Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Εντοπισμός θέσης πηγής και αποθορυβοποίηση σημάτων ομιλίας με πολυκαναλική επεξεργασία

Διπλωματική Εργασία

του

Ζήση – Ιάσονα Ε. Σκορδίλη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2013

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Εντοπισμός θέσης πηγής και αποθορυβοποίηση σημάτων ομιλίας με πολυκαναλική επεξεργασία

Διπλωματική Εργασία

του

Ζήση – Ιάσονα Ε. Σκορδίλη

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22η Ιουλίου 2013.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Πέτρος Μαραγκός	Γεράσιμος Ποταμιάνος	Κωνσταντίνος Τζαφέστας
Καθηγητής	Αναπληρωτής Καθηγητής	Επίκουρος Καθηγητής
Ε.Μ.Π.	Παν/μίου Θεσσαλίας	Ε.Μ.Π.

Αθήνα, Ιούλιος 2013

(Υπογραφή)

.....
Ζήσης – Ιάσων Ε. Σκορδίλης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Πρώτα από όλα θα ήθελα να ευχαριστήσω θερμότατα τον καθηγητή κ. Πέτρο Μαραγκό για την επίβλεψη της διπλωματικής μου εργασίας, την ευκαιρία που μου έδωσε να εργαστώ στο εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων και τις πολύτιμες συμβουλές, την καθοδήγηση και την ηθική συμπαράστασή του στις ερευνητικές μου προσπάθειες. Θα ήθελα, επίσης, να ευχαριστήσω θερμά όλα τα μέλη του εργαστηρίου για την πολύτιμη βοήθειά τους και ιδιαίτερα τους υποψήφιους διδάκτορες Ισίδωρο Ροδομαγουλάκη, Παναγιώτη Γιαννούλη και Αθανασία Ζλατίντση. Για την πολύτιμη συνεισφορά τους στην ηχογράφηση πολυκαναλικών δεδομένων, θα ήθελα να ευχαριστήσω θερμά τους υποψήφιους διδάκτορες Ισίδωρο Ροδομαγουλάκη και Παναγιώτη Γιαννούλη, καθώς και όλους τους εθελοντές–ομιλητές που συνεισέφεραν για τη συλλογή των δεδομένων αυτών. Θα ήθελα να ευχαριστήσω το Σταμάτη Λευκιμμιάτη για την ευγενική παραχώρηση του πηγαίου κώδικα MATLAB για πολυκαναλική αποθορυβοποίηση σημάτων που είχε υλοποιήσει κατά τη διάρκεια του διδακτορικού του. Ευχαριστώ, επίσης, τους Luca Spelgatti και Roberto Sannino της ST Microelectronics για την παροχή στο εργαστήριο συστοιχιών με μικρόφωνα MEMS και συμβουλών για τη χρήση τους. Για την ευγενική παραχώρηση πολυκαναλικών βάσεων δεδομένων για εντοπισμό θέσης ομιλητή θα ήθελα να ευχαριστήσω τους Piergorgio Svaizer και Alessio Brutti του Ινστιτούτου Fondazione Bruno Kessler. Θα ήθελα, τέλος, να ευχαριστήσω θερμά την οικογένεια μου για τη βοήθεια, τη συμπαράσταση, την ηθική υποστήριξη, τις συμβουλές και την καθοδήγηση που μου προσέφεραν και συνεχίζουν να μου προσφέρουν.

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως αντικείμενο τον εντοπισμό θέσης πηγής ακουστικού σήματος και την αποθορυβοποίηση ομιλίας με πολυκαναλική επεξεργασία. Για το πρόβλημα του εντοπισμού θέσης πηγής, προτείνεται μία νέα μέθοδος ελαχίστων τετραγώνων για τη βέλτιστη εκτίμηση της θέσης πηγής από τις κατευθύνσεις άφιξης (Direction of Arrival, DOA) του σήματος πηγής σε ζεύγη μικροφώνων. Η κατεύθυνση άφιξης υπολογίζεται εκτιμώντας τη διαφορά χρόνου άφιξης (Time Difference of Arrival, TDOA) του σήματος πηγής στο εκάστοτε ζεύγος μικροφώνων. Για την εκτίμηση TDOA χρησιμοποιείται η μέθοδος μετρικού συνοχής φασης ετεροφάσματος (Crosspower-spectrum Phase Coherence Measure, CSP-CM) με βελτιώσεις στην ακρίβεια και την υπολογιστική πολυπλοκότητά της. Η προτεινόμενη μέθοδος εντοπισμού θέσης με ελάχιστα τετράγωνα καταλήγει σε εκτίμηση κλειστής μορφής για τη θέση της πηγής, συνεπώς είναι υπολογιστικά αποδοτική και κατάλληλη για εφαρμογές πραγματικού χρόνου. Για το πρόβλημα της πολυκαναλικής αποθορυβοποίησης σημάτων, γίνεται μελέτη της επίδρασης της γεωμετρίας της συστοιχίας μικροφώνων στην αποτελεσματικότητα πολυκαναλικού συστήματος αποθορυβοποίησης με MVDR beamforming και post-filtering. Για το σκοπό αυτό, έγινε συλλογή βάσης δεδομένων με πολυκαναλικές ηχογραφήσεις σε πραγματικές συνθήκες για εξαγωνικές και γραμμικές διατάξεις της συστοιχίας μικροφώνων σε διάχυτο (diffuse) και εντοπισμένο (localized) θόρυβο. Για τη συλλογή της βάσης, χρησιμοποιήθηκε συστοιχία με μικρόφωνα MEMS, τα οποία είναι μία νέα τεχνολογία φορητών μικροφώνων πολύ μικρών διαστάσεων. Πέραν της πειραματικής αυτής μελέτης, προτείνεται μία θεωρητική βελτίωση στη μέθοδο εκτίμησης των παραμέτρων του post-filter για το χρησιμοποιηθέν σύστημα πολυκαναλικής αποθορυβοποίησης, όμως αποδεικνύεται ότι στην πράξη αυτή δε βελτιώνει την έξοδο του post-filter.

Λέξεις Κλειδιά

εντοπισμός θέσης πηγής, εκτίμηση διαφοράς χρόνου άφιξης, εκτίμηση κατεύθυνσης άφιξης, πολυκαναλική αποθορυβοποίηση ομιλίας, beamforming, post-filtering, πολυκαναλική επεξεργασία σημάτων, συστοιχίες μικροφώνων, μικρόφωνα MEMS (MicroElectroMechanical Systems)

Abstract

This thesis focuses on the problems of source localization and speech enhancement through multichannel signal processing. For the source localization problem, a novel least-squares (LS) method for estimating the source location from the Direction of Arrival (DOA) of the source signal to microphone pairs is proposed. To calculate the DOA, the Time Difference of Arrival (TDOA) of the source signal to the respective microphone pair is first estimated. The TDOA estimation is carried out using the Crosspower-Spectrum Phase Coherence Measure (CSP-CM) with some improvements to its computational efficiency and its accuracy. The proposed LS source localization method yields a closed-form source location estimator and is therefore efficient and suitable for real-time applications. For the multichannel speech enhancement problem, the effect of the microphone array geometry on the efficacy of a multichannel speech enhancement system with MVDR beamforming and post-filtering is studied. To this end, a multichannel database was formed by collecting real recorded data for hexagonal and linear arrangements of the microphone array in diffuse and localized noise fields. For the data collection, a microphone array consisting of MEMS (MicroElectroMechanical Systems) microphones, which are a newly developed technology of highly compact sensors, was used. Besides this experimental study, a theoretical improvement on the post-filter parameter estimation method of the multichannel speech enhancement system employed is proposed, however it is shown that in practice this does not improve the post-filter output.

Keywords

source localization, time difference of arrival estimation, direction of arrival estimation, multichannel speech enhancement, beamforming, post-filtering, multichannel signal processing, microphone arrays, MEMS (MicroElectroMechanical Systems) microphones

Περιεχόμενα

Κατάλογος Σχημάτων

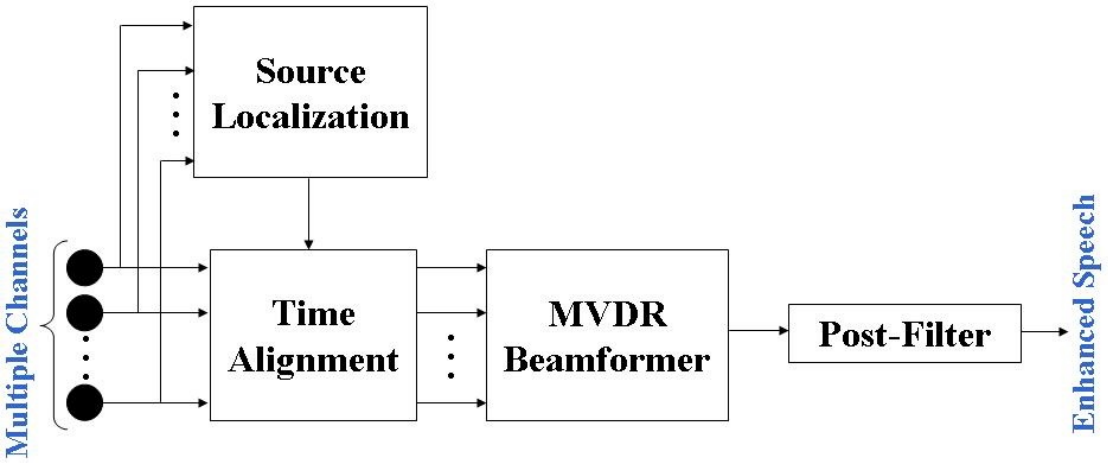
Κατάλογος Πινάκων

Κεφάλαιο 1

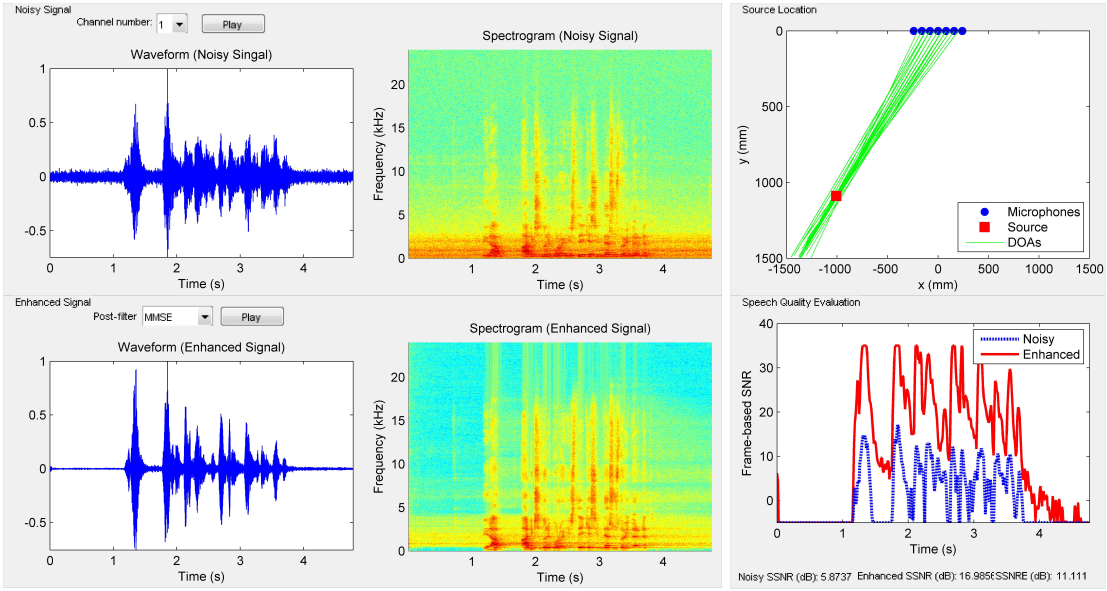
ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια σημαντική προσπάθεια έχει αφιερωθεί στην ανάπτυξη αυτοματοποιημένων συστημάτων 'διάχυτης' νοημοσύνης (ambient intelligence), τα οποία με τη χρήση διαφόρων ειδών αισθητήρων εξάγουν πληροφορία για την κατάσταση του περιβάλλοντός τους και δέχονται ερεθίσματα και εντολές. Μία δημοφιλής τέτοια εφαρμογή είναι τα έξυπνα σπίτια (Smart Homes) [9], τα οποία παρέχουν αυτοματοποιημένες υπηρεσίες ελέγχου του σπιτιού. Είναι επιθυμητό η αλληλεπίδραση του χρήστη με τέτοια αυτοματοποιημένα συστήματα να γίνεται με φυσικό τρόπο και διαφανώς ως προς το χρήστη. Ένας από τους κύριους τρόπου αλληλεπίδρασης είναι η χρήση φυσικής γλώσσας, συνεπώς η επεξεργασία φωνής και η αναγνώριση ομιλίας (Automatic Speech Recognition, ASR) είναι σημαντικό κομμάτι τέτοιων συστημάτων. Για να είναι η αλληλεπίδραση διαφανής ως προς το χρήστη, τα συστήματα αυτά χρησιμοποιούν για την καταγραφή ηχητικών σημάτων μικρόφωνα κατανεμημένα στο χώρο και όχι μικρόφωνα που ο χρήστης πρέπει να έχει πάνω του. Η μεγάλη απόσταση μικροφώνων και ομιλητή οδηγεί σε καταγραφή σημάτων ομιλίας τα οποία είναι παραμορφωμένα από φαινόμενα διάδοσης του ηχητικού σήματος στο χώρο (όπως εξασθένση και αντήχηση) και τα οποία περιέχουν σημαντική συνιστώσα θορύβου από το περιβάλλον. Η υποβάθμιση αυτή του σήματος οδηγεί σε χαμηλές επιδόσεις στην αυτόματη αναγνώριση ομιλίας (ASR). Για την αντιμετώπιση της υποβάθμισης αυτής του σήματος, χρησιμοποιούνται συστοιχίες μικροφώνων και τεχνικές πολυκαναλικής επεξεργασίας σημάτων.

Το πλεονέκτημα που παρουσιάζει η καταγραφή ακουστικών σημάτων από πολλαπλές θέσεις στο χώρο είναι η δυνατότητα αξιοποίησης χωρικών χαρακτηριστικών του ακουστικού πεδίου ώστε να εξαχθεί πληροφορία για τη χωρική θέση του ομιλητή και να αξιοποιηθούν για αποθορυβοποίηση των σημάτων, παράλληλα με τα φασματικά και χωρικά χαρακτηριστικά των σημάτων επιθυμητής πηγής και θορύβου. Τα χωρικά χαρακτηριστικά των σημάτων αξιοποιούνται για αποθορυβοποίηση μέσω τεχνικών beamforming [41, 40]. Οι τεχνικές αυτές απαιτούν γνώση της θέσης της πηγής του επιθυμητού σήματος, οπότε ο εντοπισμός θέσης πηγής είναι ένα σημαντικό πρόβλημα.



(α) Σύστημα πολυκαναλικής αποθορυβοποίησης σημάτων.



(β) Παράδειγμα λειτουργίας του συστήματος πολυκαναλικής αποθορυβοποίησης.

Σχήμα 1.1: Σύστημα πολυκαναλικής αποθορυβοποίησης σημάτων: (α) Δομή του συστήματος και (β) Παράδειγμα λειτουργίας του συστήματος με συστοιχία μικροφώνων MEMS σε γραμμική διάταξη: σε θορυβώδες περιβάλλον ένας ομιλητής σε απόσταση 1.5m και γωνία 135^o σε σχέση με τη συστοιχία εκφωνεί μία εντολή. Στην άνω σειρά γραφικών παραστάσεων διακρίνονται από αριστερά προς δεξιά: η θορυβώδης κυματομορφή του σήματος που καταγράφεται, το φασματογράφημα αυτής και η θέση στην οποία το σύστημα αυτόματα εντόπισε τον ομιλητή (κόκκινο τετράγωνο) μαζί με τις εκτιμώμενες κατευθύνσεις άφιξης (Directions of Arrival, DOAs) (πράσινες ευθείες). Στην κάτω σειρά, διακρίνονται από αριστερά προς δεξιά: η αποθορυβοποιημένη κυματομορφή του σήματος στην έξοδο του συστήματος, το φασματογράφημα αυτής και ο λόγος ισχύος σήματος προς ισχύ θορύβου (Signal to Noise Ratio) ανά πλαίσιο για τα θορυβώδη και αποθορυβοποιημένα σήματα.

Ένα παράδειγμα συστήματος πολυκαναλικής αποθορυβοποίησης με beamforming και post-filtering απεικονίζεται στο Σχήμα 1.1(α). Στο Σχήμα 1.1(β) παρουσιάζεται ένα παράδειγμα λειτουργίας του συστήματος χρησιμοποιώντας μία γραμμική συστοιχία από μικρόφωνα MEMS, τα οποία είναι μία νέα τεχνολογία μικροφώνων πολύ μικρών διαστάσεων.

Η παρούσα εργασία εστιάζει στα προβλήματα του εντοπισμού θέσης πηγής και της πολυκαναλικής αποθορυβοποίησης σημάτων.

Το πρόβλημα του εντοπισμού θέσης πηγής συνίσταται στην εκτίμηση της θέσης πηγής από τα καταγεγραμμένα από μία συστοιχία μικροφώνων σήματα. Η εκτίμηση της θέσης πηγής πρέπει να είναι ακριβής και υπολογιστικά αποδοτική, καθώς η εκτίμηση της θέσης του ομιλητή είναι μόνο το πρώτο στάδιο σε πολυκαναλικούς αλγορίθμους αποθορυβοποίησης. Για την εκτίμηση της θέσης πηγής προτείνεται στην παρούσα εργασία μία νέα μέθοδος, η οποία χρησιμοποιεί ελάχιστα τετράγωνα για την εξαγωγή μίας κλειστής μορφής εκτίμησης της θέσης πηγής. Η κλειστής μορφής εκτίμηση καθιστά τη μέθοδο υπολογιστικά αποδοτικότερη από υπάρχουσες τεχνικές που χρησιμοποιούν αλγορίθμους αναζήτησης [12, 8] για τον εντοπισμό της θέσης πηγής. Αποδεικνύεται πειραματικά οτι η νέα μέθοδος είναι ακριβής και αρκετά αποδοτική ώστε να ειναι κατάλληλη για εφαρμογές πραγματικού χρόνου.

Το πρόβλημα της πολυκαναλικής αποθορυβοποίησης σημάτων συνίσταται στη βέλτιστη εκτίμηση ενός επιθυμητού σήματος από πολυκαναλική καταγραφή θορυβωδών σημάτων. Συχνά χρησιμοποιούνται στην πράξη συστήματα πολυκαναλικής αποθορυβοποίησης με beamforming και post-filtering [25, 26]. Ένα ενδιαφέρον θέμα είναι η επίδραση της επιλεχθείσας γεωμετρίας της συστοιχίας μικροφώνων στην αποτελεσματικότητα beamforming αλγορίθμων. Στην παρούσα εργασία γίνεται μελέτη της αποτελεσματικότητας ενός πολυκαναλικού συστήματος αποθορυβοποίησης σε σχέση με τη γεωμετρία της συστοιχίας που χρησιμοποιείται και τις συνθήκες θορύβου που επικρατούν στο περιβάλλον. Για το σκοπό αυτό έγινε συλλογή μίας βάσης δεδομένων με μία συστοιχία από μικρόφωνα MEMS, τα οποία είναι μία νέα τεχνολογία μικροφώνων πολύ μικρών διαστάσεων. Έγιναν πολυκαναλικές ηχογραφήσεις για γραμμικές και εξαγωνικές διατάξεις της συστοιχίας και για διάφορα είδη θορύβων. Αποδείχτηκε πειραματικά ότι η εξαγωνική γεωμετρία υπερτερεί της γραμμικής. Πέραν της πειραματικής αυτής μελέτης, προτείνεται μία θεωρητική βελτίωση στη μέθοδο εκτίμησης των παραμέτρων του post-filter για το χρησιμοποιηθέν σύστημα πολυκαναλικής αποθορυβοποίησης.

Στο Κεφάλαιο 2, το οποίο είναι αφιερωμένο στο πρόβλημα του εντοπισμού θέσης πηγής, γίνεται μία επισκόπηση υπαρχουσών μεθόδων εντοπισμού θέσης πηγής και παρουσιάζεται η προτεινόμενη μέθοδος ελαχίστων τετραγώνων, η αποτελεσματικότητα της οποίας αποδεικνύεται πειραματικά.

