος δημιουργίας των προσομοιωμένων δεδομένων.

3.1 Τεχνικές πληροφορίες

Η συλλογή DIRHA (SimCorpus) είναι μία πολυμικροφωνική και πολυγλωσσική βάση δεδομένων που περιλαμβάνει πραγματικές και προσομοιωμένες ακουστικές ακολουθίες όπως αυτές ηχογραφήθηκαν από το προαναφερθέν περιβάλλον του διαμερίσματος. Για κάθε γλώσσα η συλλογή δεδομένων περιέχει ακουστικές ακολουθίες διάρκειας 60 δευτερολέπτων με 48kHz συχνότητα δειγματοληψίας και $16 - bit$ ακρίβειας που παρατηρήθηκαν από 40 μικρόφωνα κατανεμημένα σε πέντε δωμάτια. Η κατανομή των μικροφώνων στα δωμάτια καθώς και η κάτοψη του διαμερίσματος μπορεί να φανεί στην Εικόνα 3.2. Οι συστοιχίες των μικροφώνων που χρησιμοποιήθηκαν είναι βασισμένες σε κατανεμημένα δίκτυα μικροφώνων (δυάδες ή τριάδες αισθητήρων στους τοίχους) καθώς και πιο συμπαγείς συστοιχίες (στο ταβάνι του καθιστικού και της κουζίνας). Σε κάθε ζεύγος οι αισθητήρες έχουν 30 εκατοστά απόστασης μεταξύ τους, σε αντίθεση με τις τριάδες που έχουν 15. Οι αισθητήρες ταβανιού αποτελούνται από έξι μικρόφωνα πέντε εκ των οποίων είναι τοποθετημένα κυκλικά με ακτίνα 30 εκατοστών ενώ ένας αισθητήρας είναι τοποθετημένος στο κέντρο του κύκλου.

3.2 Η προσομοιωμένη συλλογή δεδομένων

Το κύριο στοιχείο για την δημιουργία των προσομοιωμένων δεδομένων είναι η μέτρηση της κρουστικής απόκρισης (impulse response) που πραγματοποιήθηκε στο διαμέρισμα. Οι μετρήσεις των κρουστικών αποκρίσεων έγιναν χρησιμοποιώντας διαφορετικές θέσης και κατευθύνσεις των ηχείων με στόχο την δημιουργία δεδομένων με πλούσια χωρική μεταβλητότητα. Για τις μετρήσεις των κρουστικών αποκρίσεων χρησιμοποιήθηκαν ακολουθίες εκθετικού ημίτονου σάρωσης (Exponential Sine Sweep)[14] για την διέγερση του εκάστοτε δωματίου. Συνολικά μετρήθηκαν πάνω από 9000 κρουστικές αποκρίσεις για την δημιουργία της βάσης (Πίνακας 3.1).

Επιπρόσθετα, ηχογραφήθηκαν δεδομένα καθαρής ομιλίας (χωρίς θόρυβο από το περιβάλλον του δωματίου) τα οποία συνελίχθησαν (convolved) μαζί με τις κρουστικές αποκρίσεις για την δημιουργία των προσομοιωμένων δεδομένων. Για τις ηχογραφήσεις χρησιμοποιήθηκαν για κάθε γλώσσα 20 με 30 ομιλητές, ηλικίας μεταξύ 25 και 50 ετών οι οποίοι αναπαρήγαγαν τις φωνητικές ακολουθίες που θα παρουσιαστούν παρακάτω. Ακόμη, για να ενισχυθεί η ρεαλιστικότητα της βάσης, ηχογραφήθηκαν ήχοι από διάφορες οικιακές συσκευές καθώς και επιλέχθηκαν παρασκηνιακοί ήχοι από αντίστοιχες βάσεις (Freesound, Apple Logic Pro).
Κάθε ακουστική ακολουθία περιλαμβάνει:

- Θόρυβο από το περιβάλλον ηχογράφησης
- Μία κωδική λέξη που ακολουθείται από μία εντολή
- Μία φωνητική εντολή (χωρίς κωδική λέξη)
- Μία φωνητικά πλούσια πρόταση
- Ένα τμήμα διαλόγου
- Ένας μεταβλητός αριθμός από τοπικά ακουστικά γεγονότα που δεν αποτελούν ομιλία (ραδιόφωνο, τηλεόραση, οικιακές συσκευές, κτλ)

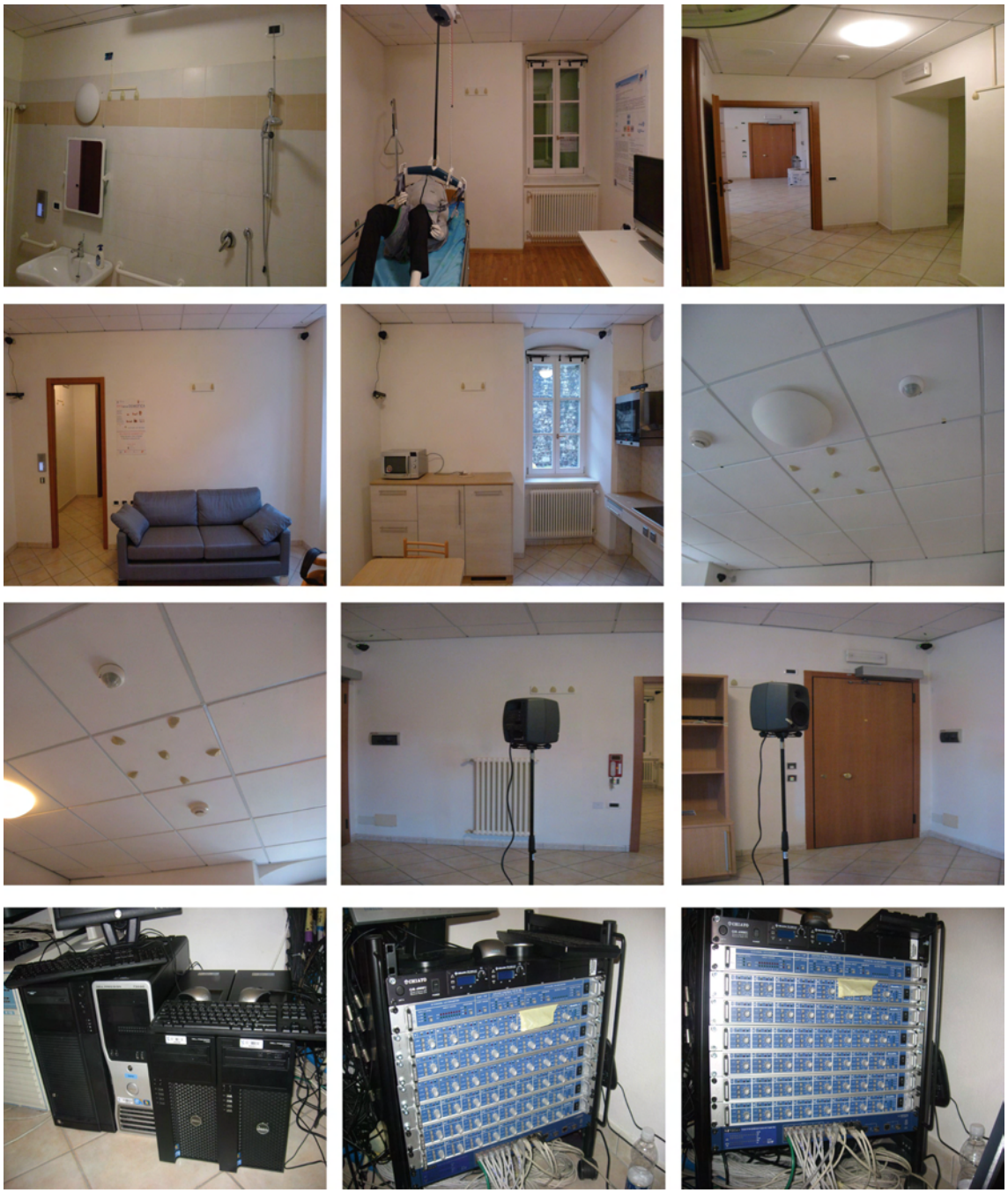
Κάθε ακολουθία επαναλήφθηκε σε τέσσερις γλώσσες διατηρώντας τους ίδιους παρασκηνιακούς θορύβους και της πηγές που δεν αποτελούν ομιλία. Το γένος και οι χρονισμοί των ενεργών ομιλητών διατηρήθηκαν σε όλες τις γλώσσες για να εξασφαλιστεί η ομοιομορφία της βάσης.

Δωμάτιο	Αριθμός μικροφώνων	Διαθέσιμες θέσεις	Κρουστικές αποκρίσεις	$T_{60}(s)$
Καθιστικό	15	18	2960	0.74
Κουζίνα	13	18	2969	0.83
Υπνοδωμάτιο	7	14	2160	0.68
Μπάνιο	3	4	640	0.75
Διάδρομος	2	3	480	0.60

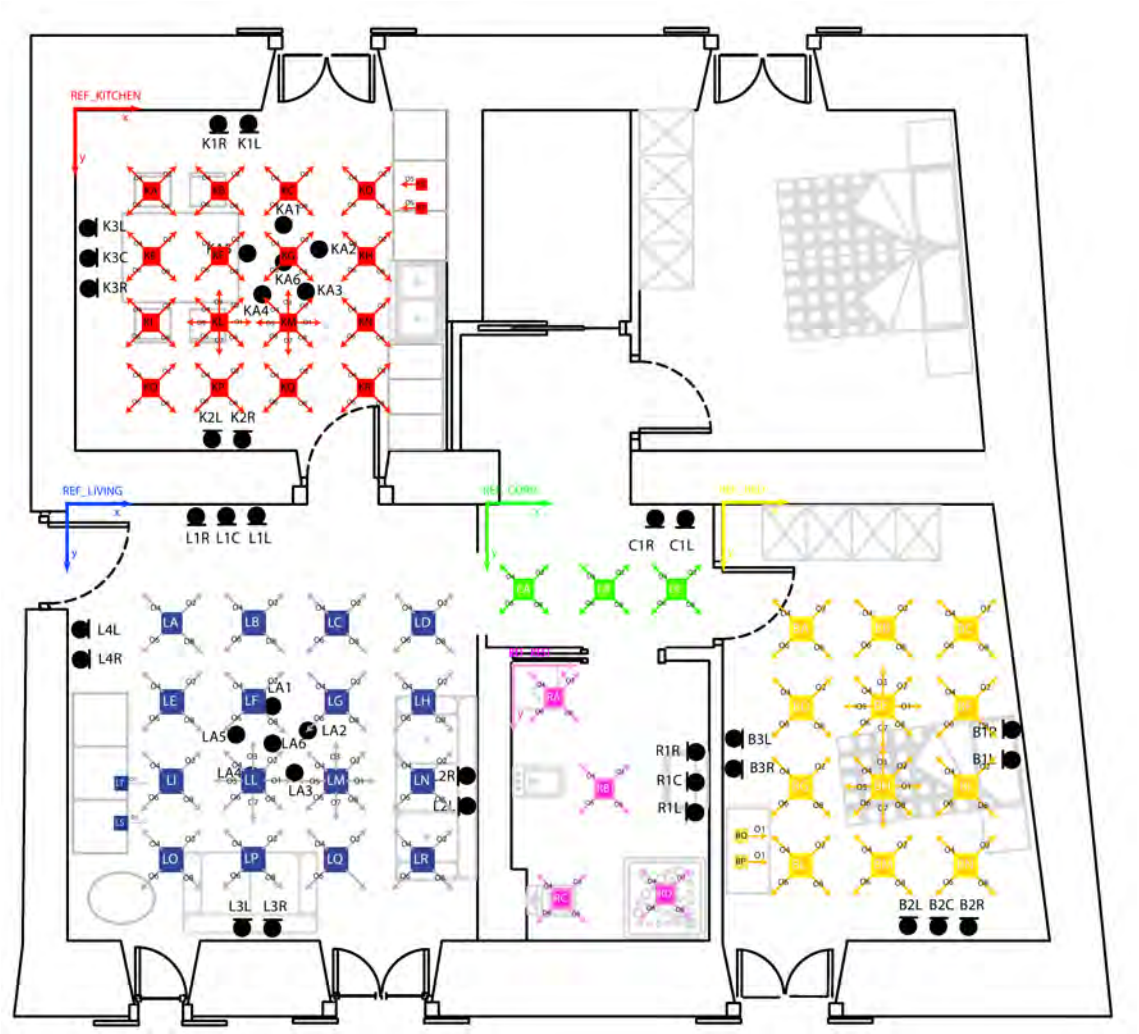
Πίνακας 3.1: Μετρήσεις κρουστικών αποκρίσεων στα αντίστοιχα δωμάτια,η τελευταία στήλη περιλαμβάνει τον εκτιμώμενο χρόνο αντήχησης T_{60} από [22].

3.3 Η διαδικασία προσομοίωσης

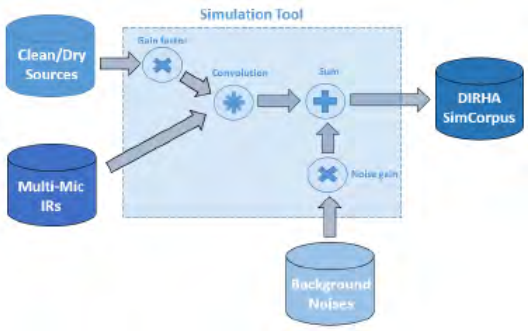
Η μέθοδος της "μόλυνσης" (contamination) χρησιμοποιήθηκε για την παραγωγή των προσομοιωμένων δεδομένων όπως αυτή απεικονίζεται στην Εικόνα 3.3. Αρχικά, επιλέγεται ένα σύνολο από ακουστικά γεγονότα (ομιλία ή τυπικοί οικιακοί ήχοι), για κάθε πηγή επιλέγεται μία τυχαία τοποθεσία στον χώρο και μετά πραγματοποιείται η συνέλιξη (convolution) του καθαρού σήματος (χρησιμοποιώντας ένα τυχαίο κέρδος (gain) μεταξύ ορίων) με την αντίστοιχη κρουστική απόκριση του δωματίου. Για την ενίσχυση της ρεαλιστικότητας των ακουστικών ακολουθιών προστίθενται παρασκηνιακοί θόρυβοι διαφόρων κερδών (noise gain). Αξίζει να σημειωθεί πως τμήματα ομιλίας μπορεί να επικαλύπτονται χρονικά με κάποια άλλη ακουστική πηγή (π.χ. ραδιόφωνο, τηλεόραση). Η επικάλυψη μεταξύ τμημάτων ομιλίας και μη-ομιλίας επιτρέπεται στο ίδιο και σε διαφορετικό δωμάτιο ενώ η επικάλυψη μεταξύ ακουστικών πηγών ομιλίας εμφανίζεται κυρίως σε διαφορετικά δωμάτια.



Σχήμα 3.1: Εικόνες των δωματίων του διαμερίσματος του προγράμματος DIRHA, όπου μπορούν να φανούν οι συστοιχίες μικροφώνων τοίχου και ταβανιού καθώς και περαιτέρω εξοπλισμός που χρησιμοποιήθηκε.



Σχήμα 3.2: Κάτοψη του διαμερίσματος από [22]. Οι μαύρες βούλες αντιπροσωπεύουν τις θέσεις των μικροφώνων και τα πολύχρωμα κουτιά αντιπροσωπεύουν τις θέσεις και τις κατευθύνσεις των ηχείων.



Σχήμα 3.3: Βασικό διάγραμμα για την δημιουργία της βάσης DIRHA SimCorpus [22].

Κεφάλαιο 4

Πειράματα και αποτελέσματα

4.1 Εισαγωγή

Στο παρόν κεφάλαιο θα γίνει η παρουσίαση του τρόπου διεξαγωγής των πειραμάτων, των εργαλείων που χρησιμοποιήθηκαν καθώς και των αποτελεσμάτων των πειραμάτων πάνω στα δεδομένα της βάσης DIRHA-SimCorpus. Πιο συγκεκριμένα θα πραγματοποιηθεί η αξιολόγηση των μεθοδολογιών που αναπτύχθηκαν στο κεφάλαιο 2 και πιο συγκεκριμένα οι μεθοδολογίες της γενικευμένης ετεροσυσχέτισης (GCC-PHat) και της κατευθυνόμενης δύναμης απόκρισης (SRP-PHat) με τους μετασχηματισμούς φάσης, στο έργο του εντοπισμού της ακουστικής πηγής. Στα πειράματα που πραγματοποιήθηκαν, χρησιμοποιήθηκαν τα δεδομένα από τους χώρους της κουζίνας και του καθιστικού μιας και σε αυτά τα δύο δωμάτια λαμβάνουν χώρα όλα τα ακουστικά γεγονότα του διαμερίσματος. Η αξιολόγηση των αποτελεσμάτων έγινε με την βοήθεια του εργαλείου βαθμολόγησης του προγράμματος DIRHA (evaluation tool) [27].

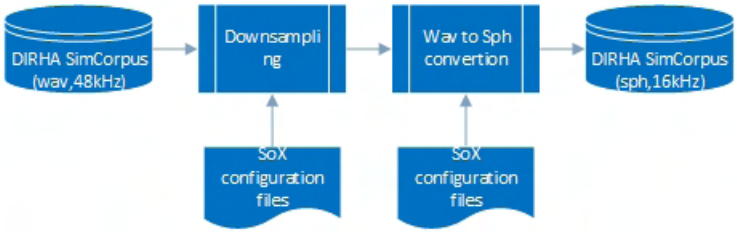
4.2 Προετοιμασία των δεδομένων

Πρώτου τα δεδομένα οδηγηθούν στα συστήματα εντοπισμού θέσης έπρεπε πρώτα να υποστούν επεξεργασία όσον αφορά την δομή και την αναπαράσταση τους. Αρχικά, αξίζει να αναφερθεί πως τα δεδομένα προσφέρονται από το πρόγραμμα DIRHA σαν αρχεία ήχου με μορφοποίηση “wav” στα 48kHz και πως ουσιαστικά ένα αρχείο αντιπροσωπεύει την παρατήρηση ενός συγκεκριμένου μικροφώνου. Για παράδειγμα στον χώρο της κουζίνας τα 13 μικρόφωνα παράγουν 13 ξεχωριστά αρχεία ήχου, συγχρονισμένα μεταξύ τους.

Το πρώτο βήμα της προετοιμασίας των δεδομένων είναι η αλλαγή της συχνότητας δειγματοληψίας τους. Για τον σκοπό αυτό χρησιμοποιήθηκε το εργαλείο επεξεργασίας αρχείων ήχου με την ονομασία “SoX” [26]. Το SoX είναι ένα λογισμικό κονσόλας, πολλαπλής πλατφόρμας, που μπορεί να αλλάξει την μορφοποίηση αρχείων ήχου καθώς και να εφαρμόσει σε αυτά διάφορα εφέ. Το παραπάνω εργαλείο σε συνδυασμό με μία ρουτίνα κονσόλας (bash script) χρησιμοποιήθηκαν για την διαπέραση του συστήματος αρχείων και για την αλλαγή της συχνότητας από τα 48 στα 16kHz.