Στο Κεφάλαιο 3, το οποίο είναι αφιερωμένο στο πρόβλημα της πολυκαναλικής αποθορυβοποίησης σημάτων, παρουσιάζεται το πολυκαναλικό σύστημα αποθορυβο-

ποίησης που χρησιμοποιήθηκε, η δομή της βάσης δεδομένων MEMS και τα αποτελέσματα των πειραμάτων για διάφορους συνδυασμούς ειδών θορύβου και γεωμετριών της συστοιχίας, τα οποία δείχνουν την υπεροχή της εξαγωνικής γεωμετρίας. Πέραν της πειραματικής αυτής μελέτης, προτείνεται μία θεωρητική βελτίωση στη μέθοδο εκτίμησης των παραμέτρων του post-filter για το χρησιμοποιηθέν σύστημα πολυκαναλικής αποθορυβοποίησης.

Κεφάλαιο 2

ΕΝΤΟΠΙΣΜΟΣ ΘΕΣΗΣ ΠΗΓΗΣ

2.1 Διατύπωση του προβλήματος

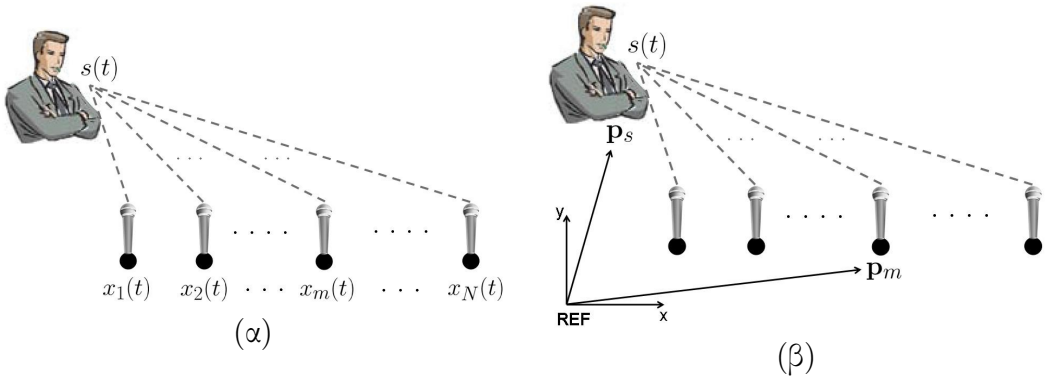
Έστω μία συστοιχία N μικροφώνων κατανεμημένων στο χώρο και μία πηγή ακουστικού σήματος, η οποία μπορεί να μοντελοποιηθεί ως σημειακή, όπως ένας ομιλητής (Σχήμα 2.1(α)). Έστω $s(t)$, όπου t ο συνεχής χρόνος, το σήμα που παράγεται από την πηγή.

Τα σήματα που καταγράφονται από τα μικρόφωνα μπορούν να μοντελοποιηθούν ως εξής:

$$x_m(t) = a_m s(t - \tau_m) + v_m(t),\ m = 1,\ 2,\ \ldots,\ N \tag{2.1}$$

όπου t ο συνεχής χρόνος, $x_m(t)$ είναι το σήμα που καταγράφεται από το μικρόφωνο m, a_m, τ_m είναι η εξασθένηση και η καθυστέρηση διάδοσης, αντίστοιχα, τις οποίες υφίσταται το ηχητικό κύμα το οποίο παράγει η πηγή μέχρι την άφιξή του στο μικρόφωνο m και $v_m(t)$ είναι η συνιστώσα θορύβου η οποία καταγράφεται στο μικρόφωνο m. Σε περίπτωση κατά την οποία στο περιβάλλον διάδοσης υπάρχει αντήχηση (reverberation), η συνιστώσα θορύβου $v_m(t)$, είναι συσχετισμένη με το σήμα πηγής $s(t)$, καθώς περιέχει και ανακλάσεις αυτού.

Το πρόβλημα εύρεσης θέσης πηγής έγκειται στην εκτίμηση της θέσης της πηγής $\mathbf{p}_s$, δεδομένης της γεωμετρίας της συστοιχίας μικροφώνων, δηλαδή των θέσεων $\mathbf{p}_m,\ m = 1,\ 2,\ \ldots,\ N$ (Σχήμα 2.1 (β)). Η εκτίμηση της θέσης της πηγής γίνεται αξιοποιώντας τη χωρική πληροφορία την οποία προσφέρει η καταγραφή του ακουστικού σήματος της πηγής σημάτων από πολλές θέσεις $\mathbf{p}_m$ στο χώρο. Η χωρική πληροφορία κωδικοποιείται στα σήματα $x_m(t)$ κυρίως μέσω των καθυστερήσεων διάδοσης τ_m.



Σχήμα 2.1: Μοντελοποίηση του προβλήματος εντοπισμού θέσης: (α) Καταγραφή σήματος ακουστικής πηγής από συστοιχία μικροφώνων και (β) Σύστημα αναφοράς και διανύσματα θέσης πηγής και μικροφώνων.

2.2 Υπάρχουσες Μέθοδοι

Οι υπάρχουσες μέθοδοι εντοπισμού θέσης πηγής μπορούν να χωριστούν σε τρεις κατηγορίες [13]: μέθοδοι που βασίζονται στην ισχύ της κατευθυνόμενης απόκρισης (Steered Response Power, SRP), μέθοδοι που βασίζονται σε υψηλής ανάλυσης φασματική εκτίμηση (High Resolution Spectral Estimation, HRSE) και μέθοδοι που βασίζονται σε εκτίμηση διαφοράς χρόνου άφιξης (Time Difference of Arrival, TDOA). Ακολουθεί μία γενική επισκόπηση της δομής και των πλεονεκτημάτων και μειονεκτημάτων των μεθόδων κάθε κατηγορίας βασισμένη στο [13].

Οι μέθοδοι SRP πραγματοποιούν beamforming κατευθυνόμενο (steered) σε διάφορες θέσεις στο χώρο και αναζητούν τη θέση που μεγιστοποιεί την ισχύ στην έξοδο του beamformer. Το πλεονέκτημα των μεθόδων αυτών είναι η αυξημένη ευρωστία (robustness). Ένα μειονέκτημα των μεθόδων αυτών, εν γένει, είναι το γεγονός ότι η έξοδος του beamformer εξαρτάται από τα φασματικά χαρακηριστικά του σήματος που παράγει η πηγή. Για την αντιμετώπιση αυτού του μειονεκτήματος προτάθηκε η μέθοδος SRP με μετασχηματισμό φάσης (SRP-PHAT) [12], η οποία χρησιμοποιείται συχνά στην πράξη. Το κύριο μειονέκτημα των μεθόδων SRP (συμπεριλαμβανομένης της SRP-PHAT) είναι το αυξημένο υπολογιστικό κόστος, το οποίο επιβάλλει η αναζήτηση που πραγματοποιείται στο χώρο για την εύρεση της θέσης πηγής.

Οι μέθοδοι της κατηγορίας HRSE χρησιμοποιούν τεχνικές από το επιστημονικό πεδίο της φασματικής ανάλυσης, όπως autoregressive μοντελοποίηση (AR modeling), minimum variance εκτίμηση φάσματος και τεχνικές βασισμένες σε ανάλυση με ιδιοδιανύσματα (όπως για παράδειγμα ο αλγόριθμος MUSIC [34]). Στην πράξη αυτές οι μέθοδοι τείνουν να είναι λιγότερο εύρωστες από τις SRP μεθόδους, λόγω ιδανικών υποθέσεων που γίνονται κατά την εξαγωγή τους και οι οποίες δεν ισχύουν με ακρίβεια σε πραγματικές συνθήκες. Επίσης, οι μέθοδοι αυτές, όπως και οι μέθοδοι

SRP, καταλήγουν σε αλγορίθμους αναζήτησης με συνέπεια το αυξημένο υπολογιστικό κόστος.

Οι μέθοδοι που βασίζονται σε εκτίμηση TDOA αποτελούνται από δύο στάδια. Στο πρώτο στάδιο, γίνεται εκτίμηση της διαφοράς χρόνου άφιξης (TDOA) $\Delta\tau_{ij} = \tau_i - \tau_j$ του σήματος πηγής σε ζεύγη μικροφώνων. Στο δεύτερο στάδιο, οι εκτιμήσεις TDOA συνδυάζονται με τη γνώση των θέσεων των μικροφώνων προκειμένου να εξαχθεί η κατεύθυνση άφιξης του σήματος πηγής σε κάθε ζεύγος και τελικά η θέση της πηγής να εξαχθεί χρησιμοποιώντας γεωμετρία. Πολλές από τις μεθόδους αυτής της κατηγορίας καταλήγουν σε λύση κλειστής μορφής (closed-form) για την εκτίμηση της θέσης πηγής. Παραδείγματα τέτοιων μεθόδων είναι η γραμμική τομή (linear intersection) [6], η σφαιρική παρεμβολή (spherical interpolation) [36], η σφαιρική τομή (spherical intersection) [33] και άλλες [5, 10, 22, 20]. Λόγω της λύσης κλειστής μορφής, το κύριο πλεονέκτημα μεθόδων είναι το μειωμένο υπολογιστικό κόστος. Το μειονέκτημα τους έγκειται στο γεγονός ότι η διαδικασία δύο σταδίων με εκτίμηση των TDOAs ανεξάρτητα από την εκτίμηση της θέσης πηγής είναι υποβέλτιστη. Ωστόσο, σε πολλές περιπτώσεις στην πράξη, η μείωση στην επίδοση είναι αποδεκτή σε σύγκριση με τις μεθόδους που χρησιμοποιούν υπολογιστικά ακριβούς αλγορίθμους αναζήτησης [13].

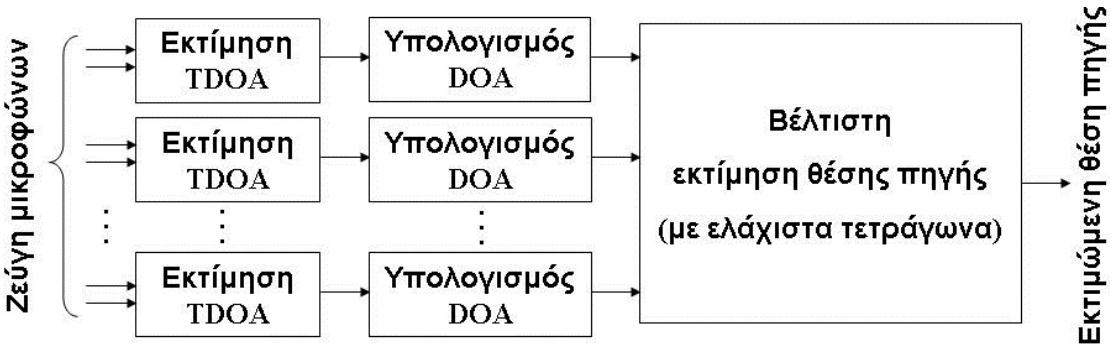
2.3 Προτεινόμενη Μέθοδος

Είναι επιθυμητό ο εντοπισμός της θέσης πηγής να είναι ακριβής, αλλά και ιδιαίτερα αποδοτικός υπολογιστικά ώστε να είναι υλοποιήσιμος σε πραγματικό χρόνο. Η ανάγκη για πολύ μικρό υπολογιστικό κόστος εντείνεται, όταν ο εντοπισμός θέσης χρησιμοποιείται ως το πρώτο στάδιο (front-end) σε πολυκαναλικούς αλγορίθμους αποθορυβοποίησης με beamforming (οι οποίοι απαιτούν την πληροφορία για τη θέση της πηγής), καθώς είναι επιθυμητό το συνολικό σύστημα να είναι υλοποιήσιμο σε πραγματικό χρόνο.

Με βάση τις παρατηρήσεις αυτές, προτείνεται μία νέα μέθοδος εντοπισμού θέσης, η οποία ανήκει στην κατηγορία μεθόδων βασισμένων σε εκτίμηση TDOA και η οποία χρησιμοποιώντας προσέγγιση με ελάχιστα τετράγωνα καταλήγει σε τύπο κλειστής μορφής για την εκτίμηση της θέσης, συνεπώς έχει μικρό υπολογιστικό κόστος, αντίθετα με διαδεδομένες μεθόδους που χρησιμοποιούν αλγορίθμους αναζήτησης [12, 8].

Το προτεινόμενο σύστημα εντοπισμού θέσης, το οποίο αποτελείται από τρία στάδια, παρουσιάζεται στο Σχήμα 2.2.

Στο πρώτο στάδιο, από τα σήματα που καταγράφονται στα μικρόφωνα, γίνεται εκτίμηση TDOA για ζεύγη μικροφώνων. Για την εκτίμηση αυτή χρησιμοποιείται μία βελτιωμένη έκδοση της μεθόδου μετρικού συσχέτισης φάσης ετεροφάσματος (Crosspower-Spectrum Phase Coherence Measure, CSP-CM). Η βασική μέθοδος CSP-CM παρουσιάζεται στο [29].



Σχήμα 2.2: Προτεινόμενη μέθοδος εντοπισμού θέσης πηγής.

Στο δεύτερο στάδιο, από τις εκτιμήσεις TDOA υπολογίζεται η κατεύθυνση άφιξης (Direction of Arrival, DOA) του σήματος πηγής σε κάθε ζεύγος μικροφώνων, η οποία, υποθέτοντας ότι η πηγή βρίσκεται στο μακρινό πεδίο (far-field) μοντελοποιείται από μία ευθεία γραμμή [5, 3]. Οι προκύπτουσες ευθείες, λόγω σφαλμάτων στις εκτιμήσεις TDOA, δεν τέμνονται σε ένα κοινό σημείο, συνεπώς η εκτίμηση της θέσης της πηγής πρέπει να γίνει ελαχιστοποιώντας κάποιο κριτήριο σφάλματος.

Στο τρίτο στάδιο, συνδυάζοντας βέλτιστα με προσέγγιση ελαχίστων τετραγώνων που οδηγεί σε κλειστό τύπο τις ευθείες DOA, γίνεται η εκτίμηση της θέσης της πηγής. Το κριτήριο σφάλματος ως προς το οποίο γίνεται η ελαχιστοποίηση είναι το άθροισμα των τετραγώνων των αποστάσεων από τις ευθείες, με αποτέλεσμα η πηγή να εκτιμάται ως το σημείο με το ελάχιστο τέτοιο άθροισμα. Διαισθητικά, αυτό το σημείο το οποίο βρίσκεται 'εγγύτερα' προς όλες τις ευθείες.

Το κύριο πλεονέκτημα της προτεινόμενης μεθόδου, έναντι ευρέως χρησιμοποιούμενων στην πράξη αλγορίθμων αναζήτησης [12, 8], είναι το μικρό υπολογιστικό κόστος που την καθιστά κατάλληλη για εφαρμογές πραγματικού χρόνου. Ταυτόχρονα, η ακρίβεια της μεθόδου είναι ικανοποιητική, όπως θα δειχθεί με πειραματισμό σε πραγματικά πολυκαναλικά σήματα.

Η προτεινόμενη μέθοδος διαφέρει από τις άλλες μεθόδους εντοπισμού θέσης κλειστής μορφής που υπάρχουν στη βιβλιογραφία [6, 36, 33, 5, 10, 22, 20] στη γεωμετρική μοντελοποίηση του προβλήματος, καθώς και στο κριτήριο ελαχιστοποίησης που χρησιμοποιείται, όπως θα εξηγηθεί στις Ενότητες 2.5 και 2.6.

Στην Ενότητα 2.4 παρουσιάζεται η βελτιωμένη μέθοδος CSP-CM για εκτίμηση TDOA που είναι το πρώτο στάδιο της προτεινόμενης μεθόδου εντοπισμού θέσης. Στην Ενότητα 2.5 παρουσιάζεται ο τρόπος υπολογισμού κατεύθυνσης άφιξης από τις εκτιμήσεις DOA που είναι το δεύτερο στάδιο της μεθόδου. Στην Ενότητα 2.6 εξάγεται η κλειστής μορφής εκτίμηση της θέσης πηγής από τις ευθείες κατεύθυνσης άφιξης του σήματος πηγής με βάση το κριτήριο ελαχιστοποίησης που προαναφέρθηκε. Στην Ενότητα 2.7 παρουσιάζεται μία πρακτική λύση για την ανίχνευση των

χρονικών διαστημάτων στα οποία η πηγή είναι ενεργή, ώστε να γίνεται εκτίμηση της θέσης της μόνο σε ενεργές περιόδους. Στην Ενότητα 2.8 παρουσιάζονται πειραματικά αποτελέσματα από την εφαρμογή της προτεινόμενης μεθόδου εντοπισμού θέσης σε πολυκαναλικά σήματα ηχογραφημένα σε πραγματικές συνθήκες και τέλος στην Ενότητα 2.9 παρουσιάζονται τα συμπεράσματα από την εφαρμογή της μεθόδου.

2.4 Εκτίμηση Διαφοράς Χρόνου Άφιξης

Το πρώτο στάδιο της προτεινόμενης μεθόδου εντοπισμού θέσης πηγής είναι η εκτίμηση της διαφοράς χρόνου άφιξης (Time Difference of Arrival, TDOA) του σήματος πηγής στα δύο μικρόφωνα ενός ζεύγους μικροφώνων. Η πληροφορία αυτή μπορεί να αξιοποιηθεί για την εύρεση της κατεύθυνσης άφιξης (Direction of Arrival, DOA) του σήματος πηγής στο ζεύγος μικροφώνων. Η εκτίμηση TDOA πρέπει να είναι ακριβής και πολύ αποδοτική υπολογιστικά, καθώς είναι μόνο το πρώτο στάδιο της προτεινόμενης μεθόδου εντοπισμού θέσης πηγής. Στην Ενότητα 2.4.1 διατυπώνεται μαθηματικά το πρόβλημα εκτίμησης TDOA, στην Ενότητα 2.4.2 παρουσιάζεται μία γνωστή κοινά χρησιμοποιούμενη μέθοδος εκτίμησης TDOA και στις Ενότητα 2.4.3 προτείνονται τρόποι μείωσης της υπολογιστικής πολυπλοκότητας και βελτίωσης της ακρίβειάς της.

2.4.1 Διατύπωση του προβλήματος

Έστω ένα ζεύγος μικροφώνων, το οποίο αποτελείται από τα μικρόφωνα $m_1 = i$, $m_2 = j$. Τα σήματα που καταγράφονται από τα μικρόφωνα αυτά είναι (εξίσωση 2.1):

$$x_i(t) = a_i s(t - \tau_i) + v_i(t), \tag{2.2}$$

$$x_j(t) = a_j s(t - \tau_j) + v_j(t), \tag{2.3}$$

Το πρόβλημα εκτίμησης διαφοράς χρόνου άφιξης (Time Difference of Arrival, TDOA) έγκειται στην εύρεση της διαφοράς των χρόνων άφιξης του σήματος πηγής στα 2 μικρόφωνα $\Delta\tau_{ij} = \tau_i - \tau_j$.

Στη βιβλιογραφία έχουν προταθεί πολλές μέθοδοι για την εκτίμηση του TDOA. Μια γενική παρουσίαση όλων των ειδών μεθόδων μπορεί να βρεθεί στο [3]. Από τις πιο διαδεδομένες και συχνά χρησιμοποιούμενες στην πράξη μεθόδους είναι η μέθοδος Crosspower Spectrum Phase Coherence Measure (CSP-CM) [28, 29, 39], η οποία βασίζεται στη μέθοδο μετασχηματισμού φάσης (Generalized Cross Correlation – Phase Transform, GCC-PHAT) [24]. Η μέθοδος CSP-CM ανήκει στην κατηγορία μεθόδων γενικευμένης ετεροσυσχέτισης (Generalized Cross Correlation, GCC) [24]. Οι μέθοδοι αυτές παρουσιάζουν το πλεονέκτημα ότι είναι υπολογιστικά αποδοτικές και κατάλληλες για εφαρμογές πραγματικού χρόνου, ενώ ταυτόχρονα παρουσιάζουν

ικανοποιητική ακρίβεια στην πράξη [3]. Μεταξύ των GCC μεθόδων, η CSP-CM μέθοδος παρουσιάζει, σύμφωνα με τη θεωρία, το πλεονέκτημα ότι δεν εξαρτάται από τα χαρακτηριστικά της κυματομορφής $s(t)$ [3]. Επίσης έχει βρεθεί και πρακτικά ότι είναι η πιο εύρωστη σε θορυβώδη περιβάλλοντα [28].

Στην Ενότητα 2.4.2 παρουσιάζεται η μέθοδος CSP-CM, ενώ στην Ενότητα 2.4.3 προτείνονται βελτιώσεις στην υπολογιστική πολυπλοκότητα και την ακρίβεια τη μεθόδου.