Επιπρόσθετα, απαραίτητη ήταν η αλλαγή της μορφής (format) των αρχείων από “wav” σε “sph”. Τα αρχεία τύπου “sph” (Sphere waveform) είναι ψηφιακές ηχογραφήσεις κάποιας ακουστικής εκπομπής (broadcast) και χρησιμοποιούνται σαν είσοδο σε συστήματα αναγνώρισης ομιλίας. Η παραπάνω μετατροπή έγινε για ακόμη μία φόρα με την ανάπτυξη μίας ρουτίνας κονσόλας σε συνδυασμό με το εργαλείο “SoX”. Μετά το τέλος των παρα-

πάνω διαδικασιών έχει ουσιαστικά δημιουργηθεί ένα αντίγραφο της αρχικής βάσης σε διαφορετική μορφή και συχνότητα δειγματοληψίας (Εικόνα 4.1).



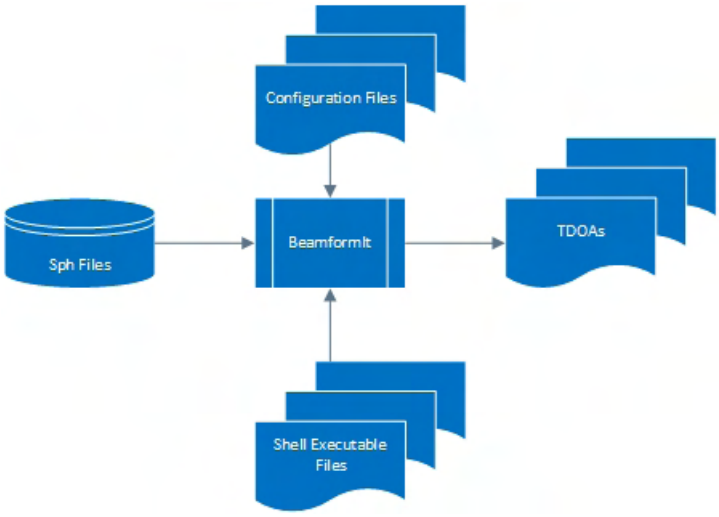
Σχήμα 4.1: Βασικό διάγραμμα για την προετοιμασία των δεδομένων.

4.2.1 Υπολογισμός των χρονικών καθυστερήσεων με τον δημιουργό δέσμης BeamformIt

Για τον υπολογισμό των χρονικών καθυστερήσεων άφιξης (TDOA) σε καθένα από τους αισθητήρες χρησιμοποιήθηκε το εργαλείο BeamformIt [3][2]. Το προαναφερθέν λογισμικό είναι ένας δημιουργός δέσμης (beamformer) ο οποίος αναπτύχθηκε με σκοπό την αναγνώριση ομιλητών σε περιβάλλοντα διασκέψεων και συνεδρίων. Το BeamformIt δέχεται σαν είσοδο έναν αριθμό από κανάλια (σε sph μορφή) και παράγει την έξοδο του χρησιμοποιώντας την μέθοδο φιλτραρίσματος-και-άθροισης (Ενότητα 2.8). Δίνοντας κατάλληλες εισόδους σε συνδυασμό με τα ανάλογα αρχεία ρυθμίσεων (configuration files) το παραπάνω λογισμικό μπορεί να παράγει τις χρονικές καθυστερήσεις μεταξύ των μικροφώνων στο δοθέν χώρο. Παρακάτω μπορούν να φανούν οι σημαντικότερες από τις ρυθμίσεις που χρησιμοποιήθηκαν:

- scroll_size [250], το μέγεθος της κύλισης (scrolling) που χρησιμοποιείται για την εφαρμογή των καθυστερήσεων και την εξαγωγή του σήματος.
- window_size [500], το παράθυρο για τον υπολογισμό της ετεροσυσχέτισης.
- do_compute_reference [0], δήλωση χρήσης αναφοράς σαν παράμετρο.
- reference_channel [i], δήλωση του καναλιού i που θα χρησιμοποιηθεί σαν κανάλι αναφοράς. Σε κάθε εφαρμογή του λογισμικού η μεταβλητή i παίρνει διαφορετικές τιμές.

Πιο συγκεκριμένα αναπτύχθηκαν δύο συστήματα (για καθένα από τα δύο δωμάτια που βρίσκονται υπό μελέτη) για την παραγωγή των χρονικών καθυστερήσεων. Σύμφωνα με την εικόνα 3.2 μπορούμε να δημιουργήσουμε τον πίνακα 4.1 με τα αντίστοιχα αναγνωριστικά των μικροφώνων. Ξεκινώντας με το δωμάτιο του καθιστικού, δημιουργήθηκαν τέσσερα αρχεία ρυθμίσεων (configuration files) έτσι ώστε κάθε ζεύγος/τριάδα μικροφώνων να έχει το αντίστοιχο μικρόφωνο αναφοράς σύμφωνα με το οποίο θα υπολογιστούν οι χρονικές καθυστερήσεις των υπολοίπων που βρίσκονται στην ίδια συστοιχία. Ουσιαστικά το αποτέλεσμα είναι η απομόνωση κάθε συστοιχίας από τις υπόλοιπες. Για παράδειγμα, στην συστοιχία L3L,L3R χρησιμοποιώντας σαν μικρόφωνο αναφοράς (reference) το μικρόφωνο με αναγνωριστικό “L3L” παίρνουμε ως αποτέλεσμα την χρονική καθυστέρηση όλων των υπολοίπων μικροφώνων σε σχέση με το “L3L”. Επαναλαμβάνοντας την ίδια μέθοδο για τα υπόλοιπα ζεύγη μικροφώνων και χρησιμοποιώντας πάντα τον αριστερότερο αισθητήρα (“L” σαν τελευταίο γράμμα του αναγνωριστικού) ως μικρόφωνο αναφοράς λαμβάνουμε ως έξοδο από το εργαλείο BeamformIt τα αρχεία που περιλαμβάνουν



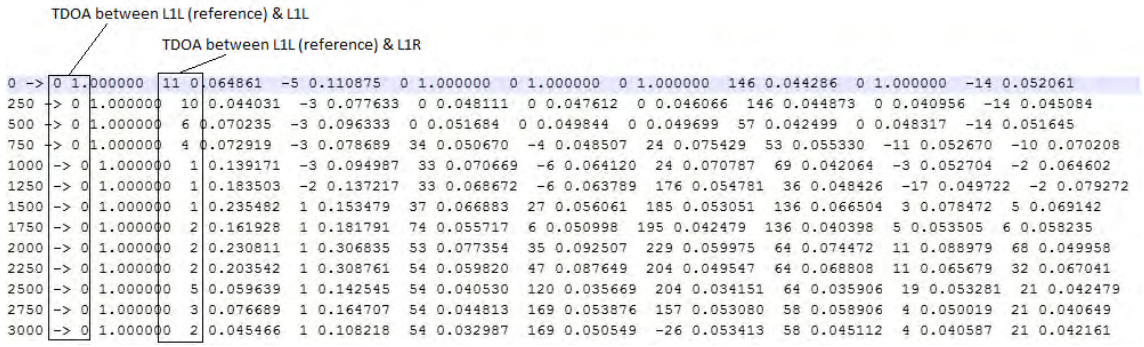
Σχήμα 4.2: Διάγραμμα παραγωγής των χρονικών καθυστερήσεων μεταξύ των μικροφώνων με το λογισμικό BeamformIt.

τις χρονικές καθυστερήσεις. Αξίζει να σημειωθεί πως στην περίπτωση των συστοιχιών με τρεις αισθητήρες (για παράδειγμα L1R,L1C,L1L) το κεντρικό μικρόφωνο (L1C) δεν χρησιμοποιήθηκε στον υπολογισμό των χρονικών καθυστερήσεων. Ένα τυπικό παράδειγμα εξόδου μπορεί να φανεί στην εικόνα 4.3.

Επαναλαμβάνοντας την παραπάνω διαδικασία και για το περιβάλλον της κουζίνας και στην συνέχεια για κάθε συνεδρία (session) της βάσης DIRHA SimCorpus έχουμε ως αποτέλεσμα τις σχετικές καθυστερήσεις κάθε μικροφώνου αναφοράς με το δεξιότερο (σε σχέση με αυτό) της συστοιχίας στην οποία ανήκουν.

	Αναγνωριστικά μικροφώνων			
Δωμάτιο	Βόρεια	Νότια	Ανατολικά	Δυτικά
Κουζίνα	K1R,K1L	K2L,K2R	-	K3L,K3C,K3R
Καθιστικό	L1R,L1C,L1L	L3L,L3R	L2L,L2R	L4L,L4R

Πίνακας 4.1: Αναγνωριστικά μικροφώνων σε σχέση με την θέση τους στο κάθε δωμάτιο.




```
0 -> 0 1.000000  11 0.064861  -5 0.110875  0 1.000000  0 1.000000  0 1.000000  146 0.044286  0 1.000000  -14 0.052061
250 -> 0 1.000000  10 0.044031  -3 0.077633  0 0.048111  0 0.047612  0 0.046066  146 0.044873  0 0.040956  -14 0.045084
500 -> 0 1.000000  6 0.070235  -3 0.096333  0 0.051684  0 0.049844  0 0.049699  57 0.042499  0 0.048317  -14 0.051645
750 -> 0 1.000000  4 0.072919  -3 0.078689  34 0.050670  -4 0.048507  24 0.075429  53 0.055330  -11 0.052670  -10 0.070208
1000 -> 0 1.000000  1 0.139171  -3 0.094987  33 0.070669  -6 0.064120  24 0.070787  69 0.042064  -3 0.052704  -2 0.064602
1250 -> 0 1.000000  1 0.183503  -2 0.137217  33 0.068672  -6 0.063789  176 0.054781  36 0.048426  -17 0.049722  -2 0.079272
1500 -> 0 1.000000  1 0.235482  1 0.153479  37 0.066883  27 0.056061  185 0.053051  136 0.066504  3 0.078472  5 0.069142
1750 -> 0 1.000000  2 0.161928  1 0.181791  74 0.055717  6 0.050998  195 0.042479  136 0.040398  5 0.053505  6 0.058235
2000 -> 0 1.000000  2 0.230811  1 0.306835  53 0.077354  35 0.092507  229 0.059975  64 0.074472  11 0.088979  68 0.049958
2250 -> 0 1.000000  2 0.203542  1 0.308761  54 0.059820  47 0.087649  204 0.049547  64 0.068808  11 0.065679  32 0.067041
2500 -> 0 1.000000  5 0.059639  1 0.142545  54 0.040530  120 0.035669  204 0.034151  64 0.035906  19 0.053281  21 0.042479
2750 -> 0 1.000000  3 0.076689  1 0.164707  54 0.044813  169 0.053876  157 0.053080  58 0.058906  4 0.050019  21 0.040649
3000 -> 0 1.000000  2 0.045466  1 0.108218  54 0.032987  169 0.050549  -26 0.053413  58 0.045112  4 0.040587  21 0.042161
```

Σχήμα 4.3: Παράδειγμα εξόδου του λογισμικού BeamformIt με μικρόφωνο αναφοράς το "L1L".