2.4.2 Η μέθοδος CSP-CM

Το σήμα $s(t)$, εφόσον είναι σήμα φωνής, δεν είναι στάσιμο (stationary). Επίσης η πηγή ενδέχεται να κινείται, με αποτέλεσμα να αλλάζει το TDOA. Συνεπώς πρέπει να ακολουθηθεί ανάλυση βραχέος χρόνου. Έστω $X_i(f,t)$, $X_j(f,t)$, όπου f η συνεχής συχνότητα, οι μετασχηματισμοί Fourier βραχέος χρόνου (Short-Time Fourier Transform, STFT) των σημάτων $x_i(t)$, $x_j(t)$, αντίστοιχα. Έστω $G_{ij}(f,t) = E[X_i(f,t)X_j^*(f,t)]$ το crosspower-spectrum των σημάτων των δύο μικροφώνων. Ο μετασχηματισμός GCC-PHAT δίνεται από [24]:

$$R_{ij}(\tau,t) = \int_{-\infty}^{\infty} \frac{G_{ij}(f,t)}{|G_{ij}(f,t)|} e^{j2\pi f\tau} df. \tag{2.4}$$

Στην ιδανική περίπτωση κατά την οποία τα σήματα θορύβου στα μικρόφωνα είναι ασυσχέτιστα με το σήμα πηγής προκύπτει [28, 3]:

$$\frac{G_{ij}(f,t)}{|G_{ij}(f,t)|} = e^{-j2\pi f\Delta\tau_{ij}(t)}. \tag{2.5}$$

Επομένως:

$$R_{ij}(\tau,t) = \delta(\tau - \Delta\tau_{ij}(t)), \tag{2.6}$$

δηλαδή το $R_{ij}(\tau,t)$ είναι ένας κρουστικός παλμός στη θέση του TDOA $\Delta\tau_{ij}$.

Η μέθοδος CSP-CM υπολογίζει το CSP-CM [28]:

$$C_{ij}(\tau,t) = \int_{-\infty}^{\infty} \frac{X_i(f,t)X_j^*(f,t)}{|X_i(f,t)||X_j(f,t)|} e^{j2\pi f\tau} df \tag{2.7}$$

και εκτιμά το TDOA ως:

$$\Delta\tau_{ij}(t) = \arg\max_{\tau} C_{ij}(\tau,t), \tag{2.8}$$

καθώς, σύμφωνα με την εξίσωση (2.6), το CSP-CM $C_{ij}(\tau,t)$ αναμένεται να εμφανίζει κορυφή ολικού μεγίστου στη θέση $\tau = \Delta\tau_{ij}$.

Αν η πηγή παραμένει ακίνητη, τότε το TDOA δεν αλλάζει με το χρόνο και μπορεί να εκτιμηθεί ως εξής [39]:

$$\Delta\tau_{ij} = \arg\max_{\tau} \int_{T} C_{ij}(\tau, t)dt, \tag{2.9}$$

όπου T το χρονικό διάστημα κατά το οποίο η πηγή είναι ενεργή.

Στο διακριτό πεδίο, το CSP-CM μπορεί να υπολογιστεί ως εξής: Έστω $x_i(n)$, $x_j(n)$ τα δειγματοληπτημένα $x_i(t)$, $x_j(t)$, αντίστοιχα. Έστω $X_i(k,\ell)$, $X_j(k,\ell)$ οι STFTs των $x_i(n)$, $x_j(n)$, αντίστοιχα, όπου k ο συχνοτικός δείκτης (frequency bin index) και ℓ ο δείκτης πλαισίου (frame index). Οι STFTs υπολογίζονται με διακριτό μετασχηματισμό Fourier (Discrete Fourier Transform, DFT). Το διακριτό CSP-CM υπολογίζεται ως [39]:

$$C_{ij}(q,\ell) = \text{IDFT}\left[\frac{X_i(k,\ell)X_j^*(k,\ell)}{|X_i(k,\ell)||X_j(k,\ell)|}\right], \tag{2.10}$$

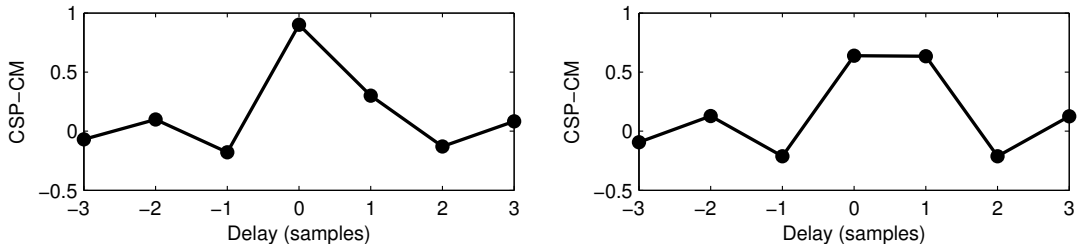
όπου IDFT ο αντίστροφος διακριτός Fourier μετασχηματισμός (Inverse DFT, IDFT) και q ο δείκτης καθυστέρησης (lag index).

2.4.3 Βελτιώσεις στη μέθοδο CSP-CM

Παρεμβολή του CSP-CM

Το διακριτό CSP-CM της εξίσωσης (2.10) είναι περιορισμένο σε ακέραια δείγματα στο πεδίο καθυστέρησης q. Όμως, για χρήση των εκτιμήσεων TDOA για εύρεση της κατεύθυνσης άφιξης και τελικά για εντοπισμό θέσης πηγής απαιτείται ακρίβεια μεγαλύτερη από ακέραια πολλαπλάσια της περιόδου δειγματοληψίας. Για την αντιμετώπιση αυτού του προβλήματος, στο [29] οι συγγραφείς προτείνουν τοπική παρεμβολή interpolation του CSP-CM στην περιοχή μεγίστου. Όμως η παρεμβολή εισάγει μη αμελητέο υπολογιστικό φόρτο και καθυστέρηση στον υπολογισμό. Για να κρατηθεί η υπολογιστική πολυπλοκότητα του αλγορίθμου εκτίμησης TDOA όσο το δυνατόν χαμηλότερα, προτείνεται εδώ μία διαφορετική προσέγγιση για την ακριβή εκτίμηση του TDOA από το διακριτό CSP-CM με αμελητέο υπολογιστικό κόστος.

Στο Σχήμα 2.3 παρουσιάζεται το CSP-CM για ένα πλαίσιο λευκής Γκαουσιανής διαδικασίας (white Gaussian process), στο οποίο εισήχθη καθυστέρηση μη ακέραιου αριθμού δειγμάτων για την παραγωγή του δεύτερου σήματος. Παρουσιάζονται τα CSP-CMs για καθυστέρηση 0.25 και 0.5 δειγμάτων. Διαισθητικά, η τιμή του CSP-CM στο δείκτη καθυστέρησης q δίνει ένα μέτρο συσχέτισης των σημάτων $x_i(n)$, $x_j(n)$ για καθυστέρηση q. Όταν το TDOA δεν ισούται με ακέραιο αριθμό δειγμάτων, τότε το CSP-CM λαμβάνει υψηλή τιμή για δύο γειτονικές τιμές του δείκτη καθυστέρησης q. Η τιμή του σε καθέναν από τους δύο αυτούς δείκτες καθυστέρησης



Σχήμα 2.3: CSP-CM για λευκή Γκαουσιανή διαδικασία για TDOA 0.25 (αριστερά) και 0.5 δείγματα (δεξιά).

δείχνει το 'ποσό συσχέτισης' των σημάτων για τη συγκεκριμένη τιμή καθυστέρησης, επομένως είναι ένα μέτρο εγγύτητας της πραγματικής (κλασματικής) τιμής TDOA στο συγκεκριμένο (ακέραιο) δείκτη καθυστέρησης. Με βάση αυτήν την παρατήρηση προτείνεται ο εξής σταθμισμένος μέσος όρος $\hat{q}_{ij}$ ως εκτίμηση του TDOA (ο δείκτης πλαισίου ℓ παραλείπεται για σαφήνεια):

$$\hat{q}_{ij} = \frac{q_0 C_{ij}(q_0) + q_1 C_{ij}(q_1)}{C_{ij}(q_0) + C_{ij}(q_1)}, \tag{2.11}$$

$$q_0 = \arg\max_q C_{ij}(q), \tag{2.12}$$

$$q_1 = \begin{cases} q_0 + 1 & \text{αν } C_{ij}(q_0+1) > C_{ij}(q_0-1) \\ q_0 - 1 & \text{αλλιώς} \end{cases} \tag{2.13}$$

Για την αξιολόγηση της ακρίβειας της εκτίμησης TDOA που δίνεται από την εξίσωση (2.11), εκτελέστηκαν δύο προσομοιώσεις.

Στην πρώτη προσομοίωση, χρησιμοποιήθηκε ένα πλαίσιο λευκής Γκαουσιανής τυχαίας διαδικασίας μήκους 512 δειγμάτων στο οποίο εισήχθη καθυστέρηση για την παραγωγή του δεύτερου σήματος. Στα δύο σήματα προστέθηκε λευκός Γκαουσιανός θόρυβος (white Gaussian noise) για την επίτευξη συγκεκριμένου λόγου ισχύος σήματος προς ισχύ θορύβου (Signal to Noise Ration, SNR). Χρησιμοποιώντας Hanning παράθυρο και μήκος DFT 1024 δειγμάτων υπολογίστηκε το CSP-CM και έγινε εκτίμηση του TDOA σύμφωνα με την εξίσωση (2.11). Στον Πίνακα 2.1 παρουσιάζεται ο μέσος όρος (mean) και η τυπική απόκλιση standard deviation της εκτίμησης TDOA για 10^4 δοκιμές σε κάθε SNR. Τα αποτελέσματα δείχνουν ότι η εκτίμηση TDOA με την εξίσωση (2.11) είναι ακριβής για όλο το εύρος συνθηκών SNR.

Στη δεύτερη προσομοίωση, χρησιμοποιήθηκε ως αρχικό σήμα ένα φώνημα (/ah/), το οποίο περιέχεται σε ένα πλαίσιο 400 δειγμάτων απομονωμένο από ηχογράφηση στα 16kHz. Τα αποτελέσματα, τα οποία προέκυψαν εκτελώντας την ίδια διαδικασία προσομοίωσης με την πρώτη προσομοίωση, εμφανίζονται στον Πίνακα 2.2. Σε αυτήν την περίπτωση παρατηρείται μία μικρή απόκλιση του μέσου όρου της εκτίμησης από

Πραγματικό TDOA (δείγματα)	SNR (dB)	Εκτίμηση TDOA	
		Mean	St. Dev.
0.25	30	0.25	0.003
	20	0.25	0.007
	10	0.25	0.022
0.50	30	0.50	0.002
	20	0.50	0.007
	10	0.50	0.021
0.75	30	0.75	0.003
	20	0.75	0.007
	10	0.75	0.022

Πίνακας 2.1: Αποτελέσματα εκτίμησης TDOA σε προσομοίωση με λευκή Γκαουσιανή τυχαία διαδικασία σε λευκό Γκαουσιανό θόρυβο: μέσος όρος και τυπική απόκλιση της εκτίμησης TDOA με την προτεινόμενη μέθοδο παρεμβολής του CSP-CM για διάφορες τιμές SNR και καθυστέρησης.

Πραγματικό TDOA (δείγματα)	SNR (dB)	Εκτίμηση TDOA	
		Mean	St. Dev.
0.25	30	0.39	0.038
	20	0.41	0.165
	10	0.26	0.492
0.50	30	0.50	0.037
	20	0.50	0.058
	10	0.50	0.463
0.75	30	0.61	0.039
	20	0.59	0.179
	10	0.73	0.493

Πίνακας 2.2: Αποτελέσματα εκτίμησης TDOA σε προσομοίωση με φώνημα σε λευκό Γκαουσιανό θόρυβο: μέσος όρος και τυπική απόκλιση της εκτίμησης TDOA με την προτεινόμενη μέθοδο παρεμβολής του CSP-CM για διάφορες τιμές SNR και καθυστέρησης.

την πραγματική τιμή TDOA, ωστόσο αυτή δεν υπερβαίνει τα 0.15 δείγματα, το οποίο είναι αρκετά ικανοποιητικό δεδομένης της απλότητας της προτεινόμενης μεθόδου για εκτίμηση TDOA με κλασματική ακρίβεια.

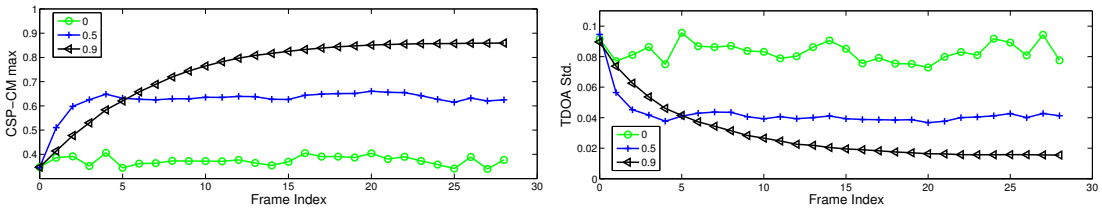
Εκτίμηση του ετεροφάσματος

Κατά τον υπολογισμό του CSP-CM μέσω της εξίσωσης (2.10), το ετεροφάσμα μεταξύ των σημάτων $x_1(n)$ και $x_2(n)$ εκτιμάται ανεξάρτητα για κάθε πλαίσιο ℓ ως $\phi_{ij}(k,\ell) = X_i(k,\ell)X_j^*(k,\ell)$. Η εκτίμηση του ετεροφάσματος μπορεί να βελτιωθεί χρησιμοποιώντας τη μέθοδο βραχέος χρόνου φασματικής εκτίμησης η οποία προτείνεται στο [1]:

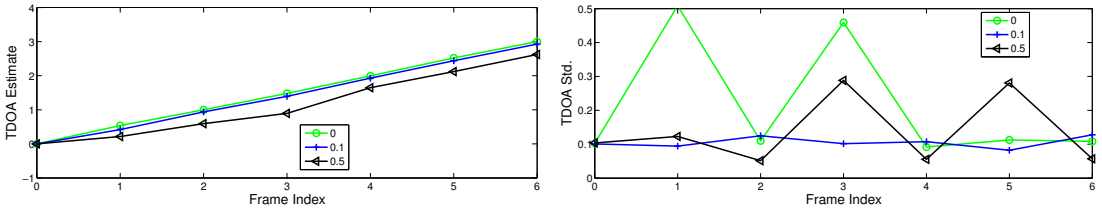
$$\hat{\phi}(k,\ell) = a\hat{\phi}(k,\ell-1) + (1-a)X_i(k,\ell)X_j^*(k,\ell), \tag{2.14}$$

όπου $0 < a < 1$. Αυτή η εκτίμηση του ετεροφάσματος μπορεί να θεωρηθεί ως αναδρομικό Welch περιοδόγραμμα και παράγει ομαλότερη, βελτιωμένη εκτίμηση του ετεροφάσματος. Παρότι αυτή η μέθοδος για εκτίμηση του ετεροφάσματος είναι γνωστή, οι υπάρχουσες εργασίες για τη μέθοδο CSP-CM [28, 29, 39] δεν κάνουν χρήση της.

Για να δειχθεί ότι η εκτίμηση του ετεροφάσματος μέσω της εξίσωσης (2.14) βελτιώνει την εκτίμηση TDOA εκτελέστηκε η εξής προσομοίωση: χρησιμοποιήθηκε μία λευκή Γκαουσιανή διαδικασία, στην οποία εισήχθη καθυστέρηση 0.75 δειγμάτων για παραγωγή ενός δεύτερου σήματος. Στα δύο σήματα προστέθηκε λευκός Γκαουσιανός θόρυβος, ώστε το SNR να γίνει 0dB. Τα θορυβώδη σήματα χωρίστηκαν σε ημιεπικαλυπτόμενα πλαίσια μήκους 512 δειγμάτων στα οποία εφαρμόστηκε παράθυρο Hanning. Για κάθε πλαίσιο, χρησιμοποιώντας Hanning παράθυρο και μήκος DFT 1024 δειγμάτων, υπολογίστηκε το CSP-CM με εκτίμηση του ετεροφάσματος σύμφωνα με την εξίσωση (2.14) και έγινε εκτίμηση του TDOA. Εκτελέστηκαν 100 δοκιμές και υπολογίστηκε η μέση μέγιστη τιμή του CSP-CM και η τυπική απόκλιση της εκτίμησης TDOA για κάθε πλαίσιο. Στο Σχήμα 2.4 παρουσιάζονται τα αποτελέσματα τα οποία προέκυψαν για τιμές $a = 0, 0.5, 0.9$. Είναι εμφανές ότι η χρήση της εξίσωσης (2.14) για εκτίμηση του ετεροφάσματος αυξάνει το ύψος της κορυφής του CSP-CM και μειώνει την τυπική απόκλιση των εκτιμήσεων TDOA με την πάροδο του χρόνου. Επίσης φαίνεται ότι η αύξηση της σταθεράς a βελτιώνει το αποτέλεσμα. Ωστόσο υπάρχει ένα trade-off: αύξηση της σταθεράς a συνεπάγεται μείωση της ικανότητας του ομαλοποιημένου ετεροφάσματος να παρακολουθεί μεταβολές βραχέος χρόνου, καθώς η εξίσωση (2.14) εκφυλίζεται σε μακροπρόθεσμο (long-term) μέσο όρο. Στην πράξη, το ετεροφάσμα ενδέχεται να μεταβάλλεται σημαντικά με την πάροδο του χρόνου, καθώς η πηγή ενδέχεται να κινείται και συνεπώς το TDOA ενδέχεται να μεταβάλλεται. Η σταθερά a πρέπει να είναι αρκετά μικρή ώστε το ομαλοποιημένο ετεροφάσμα να παρακολουθεί τέτοιες μεταβολές. Για την πρακτική επιβεβαίωση του επιχειρήματος αυτού εκτελέστηκε μία δεύτερη προσομοίωση, όμοια με την πρώτη,



Σχήμα 2.4: Αποτελέσματα εκτίμησης TDOA σε προσομοίωση με λευκή Γκαουσιανή τυχαία διαδικασία με σταθερή καθυστέρηση 0.75 δείγματα σε λευκό Γκαουσιανό θόρυβο: μέσος όρος της μέγιστης τιμής του CSP-CM και τυπική απόκλιση της εκτίμησης TDOA για κάθε πλαίσιο με χρήση της αναδρομικής μεθόδου για εκτίμηση του ετεροφάσματος για τιμές της σταθεράς $a = 0,\ 0.5,\ 0.9$.

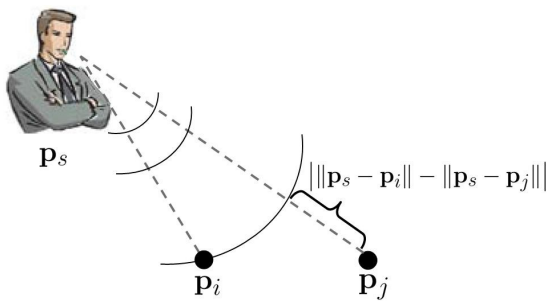


Σχήμα 2.5: Αποτελέσματα εκτίμησης TDOA σε προσομοίωση με λευκή Γκαουσιανή τυχαία διαδικασία με μεταβαλλόμενη καθυστέρηση σε λευκό Γκαουσιανό θόρυβο: μέσος όρος της μέγιστης τιμής του CSP-CM και τυπική απόκλιση της εκτίμησης TDOA για κάθε πλαίσιο με χρήση της αναδρομικής μεθόδου για εκτίμηση του ετεροφάσματος για τιμές της σταθεράς $a = 0,\ 0.1,\ 0.5$.

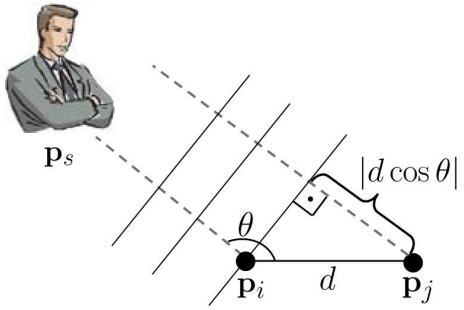
με μόνη διαφορά το γεγονός ότι το TDOA αυξάνει με το χρόνο προσομοιώνοντας μια κινούμενη πηγή. Τα αποτελέσματα για τιμές της σταθεράς $a = 0,\ 0.1,\ 0.5$ παρουσιάζονται στο Σχήμα 2.5. Είναι εμφανές ότι για να παρακολουθεί με ακρίβεια το ομαλοποιημένο ετεροφάσμα τη μεταβολή στο TDOA πρέπει η σταθερά a να είναι κοντά στο 0. Η τιμή $a = 0.1$ βελτιώνει την εκτίμηση TDOA, μειώνοντας την τυπική απόκλισή της, ενώ ταυτόχρονα επιτρέπει την παρακολούθηση των μεταβολών, συνεπώς είναι κατάλληλη για χρήση στην εκτίμηση ετεροφάσματος για τη μέθοδο CSP-CM τόσο για ακίνητες όσο και για κινούμενες πηγές.