4.3 Αξιολόγηση της μεθοδολογίας γενικευμένης ετεροσυσχέτισης

4.3.1 Υπολογισμός των κατευθύνσεων άφιξης

Μετά τον υπολογισμό όλων των χρονικών καθυστερήσεων (TDOA) σειρά έχει ο υπολογισμός της κατεύθυνσης άφιξης (DOA). Δεδομένου των σταθερών συντεταγμένων των μικροφώνων σε κάθε χώρο, αναπτύχθηκε μία ρουτίνα σε περιβάλλον MATLAB η οποία αναλύει τα αρχεία των καθυστερήσεων που παρήχθησαν στο προηγούμενο βήμα και τα αποθηκεύει σε μορφή διανυσμάτων. Επομένως, για κάθε συνεδρία της βάσης έχουμε τέσσερα διανύσματα καθυστερήσεων για το καθιστικό και τρία για την κουζίνα (ένα διάνυσμα για κάθε ξεχωριστή συστοιχία). Χρησιμοποιώντας ως σταθερές την συχνότητα δειγματοληψίας ($Fs = 16$kHz) καθώς και την ταχύτητα του ήχου στο κενό ($c = 343$ m/s) μπορούμε να υπολογίσουμε την γωνία άφιξης που ορίζεται ως προς το κέντρο του ευθύγραμμου τμήματος που ενώνει τα δύο μικρόφωνα με φορά από το δεξιό προς το αριστερό (Αλγόριθμος 1).

Data: διανύσματα καθυστερήσεων, διανύσματα αποστάσεων μεταξύ μικροφώνων
Result: διάνυσμα των γωνιών άφιξης
for *για κάθε διάνυσμα καθυστέρησης* j **do**
  for *για κάθε στοιχείο των διανυσμάτων καθυστερήσεων* i **do**
    θ(i,j)= $cos^{-1}(\frac{C*TDOA(i)}{Fs*distance})$;
  end
end

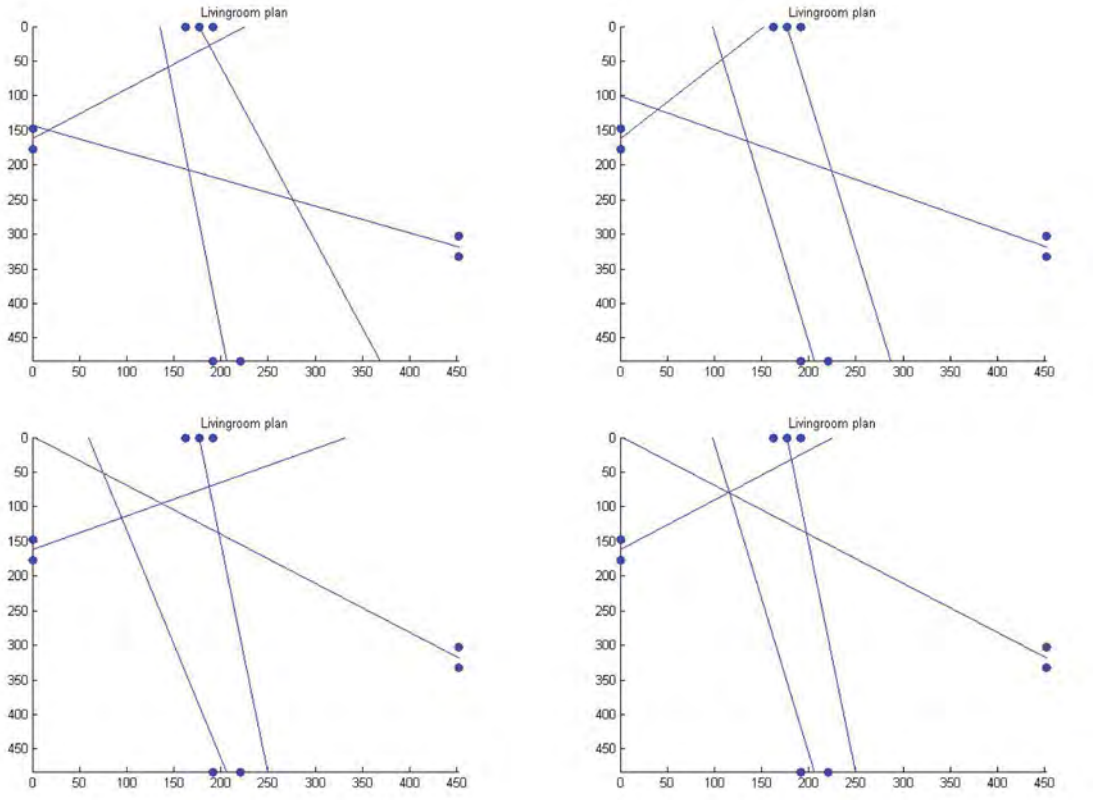
Algorithm 1: Υπολογισμός της σχετικής γωνίας άφιξης, η μεταβλητή "distance" προσδιορίζει την πραγματική απόσταση μεταξύ των μικροφώνων σε εκατοστά και η "TDOA" τα διανύσματα των χρονικών καθυστερήσεων για κάθε συστοιχία.

Γνωρίζοντας την γωνία άφιξης για κάθε συστοιχία καθώς και τις συντεταγμένες των σημείων μεταξύ των μικροφώνων (x_0, y_0) μπορούμε να υπολογίσουμε την εξίσωση της ευθείας που θα ενώνει την κάθε συστοιχία με την εκτιμώμενη ακουστική πηγή χρησιμοποιώντας την παρακάτω εξίσωση:

$$y_0 = tan(\theta)x_0 + b \tag{4.1}$$

$$b = y_0 - tan(\theta)x_0 \tag{4.2}$$

Είναι σημαντικό να αναφερθεί πως εξαιτίας της γεωμετρίας των δωματίων είναι απαραίτητη η πρόσθεση 90° μοιρών στις γωνίες άφιξης των συστοιχιών που βρίσκονται στα δυτικά και ανατολικά των δωματίων. Αυτό έχει σαν αποτέλεσμα το ταίριασμα του επιπέδου των μικροφώνων με το σύστημα αναφοράς του εκάστοτε δωματίου καθώς και καλύτερη οπτικοποίηση της ευθείας που ενώνει την πηγή με την συστοιχία. Στο σχήμα 4.4 μπορούν να φανούν τέσσερα στιγμιότυπα κατευθύνσεων άφιξης για τις τέσσερις συστοιχίες τοίχου του καθιστικού.



Σχήμα 4.4: Κατευθύνσεις άφιξης για τις τέσσερις συστοιχίες του καθιστικού, οι μπλε βούλες αντικατοπτρίζουν τα αντίστοιχα μικρόφωνα του δωματίου.

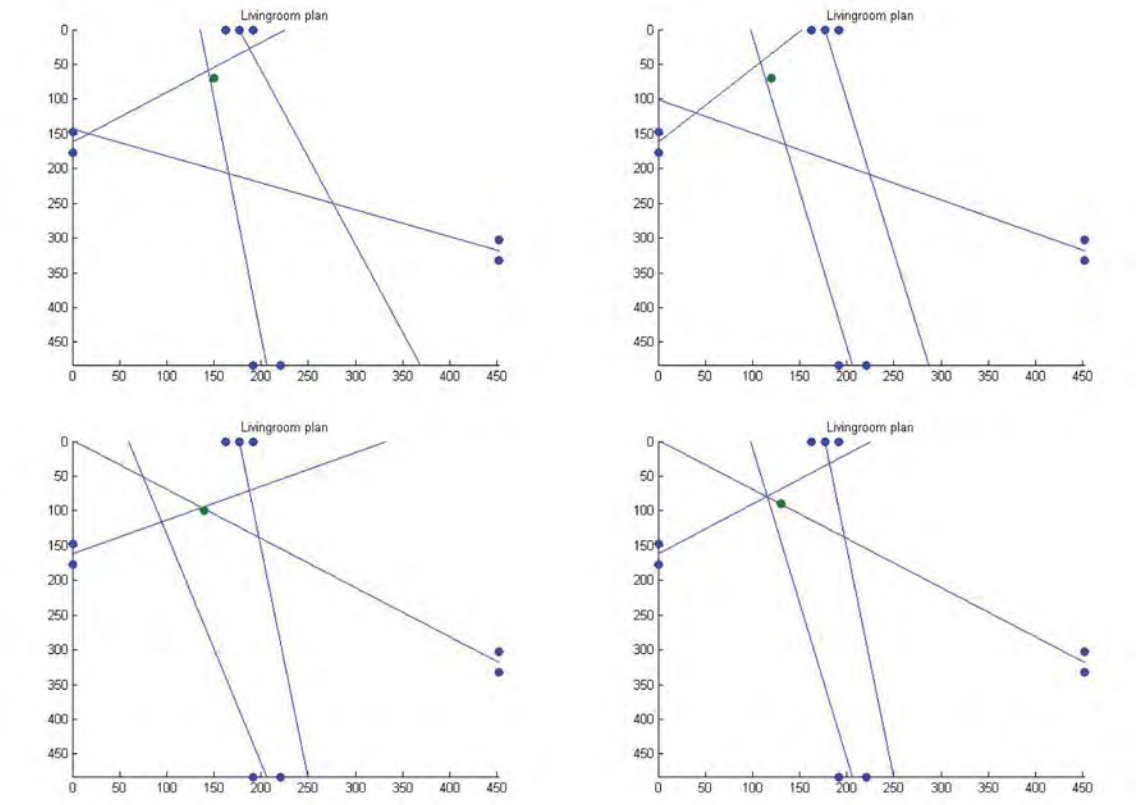
4.3.2 Εκτίμηση του σημείου της ακουστικής πηγής

Είναι εύκολο να φανταστεί κανείς, πως το σημείο που έχει την μικρότερη δυνατή απόσταση από όλες της ευθείες (DOA) που προέκυψαν είναι οι συντεταγμένες του σημείου της ακουστικής πηγής. Δεδομένου της ευθείας στην μορφή $ax + by + c = 0$ μπορούμε να

υπολογίσουμε την απόσταση d κάθε σημείου του δωματίου από την κάθε ευθεία χρησιμοποιώντας την παρακάτω εξίσωση:

$$d = \frac{|\alpha x_0 + \beta y_0 + c|}{\sqrt{a^2 + b^2}} \tag{4.3}$$

Με αυτόν τον τρόπο, διατρέχοντας όλο το πλέγμα του δωματίου ανά διαστήματα των 10 εκατοστών μπορούμε να εντοπίσουμε το σημείο που έχει την μικρότερη απόσταση από τις ευθείες. Εφαρμόζοντας αυτήν την λογική στα στιγμιότυπα εκτέλεσης του σχήματος 4.4 προκύπτει το σχήμα 4.5



Σχήμα 4.5: Εκτίμηση των συντεταγμένων της ακουστικής πηγής στο χώρο του καθιστικού, οι πράσινες βούλες αντικατοπτρίζουν την θέση της πηγής.

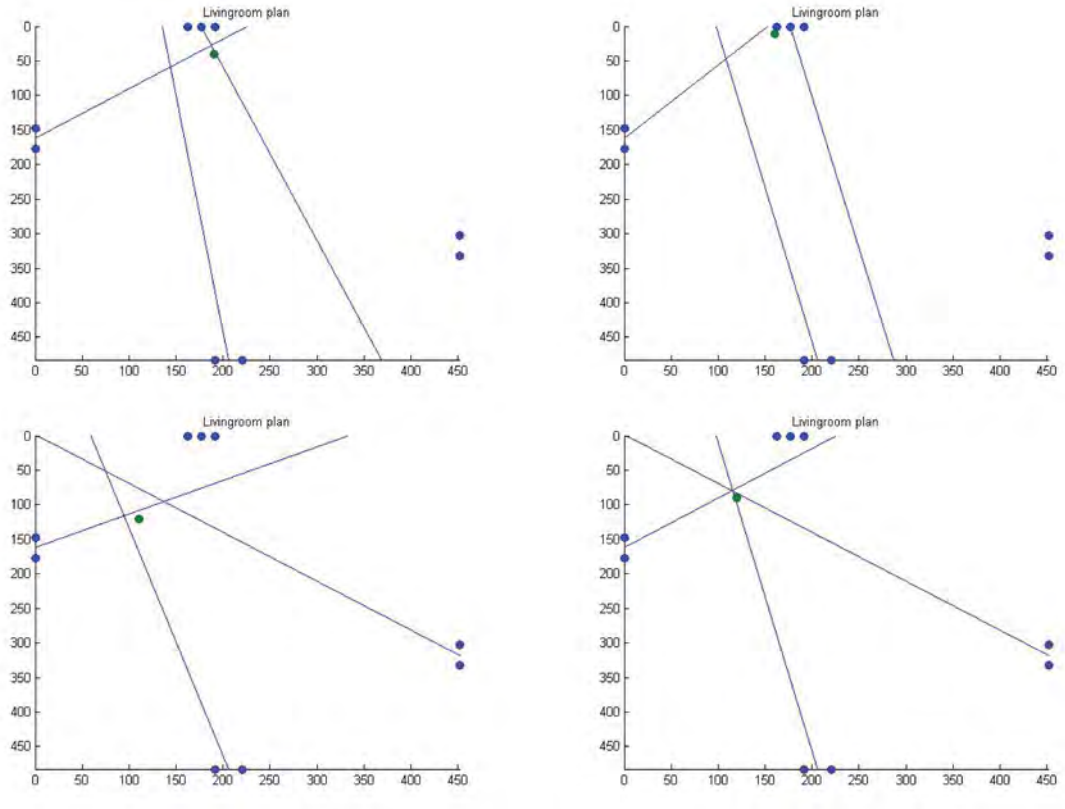
4.3.3 Βελτιστοποίηση στην εκτίμηση του σημείου της ακουστικής πηγής

Όπως έχει αναφερθεί προηγουμένως, το φαινόμενο της αντήχησης μπορεί να προκαλέσει σφάλματα στον υπολογισμό των χρονικών καθυστερήσεων άφιξης (TDOA) και κατά επέκταση στον υπολογισμό των κατευθύνσεων άφιξης (DOA). Αυτό το φαινόμενο μπορεί να φανεί ξεκάθαρα στα δύο πρώτα στιγμιότυπα των σχημάτων 4.4 και 4.5, όπου η εκτιμώμενη ευθεία της ανατολικής συστοιχίας έχει μεγάλη απόκλιση από τις υπόλοιπες. Για τον σκοπό αυτό αναπτύχθηκε ένας αλγόριθμος ο οποίος υλοποιεί την αφαίρεση της ευθείας που βρίσκεται σε μεγαλύτερη απόσταση από τις υπόλοιπες. Σε κάθε χρονικό καρέ μετά τον υπολογισμό των ευθειών, πραγματοποιείται η αποκοπή της πιο απομακρυσμένης ευθείας (εάν αυτή υπάρχει) σύμφωνα με τα παρακάτω:

Data: Αποστάσεις των ευθειών από την εκτιμημένη θέσης της πηγής
Result: Καινούρια θέσης της πηγής χωρίς τον υπολογισμό της απομακρυσμένης ευθείας
$maxDistance =$μέγιστη απόσταση ευθείας;
if $maxDistance > threshold$ **then**
  αποκοπή της ευθείας με το maxDistance και επανα-υπολογισμός της θέσης της πηγής
end

Algorithm 2: Αποκοπή της απομακρυσμένης ευθείας που συμβάλει καταστρεπτικά στην εκτίμηση της πηγής, η μεταβλητή "maxDistance" προσδιορίζει την ευθεία με την μεγαλύτερη απόσταση από την ήδη εκτιμημένη πηγή και η σταθερά "threshold" είναι ορισμένη στα 50 εκατοστά.

Η εφαρμογή του παραπάνω αλγορίθμου πάνω στα τέσσερα στιγμιότυπα του σχήματος 4.5 μπορεί να φανεί στο σχήμα 4.6

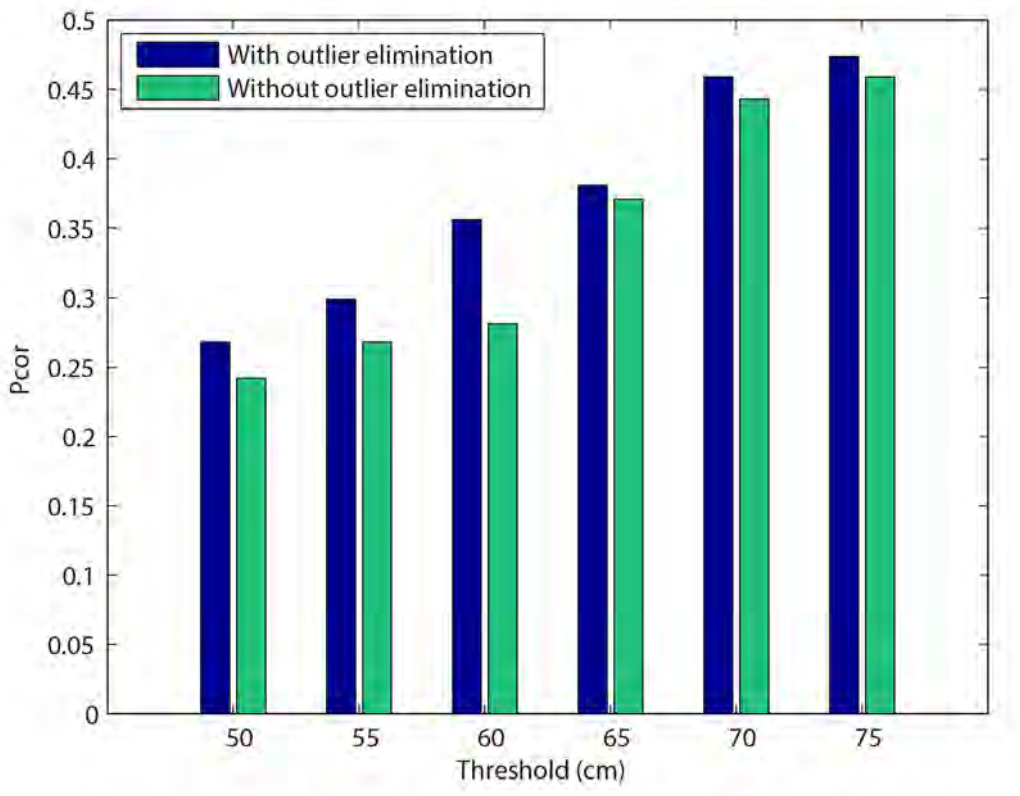


Σχήμα 4.6: Αποκοπή και υπολογισμός της καινούριας θέσης της πηγής, στα στιγμιότυπα από αριστερά προς τα δεξιά και από πάνω προς τα κάτω έχουμε τις αποκοπές της ανατολικής, ανατολικής, βόρειας και βόρειας ευθείας (DOA) αντίστοιχα.