2.5 Εκτίμηση κατεύθυνσης άφιξης

Από εκτιμήσεις TDOA για ζεύγη μικροφώνων μπορεί να εξαχθεί η κατεύθυνση άφιξης του σήματος πηγής (Direction of Arrival, DOA) στο εκάστοτε ζεύγος μικροφώνων. Έστω ένα ζεύγος μικροφώνων $\{i, j\}$, για το οποίο το TDOA ισούται με $\Delta\tau_{ij}$. Εν γένει, ο γεωμετρικός τόπος των σημείων του χώρου $\mathbf{p}$ που αποτελούν πιθανές



(α) Πηγή στο κοντινό πεδίο



(β) Πηγή στο μακρινό πεδίο

Σχήμα 2.6: Μοντέλα διάδοσης: (α) Διάδοση κοντινού πεδίου: το μέτωπο του ηχητικού κύματος θεωρείται σφαιρικό και (β) Διάδοση μακρινού πεδίου: το μέτωπο του κύματος θεωρείται επίπεδο.

θέσεις της πηγής περιγράφεται από τη σχέση (Σχήμα 2.6(α))):

$$\frac{\|\mathbf{p} - \mathbf{p}_i\| - \|\mathbf{p} - \mathbf{p}_j\|}{c} = \Delta\tau_{ij}, \tag{2.15}$$

όπου c η ταχύτητα του ήχου. Η εξίσωση αυτή (2.15) περιγράφει ένα υπερβολοειδές και είναι μη γραμμική ως προς $\mathbf{p}$. Υποθέτοντας όμως ένα μοντέλο διάδοσης μακρινού πεδίου (far-field propagation model), δηλαδή υποθέτοντας ότι η πηγή είναι επαρκώς μακριά από το ζεύγος μικροφώνων, μπορεί να θεωρηθεί ότι το μέτωπο του ηχητικού κύματος, όταν αυτό φτάνει στο ζεύγος μικροφώνων, δεν είναι πλέον σφαιρικό, αλλά έχει γίνει προσεγγιστικά επίπεδο (Σχήμα 2.6(β)). Τότε προκύπτει [5, 3]:

$$\frac{d\cos\theta}{c} = \Delta\tau_{ij}, \tag{2.16}$$

όπου d η απόσταση των μικροφώνων του ζεύγους και θ η γωνία άφιξης του ηχητικού κύματος ως προς το νοητό ευθύγραμμο τμήμα που συνδέει τα μικρόφωνα του ζεύγους. Από την εξίσωση (2.16) προκύπτει εύκολα η κατεύθυνση άφιξης (Direction of Arrival, DOA) θ του σήματος πηγής στο ζεύγος μικροφώνων ως:

$$\theta = \cos^{-1}\left(\frac{c\Delta\tau_{ij}}{d}\right). \tag{2.17}$$

2.6 Εντοπισμός θέσης πηγής με ελάχιστα τετράγωνα

Θα παρουσιαστεί μέθοδος εκτίμησης της θέσης της πηγής, η οποία συνδυάζει τις εκτιμώμενες κατευθύνσεις άφιξης (Directions of Arrival, DOAs) του σήματος πηγής

σε ζεύγη μικροφώνων με ελάχιστα τετράγωνα δίνοντας λύση κλειστής μορφής για την εκτίμηση της θέσης πηγής. Συνεπώς η προτεινόμενη αυτή μέθοδος είναι υπολογιστικά αποδοτικότερη από μεθόδους εντοπισμού θέσης οι οποίες περιλαμβάνουν υπολογιστικά ακριβούς αλγορίθμους αναζήτησης [12, 8].

Η προτεινόμενη μέθοδος υποθέτει ότι:

- Η πηγή και τα μικρόφωνα βρίσκονται στο ίδιο επίπεδο. Συνεπώς, τα διανύσματα θέσης της πηγής $\mathbf{p}_s$ και των μικροφώνων $\mathbf{p}_m,\ m = 1, 2, \ldots, N$ (Σχήμα 2.1(β)) θεωρούνται διδιάστατα.

- Η πηγή είναι επαρκώς μακριά από τα μικρόφωνα ώστε να ισχύει το μοντέλο διάδοσης μακρινού πεδίου. Συνεπώς η κατεύθυνση άφιξης του σήματος της πηγής στα μικρόφωνα κάθε ζεύγους μπορεί να υπολογιστεί χρησιμοποιώντας την εξίσωση (2.17).

Η υιοθέτηση μοντέλου διάδοσης μακρινού πεδίου διαχωρίζει την προτεινόμενη μέθοδο άλλες μεθόδους εντοπισμού θέσης με λύση κλειστής μορφής που υπάρχουν στη βιβλιγραφία, όπως [36, 33, 10, 22, 20]. Η υιοθέτηση μοντέλου διάδοσης μακρινού πεδίου δεν πλήττει σοβαρά την ακρίβεια της προτεινόμενης μεθόδου για πηγές με απόσταση μεγαλύτερη του 1m από τα μικρόφωνα, όπως θα δειχθεί πειραματικά, ενώ ταυρόχρονα απλοποιεί τη γεωμετρία του προβλήματος οδηγώντας σε γραμμικές εξισώσεις ευθειών κατεύθυνσης άφιξης, αντί των μη γραμμικών εξισώσεων υπερβολών (2.15), διευκολύνοντας την εξαγωγή απλής, αποδοτικής λύσης στο πρόβλημα εντοπισμού θέσης πηγής.

Έστω ένα σύνολο από M ζεύγη μικροφώνων $\{\{i_1, j_1\}, \{i_2, j_2\}, \ldots, \{i_k, j_k\}, \ldots, \{i_M, j_M\}\}$. Για το ζεύγος μικροφώνων k το TDOA είναι $\Delta\tau_{i_k j_k}$. Υπό την υπόθεση μοντέλου διάδοσης μακρινού πεδίου, μπορεί να θεωρηθεί ότι η πηγή βρίσκεται επί της ευθείας η οποία διέρχεται από το μέσον του ζεύγους των μικροφώνων:

$$\mathbf{p}_{0k} = \frac{\mathbf{p}_{i_k} + \mathbf{p}_{j_k}}{2} \tag{2.18}$$

και σχηματίζει με το νοητό ευθύγραμμο τμήμα το οποίο συνδέει τα μικρόφωνα του ζεύγους γωνία που προκύπτει σύμφωνα με την εξίσωση (2.17):

$$\theta_k = \cos^{-1}\left(\frac{c\Delta\tau_{i_k j_k}}{d_k}\right), \tag{2.19}$$

όπου c η ταχύτητα του ήχου, και d_k η απόσταση των μικροφώνων του ζεύγους k, $k = 1, 2, \ldots, M$ [5, 3]. Οι ευθείες αυτές στο εξής θα καλούνται DOA ευθείες (Direction of Arrival lines, DOA lines). Από τη γωνία θ_k μπορεί να υπολογιστεί το μοναδιαίο διάνυσμα $\mathbf{r}_k$, το οποίο είναι παράλληλο στην DOA ευθεία για το ζεύγος

μικροφώνων k. Συνεπώς η DOA ευθεία για το ζεύγος μικροφώνων k εκφράζεται σε παραμετρική μορφή ως εξής:

$$\mathbf{u}_k = \mathbf{p}_{0k} + \lambda \mathbf{r}_k,\ \lambda \in \mathbb{R} \tag{2.20}$$

όπου $\mathbf{u}_k$ το τυχαίο σημείο επί της DOA ευθείας για το ζεύγος μικροφώνων k. Ιδανικά, όλες οι DOA ευθείες τέμνονται σε ένα κοινό σημείο, τη θέση της πηγής. Στην πράξη, όμως, αυτό δε συμβαίνει εξ αιτίας σφαλμάτων στην εκτίμηση των TDOAs. Για το λόγο αυτό, για να εκτιμηθεί η θέση της πηγής από τις DOA ευθείες πρέπει να υιοθετηθεί μία προσέγγιση ελαχιστοποίησης κάποιου αντικειμενικού κριτηρίου σφάλματος.

Η προτεινόμενη προσέγγιση είναι να βρεθεί εκείνο το σημείο του επιπέδου, το οποίο ελαχιστοποιεί το άθροισμα των τετραγώνων των αποστάσεων από τις DOA ευθείες. Διαισθητικά, αυτό είναι το σημείο το οποίο βρίσκεται 'εγγύτερα' προς τις DOA ευθείες. Η χρήση αυτού του κριτηρίου διαχωρίζει την προτεινόμενη μέθοδο από τις μεθόδους της βιβλιογραφίας [5, 6].

Έστω $\mathbf{a}$ ένα τυχαίο σημείο στο επίπεδο. Έστω $\mathbf{a}_{\mathrm{proj}_k}$ η προβολή του σημείου $\mathbf{a}$ επί της DOA ευθείας για το ζεύγος μικρόφωνων k. Η απόσταση του $\mathbf{a}$ από την ευθεία είναι:

$$D_k(\mathbf{a}) = \|\mathbf{a} - \mathbf{a}_{\mathrm{proj}_k}\| \tag{2.21}$$

Το $\mathbf{a}_{\mathrm{proj}_k}$ ως σημείο της ευθείας αυτής ικανοποιεί τη σχέση:

$$\mathbf{a}_{\mathrm{proj}_k} = \mathbf{p}_{0k} + \lambda_{\mathbf{a}_{\mathrm{proj}_k}} \mathbf{r}_k, \tag{2.22}$$

Το διάνυσμα $\mathbf{a} - \mathbf{a}_{\mathrm{proj}_k}$ είναι κάθετο στην ευθεία, οπότε ικανοποιείται τη σχέση:

$$\mathbf{r}_k^T(\mathbf{a} - \mathbf{a}_{\mathrm{proj}_k}) = 0 \tag{2.23}$$

Από τις σχέσεις (2.22), (2.23) και από το γεγονός ότι $\|\mathbf{r}_k\| = 1$ προκύπτει:

$$\lambda_{\mathbf{a}_{\mathrm{proj}_k}} = \mathbf{r}_k^T(\mathbf{a} - \mathbf{p}_{0k}) \tag{2.24}$$

και συνεπώς

$$\mathbf{a}_{\mathrm{proj}_k} = \mathbf{p}_{0k} + \mathbf{r}_k\mathbf{r}_k^T(\mathbf{a} - \mathbf{p}_{0k}). \tag{2.25}$$

Επομένως, προκύπτει:

$$\mathbf{a} - \mathbf{a}_{\mathrm{proj}_k} = (\mathbf{I} - \mathbf{r}_k\mathbf{r}_k^T)(\mathbf{a} - \mathbf{p}_{0k}) \tag{2.26}$$

όπου $\mathbf{I}$ ο 2×2 μοναδιαίος πίνακας. Ορίζοντας:

$$\mathbf{A}_k = \mathbf{I} - \mathbf{r}_k\mathbf{r}_k^T, \tag{2.27}$$

η εξίσωση (2.21) γίνεται:

$$\begin{aligned} D_k^2(\mathbf{a}) &= \|\mathbf{a} - \mathbf{a}_{\mathrm{proj}_k}\|^2 \\ &= \|\mathbf{A}_k(\mathbf{a} - \mathbf{p}_{0k})\|^2 \\ &= (\mathbf{a} - \mathbf{p}_{0k})^T \mathbf{A}_k^T \mathbf{A}_k (\mathbf{a} - \mathbf{p}_{0k}) \\ D_k^2(\mathbf{a}) &= (\mathbf{a} - \mathbf{p}_{0k})^T \mathbf{A}_k (\mathbf{a} - \mathbf{p}_{0k}), \end{aligned} \tag{2.28}$$

όπου η τελευταία ισότητα ισχύει διότι $\mathbf{A}_k^T = \mathbf{A}_k$ και $\|\mathbf{r}_k\| = 1$ οπότε $\mathbf{A}_k^T \mathbf{A}_k = \mathbf{A}_k$.

Το κριτήριο σφάλματος το οποίο επιλέγεται να ελαχιστοποιηθεί είναι:

$$E(\mathbf{a}) = \sum_{k=1}^{M} D_k^2(\mathbf{a}) = \sum_{k=1}^{M} (\mathbf{a} - \mathbf{p}_{0k})^T \mathbf{A}_k (\mathbf{a} - \mathbf{p}_{0k}), \tag{2.29}$$

Η συνάρτηση $E(\mathbf{a})$ είναι τετραγωνική μορφή με gradient:

$$\nabla_{\mathbf{a}} E(\mathbf{a}) = 2 \sum_{k=1}^{M} \mathbf{A}_k (\mathbf{a} - \mathbf{p}_{0k}) \tag{2.30}$$

Συνεπώς, αν ο πίνακας:

$$\mathbf{A} = \sum_{k=1}^{M} \mathbf{A}_k \tag{2.31}$$

είναι αντιστρέψιμος, η συνάρτηση $E(\mathbf{a})$ έχει ολικό ελάχιστο στο σταθερό σημείο:

$$\hat{\mathbf{p}}_s = \mathbf{A}^{-1} \sum_{k=1}^{M} (\mathbf{A}_k \mathbf{p}_{0k}) \tag{2.32}$$

Η εκτίμηση της θέσης πηγής $\hat{\mathbf{p}}_s$ είναι βέλτιστη ως προς το κριτήριο αθροίσματος τετραγώνων αποστάσεων από τις DOA ευθείες, δηλαδή το $E(\hat{\mathbf{p}}_s)$ είναι το ελάχιστο δυνατό.

2.7 Ανίχνευση Ενεργού Ομιλητή

Κατά την εφαρμογή της προτεινόμενης μεθόδου εντοπισμού θέσης σε πραγματικά σήματα, απαιτείται μία μέθοδος διαχωρισμού φωνής από σιωπή, ώστε να παράγονται εκτιμήσεις TDOA και θέσης πηγής μόνο για εκείνα τα πλαίσια των σημάτων για τα οποία ο ομιλητής είναι ενεργός. Για να κρατηθεί η υπολογιστική πολυπλοκότητα της μεθόδου όσο το δυνατόν μικρότερη, δε γίνεται χρήση εξωτερικού αλγορίθμου ανίχνευσης φωνής (Voice Activity Detection, VAD). Αντ΄ αυτού ακολουθείται μία προσέγγιση παρόμοια με αυτήν που προτείνεται στο [29]: πλαίσια για τα οποία η

μέγιστη τιμή του CSP-CM υπερβαίνει ένα κατώφλι κατηγοριοποιούνται ως πλαίσια με ενεργή ομιλία, ενώ τα υπόλοιπα πλαίσια ως πλαίσια σιωπής. Το κατώφλι αυτό στο εξής θα αναφέρεται ως κατώφλι VAD (Voice Activity Detection).

Η εισαγωγή του κατωφλίου VAD δημουργεί ένα άλλο πρόβλημα: ανάλογα με τον προσανατολισμό του ομιλητή και την απόστασή του από ένα ζεύγος μικροφώνων, υπάρχει το ενδεχόμενο ένα πλαίσιο με ενεργό ομιλητή να κατηγοριοποιηθεί για κάποια ζεύγη μικροφώνων ως πλαίσιο φωνής, ενώ από άλλα ζεύγη μικροφώνων ως σιωπή. Συνεπώς για ένα δεδομένο πλαίσιο, ενδέχεται να είναι διαθέσιμες λιγότερες εκτιμήσεις TDOA και συνακόλουθα μικρότερος αριθμός DOA ευθειών από τον αριθμό των ζευγών μικροφώνων. Θεωρητικά, αρκούν δύο ευθείες για την εκτίμηση της θέσης πηγής μέσω της εξίσωσης (2.32). Στην πράξη ωστόσο η εκτίμηση της θέσης πηγής από πολύ λίγες DOA είναι πιο επιρρεπής σε σφάλματα των εκτιμήσεων TDOA (τα οποία προκαλούν απόκλιση των DOA ευθειών) και ενδέχεται να είναι αναξιόπιστη. Για την αποφυγή του προβλήματος αυτού τίθεται ένα κατώφλι αριθμού ευθειών και απορρίπτονται εκτιμήσεις θέσης πηγής για πλαίσια για τα οποία είναι διαθέσιμες λιγότερες DOA ευθείες από το κατώφλι. Το κατώφλι αυτό θα αναφέρεται στο εξής ως κατώφλι NDOAL (Number of DOA Lines).

2.8 Πειραματισμός σε πραγματικά σήματα

2.8.1 Μετρικές αξιολόγησης

Για την αξιολόγηση της ακρίβειας εντοπισμού θέσης της προτεινόμενης μεθόδου χρησιμοποιείται ο τρόπος αξιολόγησης μεθόδων εντοπισμού θέσης που περιγράφεται στο [30]. Θεωρείται ότι υπάρχει αναφορά (reference) την πραγματική θέση της πηγής σε κάθε πλαίσιο. Ως σφάλμα θέσης ορίζεται η Ευκλείδια απόσταση μεταξύ πραγματικής και εκτιμώμενης θέσης πηγής. Ορίζεται ένα κατώφλι για το σφάλμα θέσης, με το οποίο διαχωρίζονται τα σφάλματα θέσης σε μικρά (fine errors) και μεγάλα (gross errors). Σφάλματα με τιμή μικρότερη από το κατώφλι θεωρούνται μικρά, ενώ τα υπόλοιπα θεωρούνται μεγάλα. Ορίζονται οι εξής μετρικές:

- Pcor: ποσοστό επιτυχούς εντοπισμού θέσης, το οποίο ορίζεται ως το ποσοστό των πλαισίων με σφάλματα μικρότερα του κατωφλίου μεγάλων σφαλμάτων.
- Bias: ορίζεται για κάθε συντεταγμένη θέσης ως η μέση τιμή της διαφοράς (προσημασμένης, όχι απόλυτης) πραγματικής από εκτιμώμενη τιμή της συντεταγμένης.
- AAE: μέσο απόλυτο σφάλμα (Average Absolute Error)
- RMSE: ρίζα μέσου τετραγωνικού σφάλματος (Root Mean Square Error)

- DR: ρυθμός διαγραφής (Deletion Rate), ο οποίος ορίζεται ως το ποσοστό των πλαισίων τα οποία η μέθοδος εντοπισμού θέσης κατηγοριοποιεί ως πλαίσια σιωπής, παρότι ο ομιλητής είναι ενεργός, επί του συνόλου των πλαισίων για τα οποία ο ομιλητής είναι ενεργός.

- FAR: ρυθμός λάθος συναγερμών (False Alarm Rate), ο οποίος ορίζεται ως το ποσοστό των πλαισίων τα οποία η μέθοδος εντοπισμού θέσης κατηγοριοποιεί ως πλαίσια με ομιλία, παρότι ο ομιλητής δεν είναι ενεργός, επί του συνόλου των πλαισίων σιωπής.

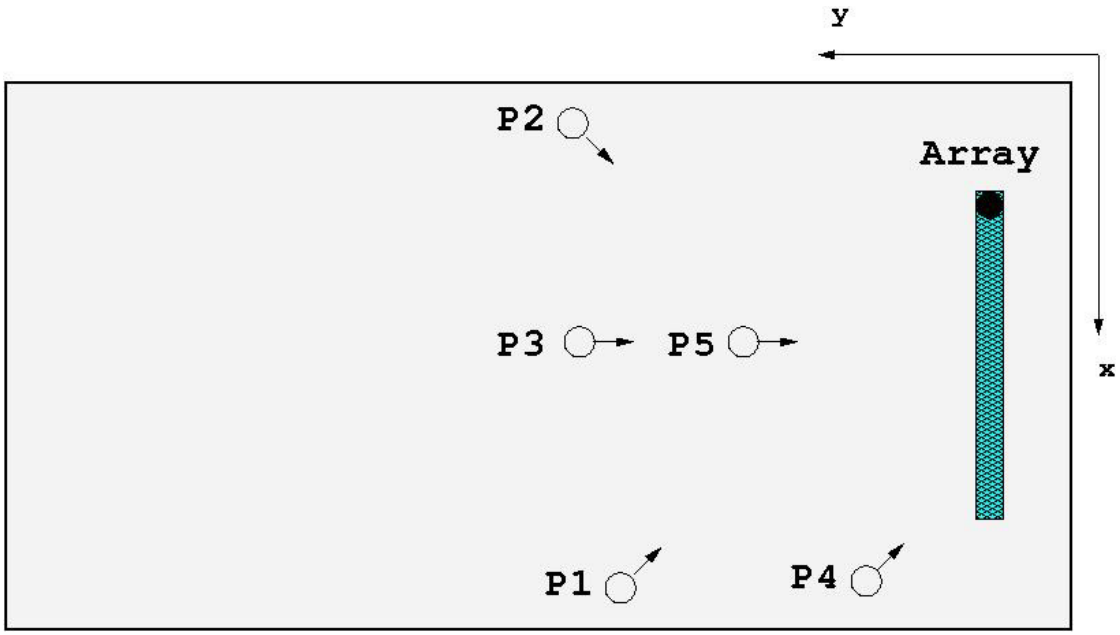
Εκτός από τις προαναφερθείες μετρικές, εισάγεται και μία άλλη μετρική για την αξιολόγηση της υπολογιστικής αποδοτικότητας του αλγορίθμου εντοπισμού θέσης και της καταλληλότητάς του για εφαρμογές πραγματικού χρόνου. Η μετρική αυτή είναι ο χρόνος επεξεργασίας, ο οποίος συμβολίζεται PT (Processing Time) και ο οποίος ισούται με το χρόνο εκτέλεσης του αλγορίθμου ως ποστοστό της διάρκειας των σημάτων υπό επεξεργασία. Η διάρκεια των σημάτων συμβολίζεται με RT (real time). Η υλοποίηση του αλγορίθμου εντοπισμού θέσης έγινε με MATLAB και συνεπώς δεν είναι υλοποίηση πραγματικού χρόνου, οπότε οι χρόνοι εκτέλεσης αναμένονται αυξημένοι σε σχέση με μια υλοποίηση πραγματικού χρόνου σε γλώσσα χαμηλότερου επιπέδου. Επιπλέον, οι χρόνοι εκτέλεσης μετρήθηκαν σε επίπεδο λογισμικού και όχι υλικού, συνεπώς η ακρίβεια της μέτρησης δεν είναι μεγαλύτερη από περίπου 1/3 του δευτερολέπτου. Παρά ταύτα, η μετρική PT παρέχει μία ένδειξη ως προς τη δυνατότητα υλοποίησης του αλγορίμου σε πραγματικό χρόνο: αν η PT είναι αρκετά μικρότερη από 1.0RT, τότε η επεξεργασία των δεδομένων είναι ταχύτερη από το ρυθμό με τον οποίο γίνονται διαθέσιμα, οπότε ο αλγόριθμος είναι δυνατόν να υλοποιηθεί σε πραγματικό χρόνο.