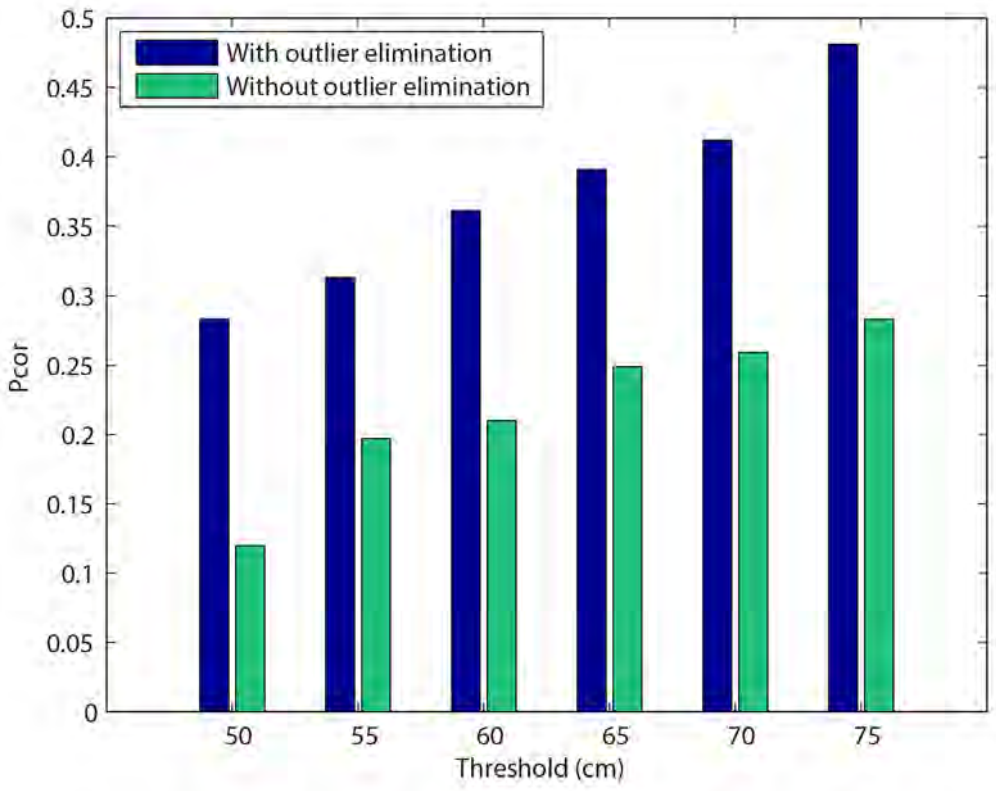
4.3.4 Αποτελέσματα της μεθόδου γενικευμένης ετεροσυσχέτισης

Όπως αναφέρθηκε και νωρίτερα, η αξιολόγηση των αποτελεσμάτων για τα δύο δωμάτια του διαμερίσματος έγινε με την βοήθεια του εργαλείου βαθμολόγησης που αναπτύχθηκε από το πρόγραμμα DIRHA. Για κάθε ακουστική σκηνή και για κάθε δωμάτιο μπορεί να βρεθεί το αντίστοιχο αρχείο αναφοράς (ground truth) το οποίο περιέχει τις πραγματικές θέσεις της πηγής όπως αυτές έχουν καταγραφεί στο διαμέρισμα κατά την πραγματοποίηση των πειραμάτων. Το εργαλείο βαθμολόγησης συγκρίνει τα αρχεία εξόδου που παρήχθησαν χρησιμοποιώντας τα συστήματα που περιγράφηκαν παραπάνω με τα αρχεία αναφοράς και παράγει ένα αρχείο εξόδου με τα στατιστικά στοιχεία της αξιολόγησης. Στην περίπτωση που οι συντεταγμένες που βρίσκονται υπό βαθμολόγηση είναι μέσα στο όριο (παράμετρος που δίνεται κατά την εκτέλεση του εργαλείου βαθμολόγησης, με προεπιλεγμένη τιμή τα 50 εκατοστά) σε σχέση με τις πραγματικές, τότε υπάρχει επιτυχία (fine error) ενώ σε διαφορετική περίπτωση υπάρχει αποτυχία (gross error). Η μετρική που δείχνει το ποσοστό επιτυχίας ($\frac{fineErrors}{Total}$) ονομάζεται Pcor.

Στο σχήμα 4.7 μπορεί να φανεί η απόδοση της μεθόδου γενικευμένης ετεροσυσχέτισης με τον μετασχηματισμό φάσης στον εντοπισμό της πηγής στο δισδιάστατο επίπεδο, στον χώρο του καθιστικού. Στο ίδιο σχήμα παρουσιάζεται η αλλαγή στην επίδοση χρησιμοποιώντας την βελτιστοποίηση με την αποκοπή των απομακρυσμένων ευθειών. Παρομοίως, το σχήμα 4.8 απεικονίζει την απόδοση της ίδιας μεθόδου στο δισδιάστατο επίπεδο στον χώρο της κουζίνας.



Σχήμα 4.7: Απόδοση της μεθόδου γενικευμένης ετεροσυσχέτισης στον χώρο του καθιστικού, με και χωρίς την χρήση της βελτιστοποίησης αποκοπής απομακρυσμένων ευθειών. Ο οριζόντιος άξονας αντικατοπτρίζει το όριο για να κριθεί επιτυχώς ο εντοπισμός της πηγής ενώ ο κάθετος τον λόγο των επιτυχιών ως προς όλα τα γεγονότα (Pcor).



Σχήμα 4.8: Απόδοση της μεθόδου γενικευμένης ετεροσυσχέτισης στον χώρο της κουζίνας, με και χωρίς την χρήση της βελτιστοποίησης αποκοπής απομακρυσμένων ευθειών. Ο οριζόντιος άξονας αντικατοπτρίζει το όριο για να κριθεί επιτυχώς ο εντοπισμός της πηγής ενώ ο κάθετος τον λόγο των επιτυχιών ως προς όλα τα γεγονότα (Pcor).

4.4 Αξιολόγηση της μεθοδολογίας κατευθυνόμενης δύναμης απόκρισης

Σε αντίθεση με τα πειράματα για την αξιολόγηση της μεθόδου γενικευμένης ετεροσυσχέτισης όπου χρησιμοποιήθηκε η μετασχηματισμένη βάση δεδομένων (μορφή sph, 16kHz συχνότητα δειγματοληψίας), στα πειράματα που διεξήχθησαν για την αξιολόγηση της κατευθυνόμενης δύναμης απόκρισης επιστρατεύτηκε η βάση στην αρχική μορφή της (μορφή wav, 48kHz συχνότητα δειγματοληψίας).

4.4.1 Εκτίμηση του σημείου της ακουστικής πηγής

Για την παραγωγή των αποτελεσμάτων χρησιμοποιήθηκε η ρουτίνα που αναπτύχθηκε από [1],[12]. Η προαναφερθείσα ρουτίνα χρησιμοποιεί την κατευθυνόμενη δύναμη απόκρισης σε συνδυασμό με τον αλγόριθμο βελτιστοποίησης της στοχαστικής συστολής περιφέρειας (SRC ή stochastic region contraction) για τον εντοπισμό μίας μοναδικής ακουστικής πηγής χρησιμοποιώντας ένα πλαίσιο (frame) από δεδομένα και έναν αριθμό από μικρόφωνα.