2.8.2 Βάσεις Δεδομένων

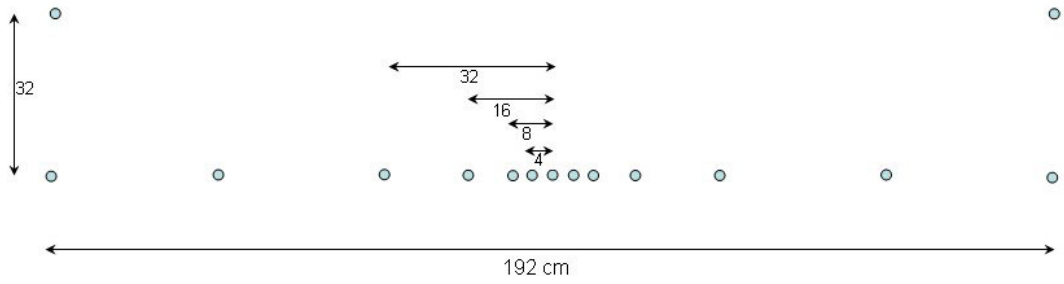
Τα πειράματα εκτελέστηκαν σε δύο βάσεις δεδομένων από το Ινστιτούτο Fondazione Bruno Kessler (FBK) [7] και στη βάση CMU (Carnegie Mellon University) [38].

Από τις βάσεις του ινστιτούτου FBK, η πρώτη βάση περιέχει πολυκαναλικές ηχογραφήσεις με μία αρμονικά φωλιασμένη γραμμική συστοιχία μικροφώνων (Harmonic Linear Array, HLA), ενώ η δεύτερη από ένα κατανεμημένο δίκτυο μικροφώνων (Distributed Microphone Network, DMN).

Η πειραματική εγκατάσταση και η γεωμετρία τοποθέτησης των μικροφώνων για τη βάση HLA παρουσιάζεται στο Σχήμα 2.7, ενώ για τη βάση DMN στο Σχήμα 2.8. Οι 5 πηγές που απεικονίζονται για κάθε βάση, δεν είναι ενεργές ταυτόχρονα. Υπάρχουν 5 διαφορετικές ηχογραφήσεις, σε καθεμία από τις οποίες ένας ακίνητος ομιλητής εκφέρει μία πρόταση στεκόμενος στην αντίστοιχη θέση με προσανατολισμό που δείχνεται από το βέλος στο σχήμα. Ο ρυθμός δειγματοληψίας των σημάτων είναι 48kHz για τη βάση HLA και 44.1kHz για τη βάση DMN.

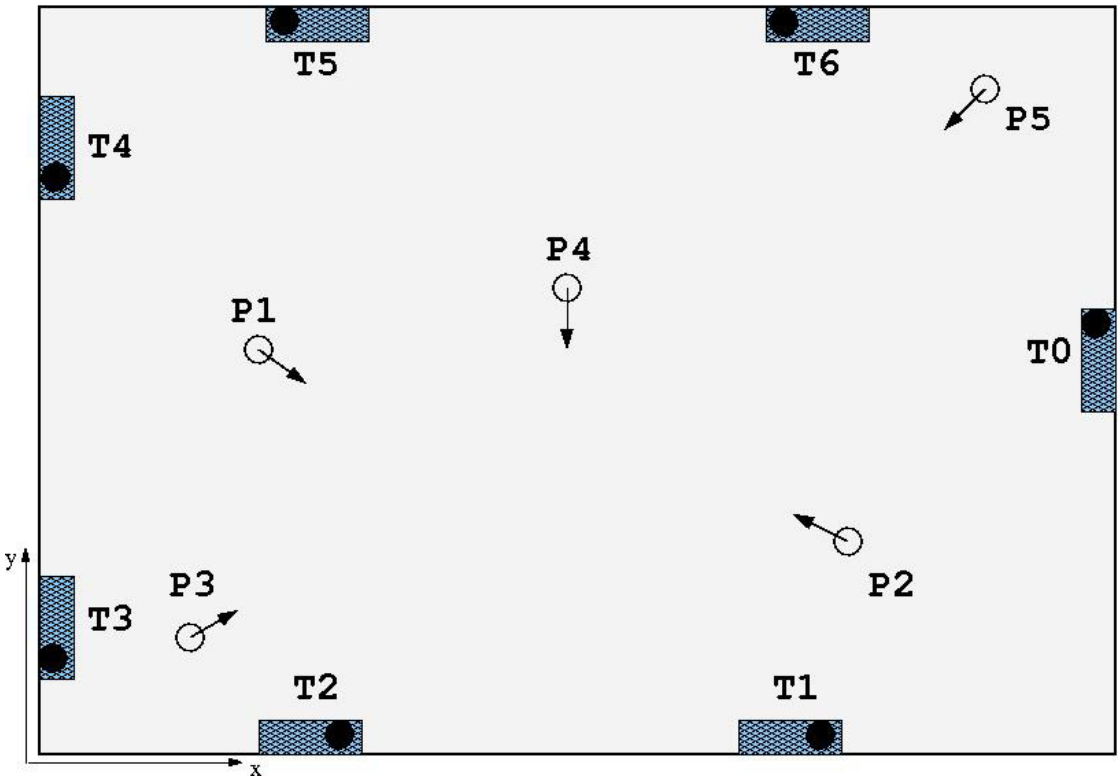


(α) Πειραματικό δωμάτιο για τη βάση HLA

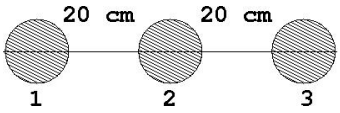


(β) Συστοιχία μικροφώνων HLA

Σχήμα 2.7: Πειραματική διάταξη για τη βάση HLA: οι ακουστικές πηγές (ομιλητές) απεικονίζονται ως δίσκοι και ο προσανατολισμός τους ως βέλος [7].



(α) Πειραματικό δωμάτιο για τη βάση DMN.



(β) Συστοιχίες μικροφώνων DMN.

Σχήμα 2.8: Πειραματική διάταξη για τη βάση DMN: οι ακουστικές πηγές (ομιλητές) απεικονίζονται ως δίσκοι και ο προσανατολισμός τους ως βέλος [7].

Η βάση CMU περιέχει πολυκαναλικές ηχογραφήσεις από μία γραμμική συστοιχία μικροφώνων αποτελούμενη από 8 μικρόφωνα με απόσταση 7cm μεταξύ τους. Οι ηχογραφήσεις έγιναν σε ένα θορυβώδες εργαστήριο υπολογιστών με πολλές πηγές θορύβου, όπως ανεμιστήρες υπολογιστών και εξαεριστήρες. Σε κάθε ηχογράφηση ένας ομιλητής ευρισκόμενος κάθετα ακριβώς μπροστά από τη συστοιχία και σε απόσταση 1m από αυτήν εκφωνεί μία πρόταση. Η βάση περιέχει 130 εκφωνήσεις από 10 ομιλητές, καθένας από τους οποίους εκφωνεί 13 από τις προτάσεις. Στη βάση CMU θα γίνουν πειράματα εντοπισμού θέσης πηγής και συνδυασμού της μεθόδου εντοπισμού θέσης πηγής με το σύστημα πολυκαναλικής αποθορυβοποίησης σημάτων που προτείνεται στο [25]. Τα συστήματα πολυκαναλικής αποθορυβοποίησης χρειάζονται ακριβή εκτίμηση θέσης πηγής για να είναι αποτελεσματικά. Θα συγκριθεί η ποιότητα του σήματος εξόδου που παράγει το πολυκαναλικό σύστημα όταν η πληροφορία της θέσης της πηγής παρέχεται από την προτεινόμενη μέθοδο εντοπισμού θέσης και όταν χρησιμοποιείται η γνώση της πραγματικής θέσης, ώστε να κριθεί αν η προτεινόμενη μέθοδος εντοπισμού θέσης είναι αρκετά ακριβής για να χρησιμοποιηθεί ως πρώτο στάδιο front-end σε πολυκαναλικά συστήματα αποθορυβοποίησης.

2.8.3 Πειραματικά αποτελέσματα

Βάσεις HLA και DMN

Για την εφαρμογή της προτεινόμενης μεθόδου εντοπισμού θέσης, τα μικρόφωνα των βάσεων πρέπει να ομαδοποιηθούν σε ζεύγη. Για τη βάση HLA χρησιμοποιήθηκαν τα 7 οριζόντια μικρόφωνα που σχηματίζουν μία γραμμική συστοιχία με ομοιόμορφη απόσταση 32cm μεταξύ μικροφώνων, ομαδοποιημένα σε 6 ζεύγη από γειτονικά μικρόφωνα. Για τη βάση DMN τα μικρόφωνα κάθε γραμμικής συστοιχίας 3 μικροφώνων ομαδοποιήθηκαν σε 2 ζεύγη γειτονικών μικροφώνων, ώστε να προκύψει ένα σύνολο από 14 ζεύγη μικροφώνων.

Στην πρώτη σειρά πειραμάτων, το γεγονός ότι ο ομιλητής είναι ακίνητος θεωρήθηκε γνωστό εκ των προτέρων (a priori), ώστε να είναι δυνατή η εκτίμηση TDOA με χρήση του αναλόγου της εξίσωσης (2.9) στο διακριτό χρόνο. Αποτέλεσμα αυτού είναι η παραγωγή μίας εκτίμησης TDOA ανά ζεύγος μικροφώνων και μίας εκτίμησης θέσης πηγής για όλη τη διάρκεια των σημάτων. Τα σήματα χωρίστηκαν σε πλαίσια διάρκειας 25ms με παραθύρωση Hanning. Το ετεροφάσμα εκτιμήθηκε με χρήση της εξίσωσης (2.14) θέτοντας $a = 0.1$. Το VAD κατώφλι τέθηκε στο 0.25 για τη βάση HLA και στο 0.2 για τη βάση DMN και η άθροιση σύμφωνα με την εξίσωση (2.9) έγινε μόνο για πλαίσια κατηγοριοποιημένα ως πλαίσια με ενεργό ομιλητή. Το κατώφλι NDOAL τέθηκε ίσο με τον αριθμό των ζευγών μικροφώνων και για τις δύο βάσεις. Τα αποτελέσματα των πειραμάτων παρουσιάζονται στον Πίνακα 2.3 για τη βάση HLA και στον Πίνακα 2.4 για τη βάση DMN. Σημειώνεται ότι στη συγκεκριμένη περίπτωση οι μετρικές Bias, AAE και RMSE που παρουσιάζονται ανά θέση ομιλητή

P	$\mathbf{p}_s$ (mm, mm)	$\hat{\mathbf{p}}_s$ (mm, mm)	Bias (mm, mm)	AAE (mm)	RMSE (mm)	PT (%RT)
1	(2800, 2400)	(2706, 2390)	(-94, -10)	95	95	30
2	(950, 3000)	(998, 3022)	(48, 22)	53	53	30
3	(1800, 3000)	(1838, 2996)	(38, -4)	38	38	30
4	(3000, 1200)	(3065, 1288)	(65, 88)	109	109	29
5	(1800, 1800)	(1830, 1780)	(30, -20)	36	36	29
Συνολικά	-	-	(17, 15)	66	73	30

Πίνακας 2.3: Πειραματικά αποτελέσματα εντοπισμού θέσης στη βάση HLA.

P	$\mathbf{p}_s$ (mm, mm)	$\hat{\mathbf{p}}_s$ (mm, mm)	Bias (mm, mm)	AAE (mm)	RMSE (mm)	PT (%RT)
1	(1280, 2450)	(1186, 2385)	(-94, -65)	114	114	60
2	(4280, 1235)	(4203, 1364)	(-77, 129)	150	150	60
3	(800, 1245)	(1000, 1259)	(200, 14)	201	201	59
4	(3080, 2440)	(3100, 2519)	(20, 79)	82	82	60
5	(4880, 3635)	(4775, 3575)	(-105, -60)	121	121	60
Συνολικά	-	-	(-11, 19)	133	139	60

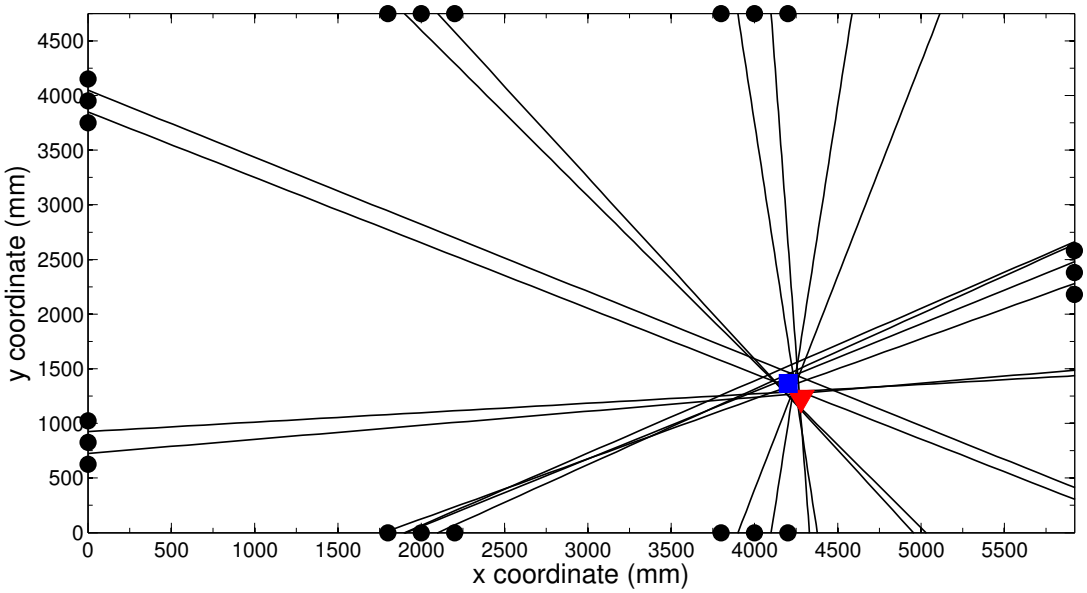
Πίνακας 2.4: Πειραματικά αποτελέσματα εντοπισμού θέσης στη βάση DMN.

δεν είναι μέσοι όροι, καθώς υπάρχει μία εκτίμηση θέσης πηγής για όλη τη διάρκεια των σημάτων σε κάθε θέση.

Η μέθοδος λειτουργεί ικανοποιητικά και στις δύο βάσεις με συνολικό μέσο και RMS σφάλμα κάτω από 15cm. Επιπλέον, ο χρόνος επεξεργασίας είναι μικρότερος από 0.75RT, γεγονός ενδεικτικό της καταλληλότητας της μεθόδου για εφαρμογές πραγματικού χρόνου.

Μία απεικόνιση της λειτουργίας της μεθόδου παρέχεται στο Σχήμα 2.9, στο οποίο απεικονίζεται ένας χάρτης του δωματίου DMN για θέση ομιλητή P2 μαζί με τις εκτιμώμενες ευθείες DOA. Τα μικρόφωνα απεικονίζονται ως μαύροι δίσκοι, η εκτιμώμενη θέση πηγής ως μπλε τετράγωνο και η πραγματική θέση πηγής ως κόκκινο τρίγωνο.

Σε ένα δεύτερο πείραμα, εφαρμόστηκε η προτεινόμενη μέθοδος εντοπισμού θέσης χωρίς υπόθεση ακίνητου ομιλητή, ώστε να παράγεται εκτίμηση θέσης ανά πλαίσιο σήματος. Για αυτό το πείραμα χρησιμοποιήθηκε η βάση DMN, η οποία όπως έδειξε το αυξημένο AAE σε σχέση με τη βάση HLA στην πρώτη σειρά πειραμάτων παρουσιάζει αυξημένη δυσκολία για το εξεταζόμενο πρόβλημα. Η προτεινόμενη μέθοδος εντοπισμού θέσης εφαρμόστηκε με την ίδια παραμετροποίηση όπως στην πρώτη σειρά πειραμάτων με μόνη διαφορά τα κατώφλια VAD και NDOAL τα οποία τέθηκαν στις τιμές 0.16 και 7 αντίστοιχα, ώστε να αυξηθεί η ευαισθησία ανίχνευσης ενεργού



Σχήμα 2.9: Χάρτης του δωματίου της βάσης DMN για θέση ομιλητή P2: τα μικρόφωνα απεικονίζονται ως μαύροι κύκλοι, οι μαύρες γραμμές δείχνουν τις εκτιμώμενες DOA ευθείες, η εκτιμώμενη με ελάχιστα τετράγωνα θέση πηγής απεικονίζεται ως μπλε τετράγωνο και η πραγματική θέση πηγής ως κόκκινο τρίγωνο.

P	Pcor	Bias		AAE		RMSE	
		fine	total	fine	total	fine	total
	(%)	(mm, mm)	(mm, cm)	(mm)	(mm)	(mm)	(mm)
1	89.90	(102, -86)	(129, -114)	213	267	232	341
2	93.62	(-61, 162)	(-88, 172)	194	228	207	274
3	91.94	(168, 64)	(181, 102)	194	245	197	310
4	93.90	(-88, -102)	(-85, -144)	170	209	191	280
5	100.00	(-100, -71)	(-100, -71)	131	131	140	140
Συνολικά	93.90	(2, -31)	(11, -39)	178	215	195	278

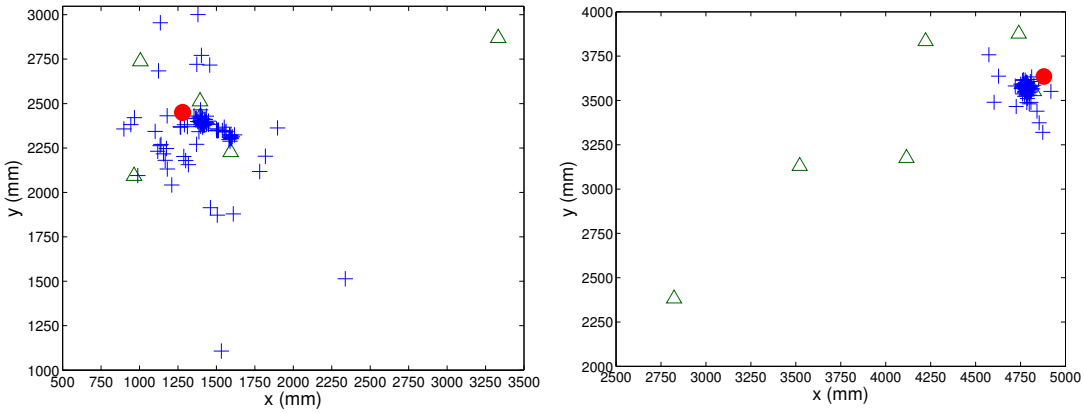
P	DR	FAR	PT
	(%)	(%)	(%RT)
1	90.43	0.90	70
2	94.07	0.57	68
3	93.82	0.22	69
4	89.84	0.25	69
5	90.49	0.49	68
Συνολικά	91.72	0.46	69

Πίνακας 2.5: Πειραματικά αποτελέσματα εντοπισμού θέσης στη βάση DMN χωρίς υπόθεση ακίνητου ομιλητή, ώστε να παράγεται εκτίμηση θέσης ανά πλαίσιο σήματος.