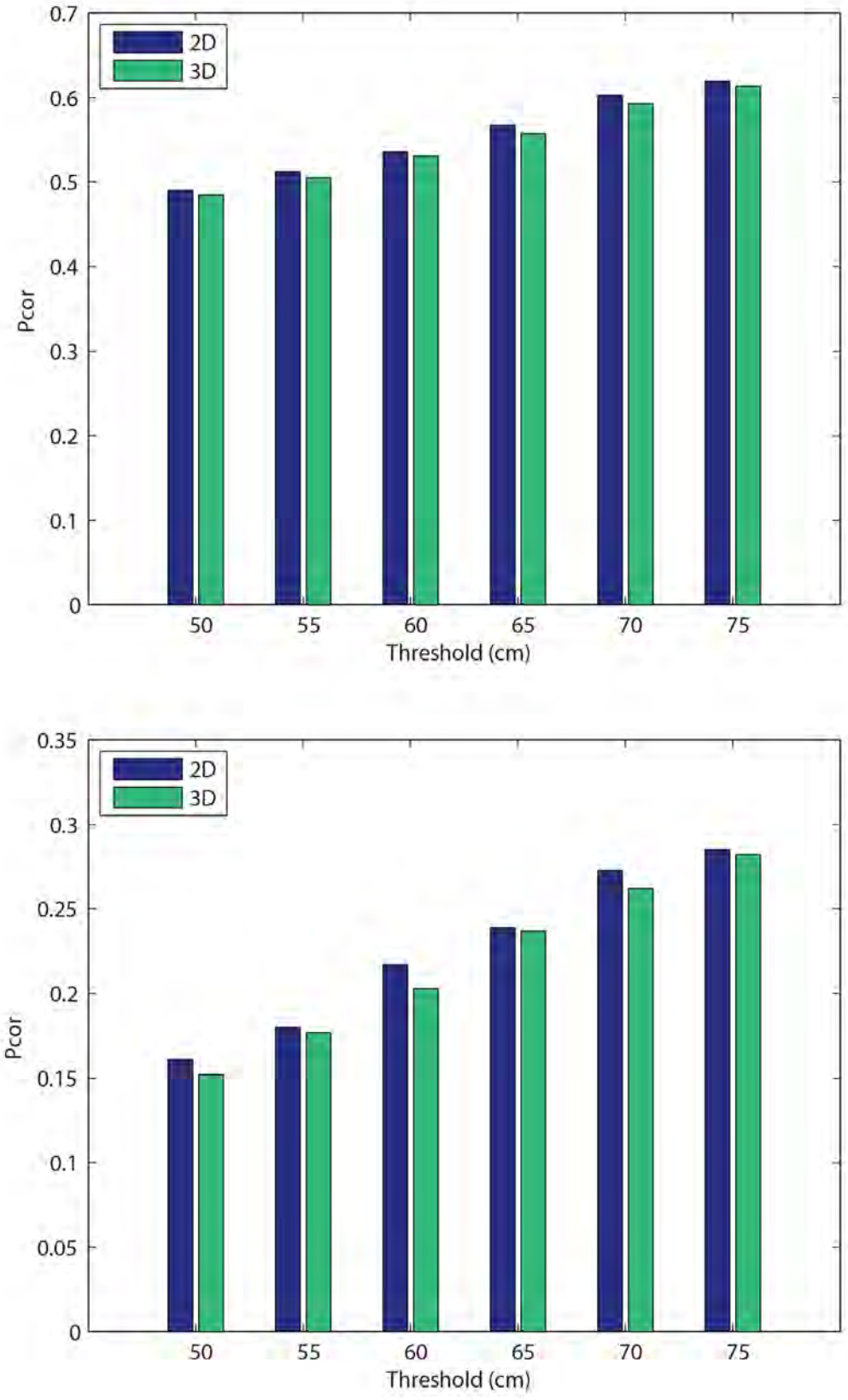
Επιπρόσθετα, κρίθηκε αναγκαία η ανάπτυξη μίας βοηθητικής ρουτίνας για την συλλογή των δεδομένων (ηχητικών σημάτων και συντεταγμένες των μικροφώνων στον χώρο)

καθώς και την είσοδο τους στο κυρίως σύστημα. Αυτή η ανάγκη προήλθε από το γεγονός του ότι η ρουτίνα που χρησιμοποιήθηκε αντιμετώπιζε όλο τον όγκο των δεδομένων σαν μία ακολουθία από την οποία έβγαζε σαν αποτέλεσμα ένα μοναδικό ζεύγος συντεταγμένων. Επομένως, ο βασικός στόχος της βοηθητικής ρουτίνας είναι ουσιαστικά να δημιουργεί υποσύνολα των παρατηρήσεων των μικροφώνων κάθε 0.05 δευτερόλεπτα ή πιο συγκεκριμένα κάθε 2400 δείγματα, δεδομένου της συχνότητας δειγματοληψίας (48kHz) και να τα προωθεί σαν είσοδο στο σύστημα εντοπισμού πηγής.

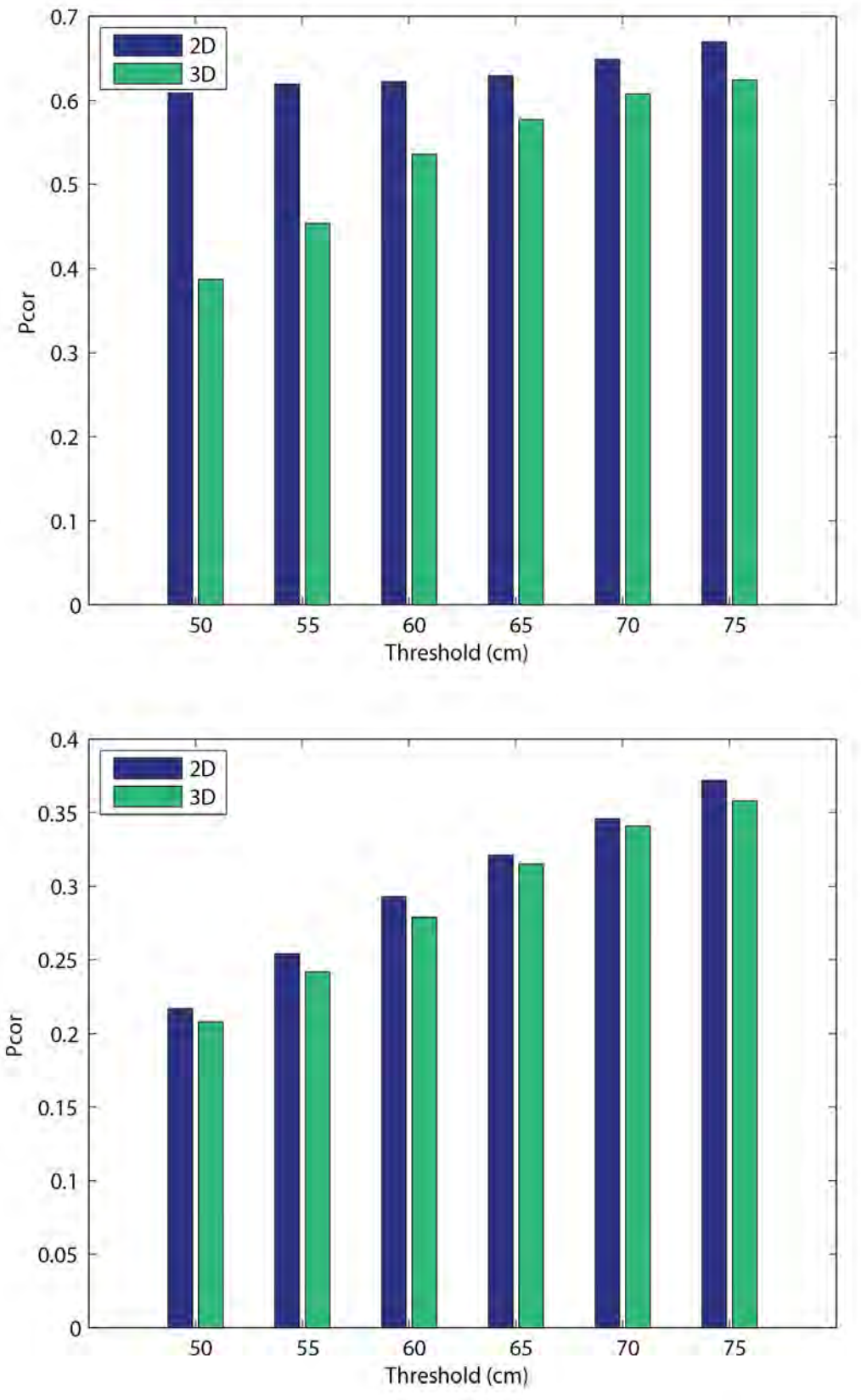
Αρχικά, μελετήθηκε η συμπεριφορά της μεθόδου χρησιμοποιώντας την πληροφορία που παρήγαγαν αποκλειστικά οι συστοιχίες μικροφώνων τοίχου, όπως στην περίπτωση των πειραμάτων γενικευμένης ετεροσυσχέτισης, και στην συνέχεια μελετήθηκε η απόδοση της μεθόδου χρησιμοποιώντας την επιπλέον πληροφορία από τις συστοιχίες ταβανιού με σκοπό την βελτίωση των αποτελεσμάτων. Συνολικά, στο πρώτο σετ πειραμάτων χρησιμοποιήθηκε η πληροφορία από 9 και 7 μικρόφωνα για τους χώρους του καθιστικού και κουζίνας αντίστοιχα και στην συνέχεια 15 και 13 μικροφώνων χρησιμοποιώντας τις εξάδες μικροφώνων στα αντίστοιχα ταβάνια των δωματίων.

4.4.2 Αποτελέσματα της μεθόδου κατευθυνόμενης δύναμης απόκρισης

Η αξιολόγηση των αποτελεσμάτων έγινε για ακόμη μια φορά, χρησιμοποιώντας το εργαλείο βαθμολόγησης DIRHA όπως στην ενότητα 4.3.4. Στο σχήμα 4.9 μπορεί να φανεί η απόδοση της μεθόδου κατευθυνόμενης δύναμης απόκρισης χρησιμοποιώντας τις συστοιχίες τοίχου στους χώρους του καθιστικού και κουζίνας αντίστοιχα, ενώ στο 4.10 απεικονίζεται η απόδοση της μεθοδολογίας χρησιμοποιώντας την επιπλέον πληροφορία από τα μικρόφωνα τοποθετημένα στο ταβάνι.



Σχήμα 4.9: Απόδοση της μεθόδου κατευθυνόμενης δύναμης απόκρισης στους χώρους του καθιστικού (πάνω) και της κουζίνας (κάτω) χωρίς τις συστοιχίες ταβανιού. Ο οριζόντιος άξονας αντικατοπτρίζει το όριο για να κριθεί επιτυχώς ο εντοπισμός της πηγής ενώ ο κάθετος τον λόγο των επιτυχιών ως προς όλα τα γεγονότα (Pcor).



Σχήμα 4.10: Απόδοση της μεθόδου κατευθυνόμενης δύναμης απόκρισης στους χώρους του καθιστικού (πάνω) και της κουζίνας (κάτω) με την χρήση της επιπρόσθετης πληροφορίας από τις συστοιχίες ταβανιού. Ο οριζόντιος άξονας αντικατοπτρίζει το όριο για να κριθεί επιτυχώς ο εντοπισμός της πηγής ενώ ο κάθετος τον λόγο των επιτυχιών ως προς όλα τα γεγονότα (Pcor).