ομιλητή και να μειωθεί το DR. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 2.5. Το κατώφλι μεγάλων λαθών τέθηκε στα 50cm. Οι μετρικές Bias, AAE, RMSE έχουν υπολογιστεί αρχικά λαμβάνοντας υπ΄ όψιν μόνο τα μικρά σφάλματα (fine) και ακολούθως λαμβάνοντας υπ΄ όψιν όλα τα σφάλματα (total = fine+gross).

Οι υψηλές τιμές επιτυχούς εντοπισμού θέσης Pcor (επιτυχής είναι ο εντοπισμός θέσης με σφάλμα μικρότερο από 50cm) δείχνουν ότι η προτεινόμενη μέθοδος εντοπισμού θέσης είναι ακριβής, ενώ ταυτόχρονα ο μικρότερος από 0.75RT χρόνος εκτέλεσης δείχνει ότι ο αλγόριθμος είναι αρκετά ταχύς για χρήση σε εφαρμογές πραγματικού χρόνου. Το συνολικό AAE και RMSE είναι λιγότερο από 30cm. Ένα μειονέκτημα που παρουσιάζει η μέθοδος είναι το υψηλό DR, το οποίο μπορεί να διορθωθεί θέτοντας μικρότερη τιμή στο κατώφλι VAD. Ωστόσο κάτι τέτοιο μειώνει την ακρίβεια, καθώς εισάγονται αναξιόπιστες εκτιμήσεις TDOA. Μέσω του κατωφλίου VAD μπορεί να γίνει ανταλλαγή ακρίβειας Pcor για μικρότερο DR.

Ως οπτικοποίηση των αποτελεσμάτων αυτού του πειράματος, στο Σχήμα 2.10 παρουσιάζονται ενδεικτικά οι εκτιμώμενες θέσεις ομιλητή ανά πλαίσιο για τις θέσεις P1 και P5 της βάσης DMN. Οι εκτιμώμενες θέσεις εμφανίζονται ως μπλε σταυροί, οι λάθος συναγερμοί ως μπλε αστερίσκοι και οι η πραγματική θέση ομιλητή ως κόκκινος δίσκος.



(α) Θέση ομιλητή P1 (β) Θέση ομιλητή P5

Σχήμα 2.10: Αποτελέσματα εντοπισμού θέσης για τις θέσεις P1, P5 της βάσης DMN: οι κόκκινοι δίσκοι δείχνουν την πραγματική θέση πηγής, οι μπλε σταυροί τις εκτιμώμενες θέσεις και τα πράσινα τρίγωνα τους λάθος συναγερμούς.

Βάση CMU

Η προτεινόμενη μέθοδος εντοπισμού θέσης πηγής εφαρμόστηκε στη βάση CMU λαμβάνοντας υπ΄ όψιν τη στασιμότητα της πηγής και εκτιμώντας τα TDOAs μέσω του διακριτού αναλόγου της εξίσωσης (2.9). Επομένως για κάθε σήμα στη βάση παράγεται μία εκτίμηση TDOA ανά ζεύγος μικροφώνων και μία εκτίμηση της θέσης πηγής. Τα σήματα χωρίστηκαν σε ημιεπικαλυπτόμενα παράθυρα Hamming διάρκειας 25ms. Τα ετεροφάσματα εκτιμήθηκαν μέσω της (2.14) με $\alpha = 0.1$. Το κατώφλι VAD τέθηκε στο 0.5. Η επιλογή της υψηλής αυτής τιμής δικαιολογείται από το γεγονός ότι υπάρχει έντονος θόρυβος στις ηχογραφήσεις που προκαλεί εμφάνιση δευτερευουσών κορυφών στο CSP-CM. Το κατώφλι NDOAL τέθηκε στο 2. Δοκιμάστηκαν διάφοροι συνδυασμοί μικροφώνων σε ζεύγη: $P_1 = \{\{1,2\},\{2,3\},\ldots,\{7,8\}\}$ (7 ζεύγη μικροφώνων με 7cm απόσταση μεταξύ μικροφώνων), $P_2 = \{\{1,3\},\{2,4\},\ldots,\{6,8\}\}$ (6 ζεύγη μικροφώνων με 14cm απόσταση μεταξύ μικροφώνων), $P_3 = \{\{1,4\},\{2,5\},\ldots,\{5,8\}\}$ (5 ζεύγη μικροφώνων με 21cm απόσταση μικροφώνων), $P_4 = \{\{1,5\},\{2,6\},\{3,7\},\{4,8\}\}$ (4 ζεύγη μικροφώνων με 28cm απόσταση μεταξύ μικροφώνων), και τέλος το σύνολο P_{all} όλων των δυνατών συνδυασμών μικροφώνων. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 2.6. Λάθη μεγαλύτερα από 15cm κατηγοριοποιούνται ως μεγάλα.

Εφόσον μόνο μία εκτίμηση θέσης προκύπτει για κάθε σήμα της βάσης, οι μέσοι όροι στον Πίνακα 2.6 αφορούν το σύνολο των σημάτων της βάσης και όχι κάθε πλάισιο των σημάτων. Το DR προκύπτει ως ποσοστό των σημάτων για τα οποία δεν παρήχθη καμμία εκτίμηση θέσης, λόγω του υψηλού VAD κατωφλίου.

Τα αποτελέσματα δείχουν ότι καθώς αυξάνει η απόσταση των μικροφώνων εντός

P	Pcor	Bias		AAE		RMSE		DR	PT
		fine	total	fine	total	fine	total		
	(%)	(mm, mm)	(mm, mm)	mm	mm	mm	mm	(%)	(%RT)
1	86.15	(-24, -33)	(-25, 35)	57	147	67	436	0.00	8
2	89.15	(-19, -27)	(-17, 91)	52	171	65	696	0.77	7
3	93.60	(-25, -4)	(-25, 78)	57	152	63	632	3.85	5
4	87.93	(-33, -44)	(-29, 134)	66	230	72	1146	10.77	4
all	93.85	(-23, -30)	(-23, 29)	54	126	65	439	0.00	30

Πίνακας 2.6: Αποτελέσματα εντοπισμού θέσης στη βάση CMU.

Post-filter	SSNRE (dB) (εκτιμώμενη θέση πηγής)	SSNRE (dB) (πραγματική θέση πηγής)
MMSE	14.2320	13.9523
STSA	13.9078	13.6467
log-STSA	14.0848	13.8164

Πίνακας 2.7: Αποτελέσματα πολυκαναλικής αποθορυβοποίησης στη βάση CMU.

κάθε ζεύγους, αυξάνει και η ακρίβεια εκτίμησης θέσης. Όταν όμως η απόσταση αυτή γίνει μεγάλη τότε η μείωση του αριθμού των διαθέσιμων ζευγών μικροφώνων μειώνει την ευρωστία της εκτίμησης και η ακρίβεια μειώνεται. Η βέλτιση επίδοση του αλγορίθμου προκύπτει όταν λαμβάνονται όλα τα δυνατά ζεύγη μικροφώνων της συστοιχίας. Σε αυτήν την περίπτωση παρατηρείται και σημαντική αύξηση του χρόνου επεξεργασίας, ωστόσο ο χρόνος εκτέλεσης παραμένει μικρότερος από το 1/3 της διάρκειας του σήματος.

Οι εκτιμήσεις της θέσης πηγής που παρήχθησαν θεωρώντας την ομαδοποίηση μικροφώνων P_{all} χρησιμοποιήθηκαν για πολυκαναλική αποθορυβοποίηση με το πολυκαναλικό σύστημα που προτείνεται στο [25] και το οποίο θα αναλυθεί με λεπτομέρεια στο Κεφάλαιο 3. Το συνδυασμένο σύστημα εντοπισμού θέσης πηγής και πολυκαναλικής αποθορυβοποίησης παρουσιάζεται στο [32]. Έγινε σύγκριση της ποιότητας αποθορυβοποίησης με αυτήν που επιτυγχάνεται με βάση την πραγματική θέση της πηγής. Για αξιολόγηση της ποιότητας του αποθορυβοποιημένου σήματος χρησιμοποιήθηκε η μετρική segmental SNR (SSNR) [21]. Η βελτίωση του SSNR (SSNR Enhancement, SSNRE), η οποία υπολογίζεται ως η διαφορά του SSNR του σήματος στο κεντρικό μικρόφωνο της συστοιχίας από το SSNR της αποθορυβοποιημένης εξόδου, παρουσιάζεται στον Πίνακα 2.7. Η χρήση των εκτιμήσεων για τη θέση πηγής δεν πλήττει την αποτελεσματικότητα του συστήματος, καθώς το SSNRE προκύπτει σχεδόν ίσο με αυτό που επιτυγχάνεται με χρήση της πραγματικής θέσης πηγής. Συνεπώς η προτεινόμενη μέθοδος εντοπισμού θέσης πηγής είναι αρκετά ακριβής για

χρήση με συστήματα πολυκαναλικής αποθορυβοποίησης.

2.9 Συμπεράσματα

Προτάθηκε μία μέθοδος εντοπισμού θέσης πηγής, η οποία χρησιμοποιώντας ελάχιστα τετράγωνα καταλήγει σε τύπο κλειστής μορφής και είναι συνεπώς αποδοτικότερη από μεθόδους που χρησιμοποιούν αλγορίθμους αναζήτησης. Η προτεινόμενη μέθοδος βασίζεται σε εκτίμηση διαφοράς χρόνου άφιξης TDOA μεταξύ ζευγών μικροφώνων. Για την εκτίμηση αυτή προτάθηκε μία βελτιωμένη έκδοση μίας συχνά χρησιμοποιούμενης στην πράξη μεθόδου εκτίμησης TDOA, της μεθόδου CSP-CM. Ο προτεινόμενος αλγόριθμος εντοπισμού θέσης πηγής απεδείχθη πειραματικά με εφαρμογή σε σήματα ηχογραφημένα υπό πραγματικές συνθήκες ότι είναι ακριβής και υπολογιστικά αποδοτικός, με χρόνους εκτέλεσης μικρότερους από 1.0RT γεγονός που τον καθιστά κατάλληλο για εφαρμογές πραγματικού χρόνου.

Κεφάλαιο 3

ΠΟΛΥΚΑΝΑΛΙΚΗ ΑΠΟΘΟΡΥΒΟΠΟΙΗΣΗ ΟΜΙΛΙΑΣ

3.1 Διατύπωση του προβλήματος

Έστω μία συστοιχία N μικροφώνων σε ένα θορυβώδες περιβάλλον και μία πηγή επιθυμητού ακουστικού σήματος, το οποίο για το πρόβλημα υπό μελέτη είναι σήμα φωνής παραγόμενο από έναν ομιλητή. Τα σήματα που καταγράφονται από τα μικρόφωνα μπορούν να μοντελοποιηθούν ως εξής:

$$x_m(n) = d_m(n) * s(n) + v_i(n),\ m = 1,\ 2,\ \ldots,\ N \tag{3.1}$$

όπου n ο διακριτός χρόνος, το $*$ συμβολίζει συνέλιξη, $x_m(n)$ είναι το σήμα που καταγράφεται στο μικρόφωνο m, $s(n)$ είναι το επιθυμητό σήμα φωνής, $d_m(n)$ είναι η κρουστική απόκριση της ακουστικής διαδρομής (acoustic path impulse response) από την πηγή στο μικρόφωνο m και $v_i(n)$ είναι η συνιστώσα ανεπιθύμητου θορύβου. Η κρουστική απόκριση της ακουστικής διαδρομής από την πηγή σε κάθε μικροφωνο μοντελοποιεί τα φαινόμενα διάδοσης όπως η καθυστέρηση διάδοσης, η αντήχηση (reverberation), κ.ό.κ. Υποθέτοντας περιβάλλον χωρίς αντήχηση, η κρουστική απόκριση λαμβάνει τη μορφή

$$d(n) = a_m \delta(n - \tau_m), \tag{3.2}$$

όπου a_m και τ_m είναι η εξασθένηση και η καθυστέρηση διάδοσης. Στο πεδίο συχνότητας πρέπει να ακολουθηθεί ανάλυση βραχέος χρόνου, καθώς το σήμα φωνής $s(n)$ δεν είναι στάσιμο (stationary). Συνεπώς, χρησιμοποιείται ο μετασχηματισμός Fourier βραχέος χρόνου (Short Time Fourier Transform, STFT). Έστω $X_m(k, \ell)$, $S(k, \ell)$, $V_m(k, \ell)$, όπου k ο συχνοτικός δείκτης (frequency bin index) και ℓ ο δείκτης πλαισίου (frame index), οι STFTs των σημάτων $x_m(n)$, $s_m(n)$, $v_m(n)$, αντίστοιχα.

Έστω επίσης $D_m(k)$ ο διακριτός μετασχηματισμός Fourier (Discrete Fourier Transform, DFT) της κρουστικής απόκρισης $d_m(n)$. Η σχέση που συνδέει τον STFT του σήματος $x_m(n)$ με τους STFTs των συνιστωσών του μπορεί να εκφραστεί συμπαγώς ως εξής:

$$\mathbf{X}(k,l) = \mathbf{D}(k)S(k,l) + \mathbf{V}(k,l), \tag{3.3}$$

όπου

$$\begin{aligned}\mathbf{X}(k,\ell) &= [X_0(k,\ell),\, X_1(k,\ell),\, \ldots,\, X_{N-1}(k,\ell)]^T \\ \mathbf{D}(k) &= [D_0(k),\, D_1(k),\, \ldots,\, D_{N-1}(k)]^T \\ \mathbf{V}(k,\ell) &= [V_0(k,\ell),\, V_1(k,\ell),\, \ldots,\, V_{N-1}(k,\ell)]^T .\end{aligned}$$

Το διάνυσμα $\mathbf{D}(k)$ κωδικοποιεί τα χωρικά χαρακτηριστικά της συστοιχίας μικροφώνων και ονομάζεται διάνυσμα κατεύθυνσης της συστοιχίας (array steering vector) [40]. Σε περιβάλλον χωρίς αντήχηση κάθε στοιχείο του διανύσματος $\mathbf{D}$ εκφράζεται ως [16]:

$$D_m(k) = a_m e^{-j\omega_k \tau_m}, \tag{3.4}$$

όπου ω_k η διακριτού χρόνου κυκλική συχνότητα (dicrete-time angular frequency) που αντιστοιχεί στο συχνοτικό δείκτη k.

Το ζητούμενο στην πολυκαναλική αποθορυβοποίηση ομιλίας είναι η βέλτιστη εκτίμηση του σήματος πηγής $s(n)$ δεδομένων των θορυβωδών σημάτων $x_m(n)$. Το πλεονέκτημα σε σχέση με τη μονοκαναλική αποθορυβοποίηση έγκειται στη διαθεσιμότητα καταγραφών του σήματος πηγής από πολλαπλές θέσεις στο χώρο, με αποτέλεσμα τη δυνατότητα αξιοποίησης, παράλληλα με τα φασματικά χαρακτηριστικά και των χωρικών χαρακτηριστικών των σημάτων πηγής και θορύβου.

Στο εξής, θα παραλείπεται η εξάρτηση ποσοτήτων από τους δείκτες k, ℓ για καθαρότητα στις εξισώσεις. Με $\phi_{x_i x_i}$ θα συμβολίζεται το φάσμα ισχύος του σήματος x_i, ενώ με $\phi_{x_i x_j}$ το ετεροφάσμα (cross-power spectrum) των σημάτων $\phi_{x_i x_j}$.

3.2 Στατιστικό μοντέλο σημάτων

Για την εξαγωγή βέλτιστων εκτιμήσεων του σήματος φωνής απαιτείται γνώση των συναρτήσεων πυκνότητας πιθανότητας (probability density functions) των Fourier συντελεστών της ομιλίας και του θορύβου. Η γνώση αυτή δεν είναι διαθέσιμη στην πράξη και η εκτίμηση των αυτών των πυκνοτήτων πιθανότητας από τα πραγματικά σήματα είναι δύσκολη. Για το λόγο αυτό, γίνονται οι εξής επιπλέον υποθέσεις για το στατιστικό μοντέλο των σημάτων, οι οποίες βασίζονται στο κεντρικό οριακό θεώρημα (central limit theorem) [18]:

- Το σήμα πηγής $s(n)$ είναι Γκαουσιανή τυχαία διαδικασία (Gaussian random process) με μηδενική μέση τιμή και φάσμα ισχύος ϕ_{ss}

- Τα σήματα θορύβου $v_m(n)$ είναι Γκαουσιανές τυχαίες διαδικασίες (Gaussian random process) με μηδενική μέση τιμή και πίνακα ετεροφασματικής πυκνότητας (cross-spectral density matrix) $\mathbf{\Phi}_{vv} = \mathrm{E}\left[\mathbf{V}\mathbf{V}^H\right]$, όπου με $\mathrm{E}[\cdot]$ συμβολίζεται η μέση τιμή και με $(\cdot)^H$ ο ερμιτιανός ανάστροφος.
- Το σήμα πηγής είναι ασυσχέτιστο με τα σήματα θορύβου και οι Fourier συντελεστές καθενός σήματος είναι ανεξάρτητοι σε διαφορετικές συχνότητες.

3.3 Μοντελοποίηση πεδίου θορύβου

Οι πηγές θορύβου που υπάρχουν στο περιβάλλον όπου καταγράφεται το επιθυμητό σήμα πηγής δημιουργούν ένα πεδίο θορύβου (noise field). Τα χωρικά χαρακτηριστικά ενός πεδίου θορύβου κωδικοποιούνται στην ετεροσυσχέτιση $\phi_{v_i v_j}$ που παρουσιάζουν τα σήματα θορύβου σε διαφορετικές θέσεις στο χώρο. Επομένως, για το χαρακτηρισμό ενός πεδίου θορύβου μπορεί να χρησιμοποιηθεί μία μετρική βασισμένη στην ετεροσυσχέτιση των σημάτων θορύβου μεταξύ μικροφώνων της συστοιχίας. Μία τέτοια μετρική είναι η συνάρτηση χωρικής συνοχής (spatial coherence function), η οποία χρησιμοποιείται συχνά για τη μοντελοποίηση των χωρικών χαρακτηριστικών πεδίων θορύβου. Πρόκειται για μία κανονικοποιημένη έκδοση του ετεροφάσματος $\phi_{v_i v_j}$ που ορίζεται ως [17]:

$$C_{v_i v_j}(\omega) = \frac{\phi_{v_i v_j}(\omega)}{\sqrt{\phi_{v_i v_i}(\omega)\phi_{v_j v_j}(\omega)}}, \tag{3.5}$$

όπου ω η διακριτού χρόνου κυκλική συχνότητα. Η γνώση της συνάρτησης συνοχής του πεδίου αξιοποιείται σε διάφορες μεθόδους για πολυκαναλική αποθορυβοποίηση σημάτων [26, 25]. Η συνάρτηση συνοχής μπορεί να υπολογιστεί αναλυτικά για διάφορα μοντέλα πεδίων θορύβου. Δύο πολύ συχνά χρησιμοποιούμενα μοντέλα θορύβου για τα οποία είναι γνωστή η συνάρτηση συνοχής είναι το διάχυτο πεδίο θορύβου (diffuse noise field) και το εντοπισμένο πεδίο θορύβου (localized noise field).

Το διάχυτο πεδίο θορύβου είναι ένα χωρικά ομογενές και σφαιρικά ισοτροπικό πεδίο το οποίο παράγεται από διάδοση σημάτων θορύβου προς όλες τις κατευθύνσεις ταυτόχρονα. Η συνάρτηση συνοχής του διάχυτου πεδίου θορύβου είναι [17]:

$$C^{\mathrm{dif}}_{v_i v_j}(\omega) = \frac{\sin(\omega f_s r_{ij}/c)}{\omega f_s r_{ij}/c}, \tag{3.6}$$

όπου r_{ij} η απόσταση μεταξύ των μικροφώνων i, j. Σε ένα διάχυτο πεδίο θορύβου τα σήματα θορύβου μεταξύ μικροφώνων είναι κυρίως συσχετισμένα στις χαμηλές συχνότητες. Το μοντέλο του διάχυτου πεδίου θορύβου χρησιμοποιείται για περιβάλλοντα θορύβου όπως γραφεία, αυτοκίνητα, κ.ό.κ. στα οποία η υπόθεση διάχυτου θορύβου ισχύει με ικανοποιητική ακρίβεια στην πράξη [26, 27].