Κεφάλαιο 5

Συμπεράσματα

5.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία παρουσιάστηκε το πρόβλημα του εντοπισμού ακουστικής πηγής σε "έξυπνα" περιβάλλοντα με την χρήση πολύ-μικροφωνικών συστοιχιών. Η εργασία ξεκίνησε με την περιγραφή του προβλήματος καθώς και συναφείς εργασίες από την ερευνητική κοινότητα στον τομέα της ανάλυσης ακουστικής σκηνής. Στην συνέχεια ακολούθησε η μοντελοποίηση του προβλήματος και το μαθηματικό/θεωρητικό υπόβαθρο των μεθοδολογιών καθώς και η μελέτη της βάσης δεδομένων που χρησιμοποιήθηκε για την διεξαγωγή των πειραμάτων. Η εργασία συνεχίστηκε με την ανάλυση του τρόπου διεξαγωγής των πειραμάτων και την παρουσίαση των αποτελεσμάτων τους.

Εξετάζοντας τα αποτελέσματα των μπορεί να φανεί ξεκάθαρα πως η μέθοδος της κατευθυνόμενης δύναμης απόκρισης (SRP-PHat) μπορεί να εκτιμήσει με πιο αποτελεσματικό τρόπο την θέσης της ακουστικής πηγής στον χώρο σε σχέση με την μέθοδο γενικευμένης ετεροσυσχέτισης (GCC-PHat) και πως η χρήση της επιπλέον πληροφορίας χρησιμοποιώντας τις παρατηρήσεις από τα μικρόφωνα ταβανιού, προκαλεί σημαντική βελτίωση των αποτελεσμάτων που λαμβάνονται. Επιπρόσθετα, τα αποτελέσματα που παράγονται από την μέθοδο της γενικευμένης ετεροσυσχέτισης καθώς και με την εφαρμογή της βελτιστοποίησης αποκοπής απομακρυσμένων ευθειών μπορούν να εκτιμήσουν σε ικανοποιητικό βαθμό την θέση της πηγής.

Αξίζει να τονισθεί πως η ανάπτυξη ενός συστήματος πραγματικού χρόνου είναι μία μη-τετριμμένη υπόθεση. Η πολυπλοκότητα που προκαλεί η διαπέραση του πλέγματος για τον εντοπισμό του βέλτιστου σημείου που αναπαριστά την ηχητική πηγή καθώς και η επεξεργασία των δεδομένων στο πεδίο της συχνότητας είναι ακριβή και δαπανηρή σε χρόνο διαδικασία.

5.2 Μελλοντικές κατευθύνσεις έρευνας

Στην έκταση της εργασίας μελετήθηκε μόνο η περίπτωση του εντοπισμού της πηγής με έναν μοναδικό ομιλητή (μοναδική ηχητική πηγή). Είναι πιθανό να επεκταθούν οι δύο βασικές τεχνικές που χρησιμοποιήθηκαν για να εφαρμοστούν στην περίπτωση των πολλαπλών ομιλητών. Αυτό μπορεί να επιτευχθεί διαχωρίζοντας τον χώρο που βρίσκεται υπό μελέτη σε έναν ορισμένο αριθμό υποσυνόλων στους οποίους στην συνέχεια μπορούν να εφαρμοστούν οι αλγόριθμοι γενικευμένης ετεροσυσχέτισης και κατευθυνόμενης δύναμης απόκρισης.

Επίσης ενδιαφέρον παρουσιάζει η έρευνα διαφορετικών συναρτήσεων βάρους (weighting functions) σε σχέση με τον μετασχηματισμό φάσης (PHat) που χρησιμοποιήθηκε στα πειράματα. Ο Brandstein [5] στην δουλεία του δείχνει πως χρησιμοποιώντας μία συνάρτηση βάρους βασισμένη στον τόνο (pitch) και εφαρμόζοντας την στην μέθοδο γενικευμένης ετεροσυσχέτισης (GCC) μπορεί να αποδώσει καλύτερα σε σχέση με τον μετασχηματισμό φάσης σε θορυβώδεις συνθήκες ενώ προσφέρει παρόμοια απόδοση σε χώρους με συνθήκες αντήχησης. Επιπλέον, θα είχε ιδιαίτερο ενδιαφέρον η μελέτη της εφαρμογής της συνάρτησης βάρους βασισμένη στον τόνο στην τεχνική κατευθυνόμενης δύναμης απόκρισης και στο αν μπορεί να βελτιώσει την απόδοση και την ευρωστία της.

Bibliography

[1] *Acoustic source localization using SRP-PHAT*. URL: http://www.mathworks.com/matlabcentral/fileexchange/24352-acoustic-source-localization-using-srp-phat.

[2] X. Anguera. "Robust Speaker Diarization for Meetings". PhD thesis. UPC Barcelona, 2006.

[3] *BeamformIt, the fast and robust acoustic beamformer*. URL: http://www.xavieranguera.com/beamformit.

[4] J. Benesty. "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization". In: *J. Acoust. Soc. Am.* 107 (2000), pp. 384–391.

[5] M. S. Brandstein. "Time-delay estimation of reverberated speech exploiting harmonic stracture". In: *J. Acoust. Soc. Amer.* (1999).

[6] M. S. Brandstein, J. Adcock, and H. Silverman. "A practical time-delay estimator for localizing speech sources with a microphone array". In: *Computer, Speech, and Language* (1995), pp. 153–169.

[7] M. S. Brandstein and H. F. Silverman. "A robust method for speech signal time-delay estimation in reverberant rooms". In: *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.* (1997).

[8] M.S. Brandstein. "A pitch-based approach to time-delay estimation of reverberant speech". In: *Proc. IEEE ASSP Workshop Appls, Signal Processing Audio Acoustics* (1997).

[9] B. Champagne, S. Bedard, and A. Stephenne. "Performance of time-delay estimation in the presence of room reverberation". In: *IEEE Trans. Speech Audio Proc.* (1996), pp. 148–152.

[10] J. Chen, J. Benesty, and Y. Huang. "Time delay estimation using spatial correlation techniques". In: *Proc. 8th IEEE International Workshop on Acoustic Echo and Noise Control* (2003).

[11] J. H. DiBiase. "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays". PhD thesis. Brown University, 2000.

[12] H. Do, H.F. Silverman, and Y. Yu. "A real-time SRP-PHAT source location implementation using stochastic region contraction on a large-aperture microphone array". In: *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.* (2007), pp. 121–124.

[13] Hoang Tran Huy Do. "Real-time SRP-PHAT source location implementations on a large-aperture microphone array". MA thesis. Brown university, 2009.

[14] A. Farina. “Simulataneus Measurements of Impulse Response and Distortion with a Swept-Sine Technique”. In: *Proc. AES Convention* (2000).

[15] J. Huang, M. Epstein, and M. Matassoni. “Effective Acoustic Adaptation for a Distant-talking Interactive TV System”. In: *Proc. Interspeech* (2008).

[16] Y. Huang, J. Benesty, and G. W. Elko. “Microphone arrays for video camera steering”. In: *Acoustic Signal Processing for Telecommunication* (2000), pp. 239–259.

[17] C. H. Knapp and G. C. Carter. “The generalized correlation method for estimation of time delay”. In: *IEEE Trans. Acoust., Speech, Signal Processing* 24 (1976), pp. 320–327.

[18] M. Matassoni et al. “HMM Training with Contaminated Speech Material for Distant-Talking Speech Recognition”. In: *Computer Speech and Language* (2002).

[19] M. Morf, J. Delosme, and B. Friedlander. “A linear equation approach to locating sources from time-difference-of-arrival measurments”. In: *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.* (1980), pp. 818–824.

[20] M. Omologo and P. Svaizer. “Acoustic event localization using a crosspower-spectrum phase based technique”. In: *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.* (1994), pp. 273–276.

[21] M. Omologo and P. Svaizer. “Acoustic source location in noisy and reverberant environment using CSP analysis”. In: *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.* (1996), pp. 921–924.

[22] M. Ravanelli et al. “The DIRHA simulated corpus”. In: *Proc. LREC* (2014).

[23] H. Schau and A. Robinson. “Passive source localization employing intersecting spherical surfaces from time-of-arrival differences”. In: *IEEE Trans. Acoust., Speech, Signal Process.* (1987), pp. 1223–1225.

[24] H. F. Silverman et al. “Performance of real-time source-location estimators for a large-aperture microphone array”. In: *IEEE Trans. Speech, Audio Process.* (2005), pp. 593–606.

[25] J. Smith and J. Abel. “The spherical interpolation method for closed-form passive source localization using range difference measurments”. In: *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.* (1987).

[26] *SoX, the Swiss Army knife of sound processing programs*. URL: `http://sox.sourceforge.net`.

[27] *Speech detection and speaker localization in domestic environments*. URL: `http://dirha.fbk.eu/hscma`.

[28] A. Stephenne and B. Champagne. “A new cepstral prefiltering technique for time delay estimation under reverberant conditions”. In: *Signal Processing* 59 (1997), pp. 253–266.

[29] *The DIRHA (Distant-speech interaction for robust home applications) project*. URL: `http://dirha.fbk.eu`.

[30] A. Tsiami et al. “Experiments in Acoustic Source Localization Using Sparce Arrays in Adverse Indoors Environments”. In: *Proc. EUSIPCO* (2014).

[31] H. Wang and P. Chu. “Voice source localization for automatic camera pointing system in videoconferencing”. In: *Proc. IEEE ASSP Workshop Apps, Signal Processing Audio Acoustics* (1997).