Το εντοπισμένο πεδίο θορύβου είναι το πεδίο θορύβου που παράγεται από μία σημειακή πηγή θορύβου σε συγκεκριμένη θέση στο χώρο. Αν η πηγή θορύβου παράγει το σήμα $v(n)$ τότε στα μικρόφωνα λαμβάνονται (σύμφωνα με την υπόθεση περιβάλλοντος χωρίς αντήχηση) τα σήματα θορύβου $v_m(n) = a_{v_m} v(n - \tau_{v_m})$, όπου a_{v_m} και τ_{v_m} η εξασθένηση και η καθυστέρηση διάδοσης του σήματος θορύβου προς το μικρόφωνο m, αντίστοιχα. Συνεπώς προκύπτει [15]:

$$C^{\text{loc}}_{v_i v_j}(\omega) = e^{-j\omega(\tau_{v_i} - \tau_{v_j})}. \tag{3.7}$$

Το μοντέλο εντοπισμένου θορύβου είναι κατάλληλο για εφαρμογές στις οποίες στο επιθυμητό σήμα πηγής υπάρχει παρεμβολή παραγόμενη από μία εντοπισμένη στο χώρο πηγή, όπως δεύτερος ομιλητής, ραδιόφωνο, κ.ό.κ.

Στο εξής, γίνεται η υπόθεση ότι η συνάρτηση συνοχής του θορύβου είναι γνωστή εκ των προτέρων. Επίσης γίνεται η υπόθεση ότι **το πεδίο θορύβου είναι ομογενές**, δηλαδή το φάσμα ισχύος του θορύβου είναι ίδιο σε κάθε μικρόφωνο:

$$\phi_{v_m v_m} = \phi_{vv}, \forall m, \tag{3.8}$$

συνεπώς προκύπτει:

$$C_{v_i v_j} = \frac{\phi_{v_i v_j}}{\phi_{vv}}, \tag{3.9}$$

Με τη βοήθεια της συνάρτησης συνοχής του θορύβου ορίζεται ο πίνακας συνοχής του θορύβου (noise field coherence matrix) ως:

$$\mathbf{C}_{vv} = \begin{pmatrix} 1 & C_{v_0 v_1} & \dots & C_{v_0 v_{N-1}} \\ C_{v_1 v_0} & 1 & & \\ \vdots & & \ddots & \\ C_{v_{N-1} v_0} & \dots & & 1 \end{pmatrix}. \tag{3.10}$$

Ο πίνακας ετεροφασματικής πυκνότητας του θορύβου με βάση τα προαναφερθέντα μπορεί να γραφεί:

$$\mathbf{\Phi}_{vv} = \phi_{vv} \mathbf{C}_{vv}. \tag{3.11}$$

3.4 Υπάρχουσες μέθοδοι

Η χωρική πληροφορία που παρέχεται από την καταγραφή του σήματος ομιλίας από πολλαπλές θέσεις στο χώρο μπορεί να αξιοποιηθεί μέσω τεχνικών beamforming [41, 40]. Το beamforming συνίσταται σε γραμμικό φιλτράρισμα κάθε σήματος $x_m(n)$ και ακολούθως άθροιση των εξόδων των φίλτρων για την παραγωγή της τελικής εξόδου, διαδικασία γνωστή ως και ως filter-and-sum beamforming [23]. Στο πεδίο συχνότητας το beamforming μοντελοποιείται ως:

$$Y = \mathbf{W}^H \mathbf{X}, \tag{3.12}$$

όπου Y η έξοδος του beamformer στο πεδίο συχνότητας και

$$\mathbf{W}(k) = [W_0(k), W_1(k), \ldots, W_{N-1}(k)]^T, \tag{3.13}$$

όπου W_m η απόκριση συχνότητας του φίλτρου που εφαρμόζεται στο σήμα του μικροφώνου m.

Η αποθορυβοποίηση που παρέχεται μόνο από beamforming δεν είναι συνήθως επαρκής και στην πράξη συχνά εφαρμόζεται post-filtering στην έξοδο του beamformer. Για τη βέλτιστη εκτίμηση του σήματος φωνής $s(n)$ από τα θορυβώδη σήματα $v_m(n)$, μερικά συχνά χρησιμοποιούμενα κριτήρια βελτιστοποίησης είναι το μέσο τετραγωνικό σφάλμα (Mean Square Error, MSE), το MSE του φασματικού πλάτους (MSE of spectral amplitude) και το MSE του λογαρίθμου του φασματικού πλάτους (MSE of log-spectral amplitude). Από την ελαχιστοποίηση των κριτηρίων αυτών προκύπτουν ο γραμμικός εκτιμητής (estimator) Wiener ή ελάχιστου MSE (Minimum MSE, MMSE) και οι μη γραμμικοί εκτιμητές βραχέος χρόνου φασματικού πλάτους (Short-Time Spectral Amplitude, STSA) [18] και βραχέος χρόνου λογαριθμικού φασματικού πλάτους (log-STSA) [19]. Έχει αποδειχθεί ότι καθένα από αυτά τα πολυκαναλικά φίλτρα είναι ισοδύναμο με έναν beamformer ελάχιστης μεταβλητότητας με απόκριση χωρίς παραμόρφωση (Minimum Variance Distortionless Response, MVDR) ακολουθούμενο από ένα μονοκαναλικό μεταφίλτρο (post-filter), το οποίο δέχεται ως είσοδο την έξοδο του beamformer [35, 40, 2]. Το γεγονός αυτό δίνει κίνητρο για χρήση πολυκαναλικών συστημάτων αποθορυβοποίησης που χρησιμοποιούν MVDR beamforming με post-filtering.

3.4.1 MVDR beamformer

Ο MVDR beamformer προκύπτει ελαχιστοποιώντας το φάσμα ισχύος της εξόδου του beamformer υπό τον περιορισμό το επιθυμητό σήμα πηγής να παραμένει αναλλοίωτο. Από τις εξισώσεις (3.12) και (3.3) προκύπτει ότι η έξοδος του beamformer είναι:

$$Y = \mathbf{W}^H \mathbf{D} S + \mathbf{W}^H \mathbf{V} \tag{3.14}$$

Συνεπώς ο MVDR beamformer προκύπτει από την ελαχιστοποίηση του φάσματος ισχύος:

$$\phi_{yy} = \mathbf{W}^H \mathbf{\Phi}_{vv} \mathbf{W}^H \tag{3.15}$$

υπό τον περιορισμό:

$$\mathbf{W}^H \mathbf{D} = 1. \tag{3.16}$$

Από τη λύση του προβλήματος αυτού προκύπτει ο MVDR beamformer [4]:

$$\mathbf{W}^H_{\text{MVDR}} = \frac{\mathbf{D}^H \mathbf{\Phi}_{vv}^{-1}}{\mathbf{D}^H \mathbf{\Phi}_{vv}^{-1} \mathbf{D}} = \frac{\mathbf{D}^H \mathbf{C}_{vv}^{-1}}{\mathbf{D}^H \mathbf{C}_{vv}^{-1} \mathbf{D}}, \tag{3.17}$$

όπου η δεύτερη ισότητα προκύπτει από την εξίσωση (3.9). Η έξοδος του MVDR beamformer προκύπτει με βάση την εξίσωση (3.14):

$$Y = S + \frac{\mathbf{D}^H \mathbf{C}_{vv}^{-1} \mathbf{V}}{\mathbf{D}^H \mathbf{C}_{vv}^{-1} \mathbf{D}} \tag{3.18}$$

Το φάσμα ισχύος του θορύβου στην έξοδο του MVDR beamformer, το οποίο θα συμβολίζεται ϕ_{nn} προκύπτει συνεπώς:

$$\phi_{nn} = \mathbf{W}_{\text{MVDR}}^H \mathbf{\Phi}_{vv} \mathbf{W}_{\text{MVDR}} = \frac{\phi_{vv}}{\mathbf{D}^H \mathbf{C}_{vv}^{-1} \mathbf{D}}, \tag{3.19}$$

όπου η δεύτερη ισότητα προκύπτει με χρήση της εξίσωσης (3.9).

3.4.2 Post-filters

MMSE

Το πολυκαναλικό Wiener φίλτρο $\mathbf{W}_{\text{MMSE}}$ παράγει την εκτίμηση $Y = \mathbf{W}_{\text{MMSE}}^H \mathbf{X}$ για το σήμα εισόδου που ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα $E\left[(Y - X)^2\right]$. Το πολυκαναλικό αυτό φίλτρο έχει αποδειχθεί ισοδύναμο με MVDR beamformer ακολουθούμενου από μονοκαναλικό Wiener φιλτράρισμα της εξόδου του MVDR beamformer [35, 40]. Συνεπώς, η απόκριση συχνότητας του MMSE post-filter δίνεται από:

$$G_w = \frac{\phi_{ss}}{\phi_{ss} + \phi_{nn}} \tag{3.20}$$

STSA

Για την υλοποίηση πολυκαναλικής STSA εκτίμησης, η συνάρτηση μεταφοράς του post-filter που επιδρά στην έξοδο του beamformer είναι [18]:

$$G(u) = \Gamma(1.5) \frac{\sqrt{u}}{\gamma} \exp\left(-\frac{u}{2}\right) \left[(1+u)\, I_0\left(\frac{u}{2}\right) + u I_1\left(\frac{u}{2}\right)\right], \tag{3.21}$$

όπου Γ είναι η γάμμα συνάρτηση και I_0, I_1 είναι οι τροποποιημένες συναρτήσεις Bessel μηδενικής και πρώτης τάξης, αντίστοιχα.. Η μεταβλητή u ορίζεται ως

$$u = \frac{\xi}{1+\xi} \cdot \gamma, \tag{3.22}$$

όπου τα ξ και γ ορίζονται ως:

$$\xi = \frac{\phi_{ss}}{\phi_{nn}}, \quad \gamma = \frac{R^2}{\phi_{nn}}, \tag{3.23}$$

όπου R είναι το φασματικό πλάτος του Y, $Y(k,\ell) = R(k,\ell) e^{j\vartheta(k,\ell)}$.

log-STSA

Για την υλοποίηση του πολυκαναλικού log-STSA εκτιμητή, η συνάρτηση μεταφοράς του post-filter προκύπτει [19]:

$$G_{log}(u) = \frac{\xi}{1+\xi} \exp\left(\frac{1}{2}\int_u^\infty \frac{e^{-t}}{t}\, dt\right) \tag{3.24}$$

3.4.3 Εκτίμηση παραμέτρων των post-filters

Για τον υπολογισμό των συναρτήσεων μεταφοράς των post-filters που προαναφέρθηκαν είναι απαραίτητη η εκτίμηση των φασμάτων ισχύος ϕ_{ss} και ϕ_{nn}. Για την εκτίμηση των φασμάτων ισχύος αυτών έχουν προταθεί στη βιβλιογραφία διάφορες μέθοδοι, από τις οποίες θα παρουσιαστούν αυτές που προτείνονται στα [42, 26, 25].

Κοινό στοιχείο αυτών των μεθόδων είναι ότι, αρχικά, γίνεται χρονική ευθυγράμμιση των σημάτων $x_m(n)$ ώστε να αντισταθμιστεί η καθυστέρηση διάδοσης τ_m του σήματος πηγής $s(n)$ στα μικρόφωνα. Κατόπιν της χρονικής ευθυγράμμισης προκύπτουν τα σήματα:

$$x'_m(n) = x_m(n+\tau_m) = s(n) + v'_m(n),\ m = 1,\, 2,\, \ldots,\, N, \tag{3.25}$$

όπου $v'_m(n) = v_m(n+\tau_m)$. Για τα ευθυγραμμισμένα σήματα, στο πεδίο συχνότητας ισχύει (κατ΄ αναλογία με την εξίσωση (3.3)):

$$\mathbf{X}' = \mathbf{1}S + \mathbf{V}', \tag{3.26}$$

όπου $\mathbf{1}$ είναι διάνυσμα μονάδων διάστασης $N \times 1$.

Σύμφωνα με τις υποθέσεις που έχουν γίνει (Ενότητα 3.2) ότι το σήμα πηγής και ο θόρυβος είναι ασυσχέτιστα και ότι το πεδίο θορύβου είναι ομογενές, για τα φάσματα ισχύος και την ετεροφαματική πυκνότητα ισχύος των ευθυγραμμισμένων θορυβωδών σημάτων προκύπτουν τα εξής:

$$\phi_{x'_i x'_i} = \phi_{ss} + \phi_{vv} \tag{3.27}$$

$$\phi_{x'_i x'_j} = \phi_{ss} + \phi_{v'_i v'_j}, \tag{3.28}$$

Η πρώτη εξίσωση προκύπτει λαμβάνοντας υπ΄ όψιν ότι $\phi_{v'_i v'_i} = \phi_{v_i v_i} = \phi_{vv}$, καθώς η χρονική ολίσθηση δεν επηρεάζει το φάσμα ισχύος μιας τυχαίας διαδικασίας.

Ο Zelinski στο [42] θεωρεί ότι τα σήματα θορύβου μεταξύ μικροφώνων είναι ασυσχέτιστα (ισοδύναμα $C_{v_i v_j} = 0$ για $i \neq j$) και καταλήγει στις εξισώσεις:

$$\phi_{x'_i x'_i} = \phi_{ss} + \phi_{vv} \tag{3.29}$$

$$\phi_{x'_i x'_j} = \phi_{ss} \tag{3.30}$$

Στη συνέχεια, εκτιμά την απόκριση συχνότητας του Wiener post-filter (εξίσωση (3.20)) ως:

$$\hat{G}_w = \frac{\frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{Re}\{\hat{\phi}_{x'_i x'_j}\}}{\frac{1}{N} \sum_{i=1}^{N} \hat{\phi}_{x'_i x'_i}}, \tag{3.31}$$

όπου $\hat{\phi}_{x'_i x'_i}$ και $\hat{\phi}_{x'_i x'_j}$ τα εκτιμώμενα φάσματα ισχύος και $Re\{\cdot\}$ το πραγματικό μέρος. Οι μέσοι όροι για όλους τους δυνατούς συνδυασμούς μικροφώνων λαμβάνονται για αυξημένη ευρωστία στην εκτίμηση. Ένα μειονέκτημα της εκτίμησης (3.31) είναι ότι στον παρονομαστή υπερεκτιμάται η ισχύς του θορύβου στην έξοδο του MVDR beamformer, καθώς η εκτίμηση γίνεται με βάση τα θορυβώδη σήματα στην είσοδο του beamformer. Ένα δεύτερο μειονέκτημα είναι η υπόθεση ασυσχέτιστων σημάτων θορύβου μεταξύ μικροφώνων, η οποία για πολλά περιβάλλοντα θορύβου δεν ισχύει στην πράξη [26]. Το τρίτο μειονέκτημα της μεθόδου είναι ότι η εκτίμηση δεν μπορεί να γενικευτεί για άλλα είδη post-filters πέραν του Wiener.

Οι McCowan και Bourlard στο [26] γενίκευσαν τη μέθοδο του Zelinski [42], θεωρώντας ότι είναι γνωστό ένα μοντέλο για το περιβάλλον θορύβου μέσω της συνάρτησης συνοχής $C_{v_i v_j}$. Υπό την προϋπόθεση ομογενούς πεδίου θορύβου χρησιμοποίησαν τη σχέση:

$$\phi_{v'_i v'_j} = \phi_{vv} C_{v_i v_j} \tag{3.32}$$

ώστε από τις εξισώσεις (3.27), (3.28) να καταλήξουν στις εξισώσεις:

$$\phi_{x'_i x'_i} = \phi_{ss} + \phi_{vv} \tag{3.33}$$

$$\phi_{x'_i x'_j} = \phi_{ss} + \phi_{vv} C_{v_i v_j}, \tag{3.34}$$

Λύνοντας το γραμμικό σύστημα των εξισώσεων (3.33), (3.34) κατέληξαν στην εξής εκτίμηση για το φάσμα ισχύος του σήματος πηγής από τα σήματα των μικροφώνων i, j:

$$\hat{\phi}_{ss}^{ij} = \frac{\mathrm{Re}\left\{\hat{\phi}_{x'_i x'_j}\right\} - \frac{1}{2}\left(\hat{\phi}_{x'_i x'_i} + \hat{\phi}_{x'_j x'_j}\right) \mathrm{Re}\left\{C_{v_i v_j}\right\}}{1 - \mathrm{Re}\left\{C_{v_i v_j}\right\}} \tag{3.35}$$

Στον αριθμητή, αντί του $\hat{\phi}_{x'_i x'_i}$ λαμβάνεται ο μέσος όρος $\frac{1}{2}\left(\hat{\phi}_{x'_i x'_i} + \hat{\phi}_{x'_j x'_j}\right)$ ως πιο εύρωστη εκτίμηση του φάσματος ισχύος των σημάτων $x'_i(n)$, το οποίο είναι ίδιο για όλα τα σήματα x'_i (εξίσωση (3.27)). Για την τελική εκτίμηση του φάσματος ισχύος του σήματος πηγής και αριθμητή του Wiener post-filter έλαβαν το μέσο όρο για όλους τους συνδυασμούς μικροφώνων:

$$\hat{\phi}_{ss} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\phi}_{ss}^{ij}. \tag{3.36}$$

Για την εκτίμηση του παρονομαστή του Wiener post-filter ακολούθησαν την ίδια διαδικασία με τον Zelinski [42]. Η μέθοδος των McCowan και Bourlard, παρότι γενικεύει την εκτίμηση του αριθμητή του Wiener post-filter για οποιοδήποτε περιβάλλον θορύβου, εξακολουθεί να είναι υποβέλτιστη, λόγω της υπερεκτίμησης της ισχύος του θορύβου στον παρονομαστή. Επίσης, δε γενικεύεται για άλλα post-filters.

Οι Lefkimmiatis και Maragos στο [25] γενίκευσαν τη μέθοδο των McCowan και Bourlard [26], χρησιμοποιώντας την ίδια εκτίμηση για το φάσμα ισχύος της πηγής ϕ_{ss}, αλλά λύνοντας το γραμμικό σύστημα των εξισώσεων (3.33), (3.34) και ως προς το φάσμα ισχύος του θορύβου ϕ_{vv}. Κατέληξαν στην εξής εκτίμηση για το ϕ_{vv} από τα σήματα των μικροφώνων i, j:

$$\hat{\phi}_{vv}^{ij} = \frac{\frac{1}{2}\left(\hat{\phi}_{x_i' x_i'} + \hat{\phi}_{x_j' x_j'}\right) - \mathrm{Re}\left\{\hat{\phi}_{x_i' x_j'}\right\}}{1 - \mathrm{Re}\left\{C_{v_i v_j}\right\}}, \tag{3.37}$$

Η τελική εκτίμηση προκύπτει από μέσο όρο για όλα τα ζεύγη μικροφώνων:

$$\hat{\phi}_{vv} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N-1} \hat{\phi}_{vv}^{ij}. \tag{3.38}$$

Χρησιμοποιώντας την εκτίμηση για το ϕ_{vv}, για την εκτίμηση του φάσματος ισχύος ϕ_{nn} του θορύβου στην έξοδο του MVDR beamformer χρησιμοποίησαν την εξίσωση (3.19). Για την εκτίμηση του παρονομαστή $\phi_{ss} + \phi_{nn}$ του Wiener post-filter χρησιμοποίησαν τη διαθέσιμη πλέον εκτίμηση φάσματος ισχύος του θορύβου στην έξοδο του MVDR beamformer καταλήγοντας σε βέλτιστη λύση. Παράλληλα, η ύπαρξη εκτίμησης για το ϕ_{nn} επιτρέπει γενίκευση της μεθόδου και για άλλα post-filters εκτός του Wiener, όπως τα μη γραμμικά STSA και log-STSA post-filters.

3.5 Βελτίωση της εκτίμησης παραμέτρων των post-filters

Οι μέθοδοι εκτίμησης φασμάτων ισχύος πηγής ϕ_{ss} και θορύβου ϕ_{nn} που προτείνονται στα [26, 25] είναι θεωρητικά υποβέλτιστες, καθώς αγνοούν τη μεταβολή που προκαλεί η χρονική ευθυγράμμιση των σημάτων εισόδου $x_m(n)$ στο ετεροφάσμα του θορύβου. Η χρονική μετατόπιση των σημάτων που καταγράφονται από τα μικρόφωνα ώστε να ευθυγραμμιστεί χρονικά το σήμα πηγής σε όλα τα μικρόφωνα πριν από περαιτέρω επεξεργασία είναι συνήθης πρακτική, ωστόσο πρέπει να συνοδεύεται από επανεκτίμηση των φασματικών χαρακτηριστικών των σημάτων [4].

Συγκεκριμένα, μετά από τη χρονική ευθυγράμμιση των σημάτων των μικροφώνων, η συνιστώσα θορύβου σε κάθε μικρόφωνο μετατρέπεται σε:

$$v_m'(n) = v_m(n + \tau_m),\ m = 1, 2, \ldots, N \tag{3.39}$$

Συνεπώς προκύπτει:

$$\phi_{v'_i v'_j} = \mathrm{E}\left[V_i e^{+\jmath\omega_k\tau_i} V_j^* e^{-\jmath\omega_k\tau_j}\right] = \phi_{v_i v_j} e^{-\jmath\omega_k(\tau_j-\tau_i)}, \tag{3.40}$$

όπου ω_k η διακριτού χρόνου κυκλική συχνότητα που αντιστοιχεί στο συχνοτικό δείκτη k. Άρα χρησιμοποιώντας την εξίσωση (3.9) προκύπτει ότι:

$$\phi_{v'_i v'_j} = \phi_{vv} C_{v_i v_j} e^{-\jmath\omega_k(\tau_j-\tau_i)} = \phi_{vv} C_{v'_i v'_j}, \tag{3.41}$$

όπου $C_{v'_i v'_j}$ η συνάρτηση συνοχής των σημάτων θορύβου μετά τη χρονική ευθυγράμμιση:

$$C_{v'_i v'_j} = \frac{\phi_{v'_i v'_j}}{\phi_{vv}} = C_{v_i v_j} e^{-\jmath\omega_k(\tau_j-\tau_i)} \tag{3.42}$$

Στα [26, 25] κατά την εφαρμογή των εκτιμήσεων (3.35), (3.37) χρησιμοποιείται η συνάρτηση συνοχής θορύβου $C_{v_i v_j}$ πριν τη χρονική ευθυγράμμιση των σημάτων. Συγκεκριμένα, χρησιμοποιείται η συνάρτηση συνοχής του διάχυτου πεδίου θορύβου που δίνεται από την εξίσωση (3.6). Δηλαδή, έχει θεωρηθεί ότι ισχύει η εξίσωση (3.32) αντί της θεωρητικά ορθής εξίσωσης (3.41).

Η θεωρητικά ορθή εκτίμηση των ϕ_{ss} και ϕ_{vv} από τα ευθυγραμμισμένα χρονικά σήματα είναι:

$$\hat{\phi}_{ss}^{ij} = \frac{\mathrm{Re}\left\{\hat{\phi}_{x'_i x'_j}\right\} - \frac{1}{2}\left(\hat{\phi}_{x'_i x'_i} + \hat{\phi}_{x'_j x'_j}\right)\mathrm{Re}\left\{C_{v'_i v'_j}\right\}}{1 - \mathrm{Re}\left\{C_{v'_i v'_j}\right\}} \tag{3.43}$$

$$\hat{\phi}_{vv}^{ij} = \frac{\frac{1}{2}\left(\hat{\phi}_{x'_i x'_i} + \hat{\phi}_{x'_j x'_j}\right) - \mathrm{Re}\left\{\hat{\phi}_{x'_i x'_j}\right\}}{1 - \mathrm{Re}\left\{C_{v'_i v'_j}\right\}} \tag{3.44}$$

$$C_{v'_i v'_j} = C_{v_i v_j} e^{-\jmath\omega_k(\tau_j-\tau_i)} \tag{3.45}$$

Στην περίπτωση του διάχυτου πεδίου θορύβου (που είναι η περίπτωση που εξετάζεται στα [26, 25]), επειδή η συνάρτηση $C_{v_i v_j}$ είναι πραγματική, προκύπτει:

$$Re\{C_{v'_i v'_j}\} = C_{v_i v_j} \cos\left(\omega_k(\tau_j - \tau_i)\right) \tag{3.46}$$

Για τις χαμηλές συχνότητες ή/και μικρό $|\tau_j - \tau_i|$ θα ισχύει $C_{v_i v_j} > 0$ (βλ. εξίσωση (3.6)) και $\cos\left(\omega_k(\tau_j - \tau_i)\right) > 0$ επομένως:

$$\mathrm{Re}\{C_{v'_i v'_j}\} \leq C_{v_i v_j} = \mathrm{Re}\{C_{v_i v_j}\} \tag{3.47}$$

Συνεπώς, στην περιοχή χαμηλών συχνοτήτων (στην οποία συγκεντρώνεται το μεγαλύτερο μέρος της ενέργειας για σήματα φωνής) η χρήση της εξίσωσης (3.37) αντί

της εξίσωσης (3.43) οδηγεί σε επερεκτίμηση του φάσματος ισχύος του θορύβου. και άρα σε υποβέλτιση εκτίμηση των post-filters.

Στην πράξη, ωστόσο, η χρήση των υποβέλτιστων εκτιμήσεων (3.35), (3.37) οδηγεί σε καλύτερα αποτελέσματα. Το γεγονός αυτό ερμηνεύεται με βάση τις εξής παρατηρήσεις:

Έστω $\widetilde{C}_{v'_iv'_j}$ η συνάρτηση συνοχής των σημάτων θορύβου στα ευθυγραμμισμένα σήματα των μικροφώνων i, j που ισχύει στην πραγματικότητα. Η χρήση ενός μοντέλου $C_{v'_iv'_j}$ για την $\widetilde{C}_{v'_iv'_j}$ στην πράξη θα εμφανίζει ένα σφάλμα:

$$\mathrm{Re}\{C_{v'_iv'_j}\} = \mathrm{Re}\{\widetilde{C}_{v'_iv'_j}\} + \epsilon \tag{3.48}$$

Με βάση τις εξισώσεις (3.27) και (3.28), η εκτίμηση (3.43) γίνεται:

$$\begin{aligned}
\hat{\phi}_{ss}^{ij} &= \frac{\phi_{ss} + \mathrm{Re}\left\{\phi_{v'_iv'_j}\right\} - (\phi_{ss}+\phi_{vv})\left(\mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\} + \epsilon\right)}{1 - \mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\} - \epsilon} \\
\hat{\phi}_{ss}^{ij} &= \phi_{ss} + \frac{\phi_{vv}\mathrm{Re}\{\widetilde{C}_{v_iv_j}\} - \phi_{vv}\left(\mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\} + \epsilon\right)}{1 - \mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\} - \epsilon} \\
\hat{\phi}_{ss}^{ij} &= \phi_{ss} - \frac{\epsilon}{1 - \mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\} - \epsilon}\phi_{vv}
\end{aligned} \tag{3.49}$$

Επίσης η εξίσωση (3.44) γίνεται:

$$\begin{aligned}
\hat{\phi}_{vv}^{ij} &= \frac{\phi_{ss} + \phi_{vv} - \left(\phi_{ss} + \phi_{vv}\mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\}\right)}{1 - \mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\} - \epsilon} \\
\hat{\phi}_{vv}^{ij} &= \frac{1 - \mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\}}{1 - \mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\} - \epsilon}\phi_{vv} \\
\hat{\phi}_{vv}^{ij} &= \phi_{vv} + \frac{\epsilon}{1 - \mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\} - \epsilon}\phi_{vv}
\end{aligned} \tag{3.50}$$

Σύμφωνα με τις εξισώσεις (3.49) και (3.50), στις συχνότητες όπου $\epsilon > 0$, γίνεται υποεκτίμηση του φάσματος ισχύος της πηγής και υπερεκτίμηση του θορύβου, ενώ στις συχνότητες όπου $\epsilon < 0$ γίνεται υπερεκτίμηση του φάσματος ισχύος της πηγής και υποεκτίμηση του φάσματος ισχύος του θορύβου. Το απόλυτο σφάλμα εκτίμησης του φάσματος ισχύος της πηγής είναι:

$$\Delta\hat{\phi}_{ss} = \left|\frac{\epsilon}{1 - \mathrm{Re}\left\{\widetilde{C}_{v'_iv'_j}\right\} - \epsilon}\right|\phi_{vv}, \tag{3.51}$$

το οποίο δεν εξαρτάται μόνο από την απόλυτη τιμή του ϵ, αλλά και από το πλάτος του φάσματος ισχύος του θορύβου. Όταν $\epsilon > 0$, ανεξάρτητα από την απόλυτη τιμή του $\Delta\hat{\phi}_{ss}$, η υποεκτίμηση του φάσματος πηγής και η υπερεκτίμηση του φάσματος θορύβου οδηγούν σε post-filters τα οποία συμπιέζουν αποτελεσματικά το θόρυβο. Το μειονέκτημα είναι η παραμόρφωση του σήματος πηγής. Αντίθετα, όταν $\epsilon < 0$ παρεισφρύει στην εκτίμηση του φάσματος ισχύος της πηγής και ένα ποσοστό φάσματος ισχύος του θορύβου, το οποίο ποσοστό είναι μεγαλύτερο εκεί όπου το φάσμα ισχύος του θορύβου είναι μεγαλύτερο (εξίσωση (3.51)). Το γεγονός αυτό προκαλεί διέλευση περισσότερου θορύβου μέσα από το post-filter σε σχέση με την περίπτωση $\epsilon > 0$. Συνεπώς, μεγαλύτερο ρόλο στην αποτελεσματικότητα του post-filtering παίζει το πρόσημο και όχι η απόλυτη τιμή του ϵ. Για $\epsilon > 0$ το post-filtering είναι αποτελεσματικό, ενώ για $\epsilon < 0$ διέρχεται περισσότερος θόρυβος.

Για διάχυτα πεδία θορύβου, η χρήση της συνάρτησης συνοχής $C_{v_i v_j}$ αντί της $C_{v_i' v_j'}$ είναι μία υπερεκτίμηση του αληθινού $\widetilde{C}_{v_i' v_j'}$ στις χαμηλές συχνότητες (λόγω εξίσωσης (3.47)), στις οποίες συγκεντρώνεται και η μεγαλύτερη ενέργεια του σήματος, όταν πρόκειται για σήματα φωνής. Επομένως, για αυτές τις συχνότητες θα ισχύει συστηματικά $\epsilon > 0$ και άρα θα προκύπτει, όπως εξηγήθηκε, αποτελεσματική συμπίεση του θορύβου μετά το post-filtering.

Κατά τη χρήση της θεωρητικά ορθής συνάρτησης συνοχής $C_{v_i' v_j'}$, το πρόσημο του λάθους ϵ δεν θα είναι συστηματικά θετικό. Κατ΄ απόλυτη τιμή αναμένεται να είναι μικρότερο από το λάθος που παράγει η χρήση της $C_{v_i v_j}$, όμως θα υπάρχουν και περιπτώσεις στις οποίες $\epsilon < 0$ και τότε θα υπάρχει διέλευση περισσότερου θορύβου μέσα από το post-filter. Επομένως, στην πράξη λειτουργούν καλύτερα οι υποβέλτιστες εκτιμήσεις (3.35) και (3.37) που χρησιμοποιούν το $C_{v_i v_j}$. Η παραπάνω ανάλυση, αν και βασίστηκε στη συνάρτηση συνοχής για το διάχυτο πεδίο θορύβου, για το οποίο προκύπτει άμεσα η σχέση (3.47), ισχύει και σε άλλες περιπτώσεις. Για παράδειγμα, για ένα εντοπισμένο πεδίο θορύβου προκύπτει με βάση τις εξισώσεις (3.7) και (3.45):

$$C_{v_i' v_j'} = e^{-\jmath\omega_k(\tau_j - \tau_i + \tau_{v_i} - \tau_{v_j})} \tag{3.52}$$

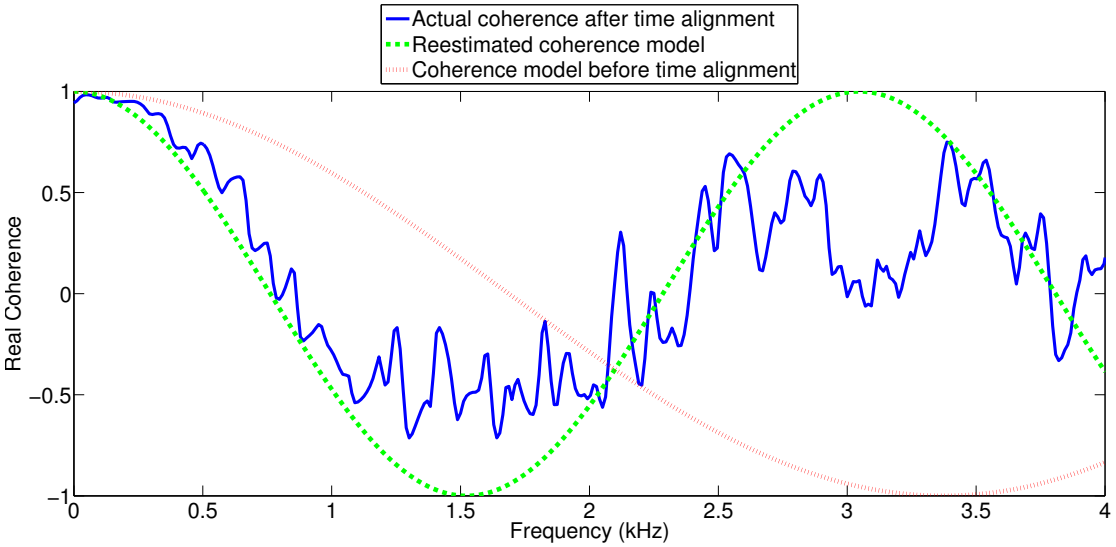
οπότε:

$$\mathrm{Re}\{C_{v_i' v_j'}\} = \cos\left(\omega_k(\tau_j - \tau_i + \tau_{v_i} - \tau_{v_j})\right) \tag{3.53}$$

Αν $|\tau_j - \tau_i + \tau_{v_i} - \tau_{v_j}| \geq |\tau_{v_j} - \tau_{v_i}|$ (γεωμετρικά αυτό σημαίνει η πηγή θορύβου και η πηγή επιθυμητού σήματος ομιλίας να βρίσκονται εκατέρωθεν της μεσοκαθέτου του ευθυγράμμου τμήματος που συνδέει τα μικρόφωνα i, j) τότε για χαμηλές συχνότητες θα ισχύει:

$$\mathrm{Re}\{C_{v_i' v_j'}\} \leq \cos\left(\omega_k(\tau_{v_j} - \tau_{v_i})\right) = \mathrm{Re}\{C_{v_i v_j}\} \tag{3.54}$$

Επομένως και σε αυτήν την περίπτωση η χρήση του $C_{v_i v_j}$ προκαλεί υπερεκτίμηση του φάσματος του θορύβου για τα ευθυγραμμισμένα σήματα. Ένα πραγματικό παράδειγμα που δείχνει αυτήν την περίπτωση φαίνεται στο Σχήμα 3.1. Με μία γραμμική



Σχήμα 3.1: Σύγκριση μεταξύ πραγματικής συνάρτησης συνοχής θορύβου μετά από ευθυγράμμιση των σημάτων, $C_{v'_i v'_j}$ (πράσινη διακεκομμένη) $C_{v_i v_j}$ (κόκκινη στικτή).

συστοιχία μικροφώνων με 8cm απόσταση ηχογραφήθηκε σήμα θορύβου προερχόμενο από εντοπισμένη πηγή (ραδιόφωνο) σε κατεύθυνση 135^o σε σχέση με τον άξονα της συστοιχίας. Για τα δύο μικρόφωνα της συστοιχίας που είναι πλησιέστερα προς την πηγή θορύβου έγινε ευθυγράμμιση των σημάτων για υποτιθέμενο σήμα πηγής που φτάνει στη συστοιχία από κατεύθυνση 45^o. Στη συνέχεια εκτιμήθηκε από τα ευθυγραμμισμένα αυτά σήματα θορύβου η συνάρτηση συνοχής, η οποία απεικονίζεται στο σχήμα με τη μπλε συνεχή καμπύλη. Στο σχήμα απεικονίζονται επίσης οι συναρτήσεις $C_{v_i v_j}$ (κόκκινη στικτή γραμμή) και $C_{v'_i v'_j}$ (πράσινη διακεκομμένη γραμμή). Εμφανώς, η $C_{v_i v_j}$ υπερεκτιμά για συχνότητες έως 2kHz την πραγματική συνάρτηση συνοχής του θορύβου, ενώ η $C_{v'_i v'_j}$ ακολουθεί με μεγαλύτερη ακρίβεια την πραγματική συνάρτηση συνοχής. Η χρήση της $C_{v'_i v'_j}$ υποεκτιμά ελαφρώς (αν και λιγότερο από ότι η $C_{v_i v_j}$ υπερεκτιμά) την πραγματική συνάρτηση συνοχής και θα οδηγήσει, για τους λόγους που εξηγήθηκαν, σε post-filter από το οποίο θα διέρχεται περισσότερος θόρυβος. Αντίθετα η χρήση της $C_{v_i v_j}$ αν και υποβέλτιση θα οδηγήσει πρακτικά σε μεγαλύτερη συμπίεση θορύβου[1].

[1]Για συχνότητες άνω των 2kHz, η $C_{v_i v_j}$ είναι αυτή που υποεκτιμά την πραγματική συνάρτηση συνοχής. Ωστόσο, στην περίπτωση υπό μελέτη, το μεγαλύτερο ποσοστό της ενέργειας του θορύβου, ο οποίος είναι σήμα φωνής παραγόμενο από ηχείο (ραδιόφωνο), είναι συγκεντρωμένο στις χαμηλές συχνότητες. Συνεπώς, για συχνότητες άνω των 2kHz το φάσμα ισχύος του θορύβου είναι ήδη μικρό και η διέλευση περισσότερου θορύβου από αυτήν την περιοχή συχνοτήτων δεν προκαλεί υποβάθμιση της ποιότητας του αποθορυβοποιημένου σήματος. Η συγκέντρωση της ενέργειας του θορύβου στις χαμηλές συχνότητες ισχύει και σε διάχυτα πεδία θορύβου. Οπότε για όλες τις περιπτώσεις που μελετώνται εδώ, ενδιαφέρει μόνο η χαμηλόσυχνη περιοχή.

3.6 Μελέτη της αποτελεσματικότητας πολυκαναλικής αποθορυβοποίησης σε διάφορες συνθήκες

Η επίδοση των μεθόδων πολυκαναλικής επεξεργασίας εξαρτάται από πολλούς παράγοντες, όπως τη γεωμετρία της συστοιχίας μικροφώνων που χρησιμοποιείται, τα χαρακτηριστικά του θορύβου που υπάρχει στο περιβάλλον, τη θέση του ομιλητή που παράγει το επιθυμητό σήμα φωνής, τα χαρακτηριστικά του χώρου στον οποίο διαδίδεται το σήμα φωνής κ.ά. Είναι ενδιαφέρον, συνεπώς, να μελετηθεί πειραματικά πώς κάποιοι από αυτούς τους παράγοντες επηρεάζουν την αποτελεσματικότητα της πολυκαναλικής επεξεργασίας. Ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη της αποτελεσματικότητας της πολυκαναλικής αποθορυβοποίησης σε σχέση με τη γεωμετρία της συστοιχίας μικροφώνων. Για το σκοπό αυτό, έγινε συλλογή μίας βάσης δεδομένων, η οποία περιέχει πολυκαναλικά σήματα για διάφορους συνδυασμούς των εξής παραμέτρων: γεωμετρία της συστοιχίας μικροφώνων, είδος του πεδίου θορύβου και θέση του ομιλητή. Χρησιμοποιήθηκε συστοιχία με μικρόφωνα MEMS (MicroElectroMechanical Systems), τα οποία είναι μία καινούρια τεχνολογία μικροφώνων πολύ μικρών διαστάσεων. Στα σήματα της βάσης έγινε πολυκαναλική αποθορυβοποίηση με το σύστημα που περιγράφεται στο [25] και εξήχθησαν συμπεράσματα για την αποτελεσματικότητα της πολυκαναλικής επεξεργασίας για κάθε συνδυασμό συνθηκών αξιολογώντας τη συμπίεση του θορύβου που επιτεύχθηκε στα αποθορυβοποιημένα σήματα.

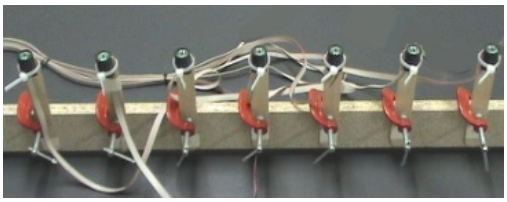
Στην Ενότητα 3.6.1 παρουσιάζεται η βάση δεδομένων που συνελέγη με τα μικρόφωνα MEMS. Στην Ενότητα 3.6.2 παρουσιάζεται συνοπτικά η δομή του πολυκαναλικού συστήματος που χρησιμοποιήθηκε για αποθορυβοποίηση. Στην Ενότητα 3.6.3 παρουσιάζονται τα πειραματικά αποτελέσματα. Τέλος, στην Ενότητα 3.6.3 παρουσιάζονται τα συμπεράσματα της μελέτης της αποτελεσματικότητας της πολυκαναλικής επεξεργασίας για τις διάφορες γεωμετρίες της συστοιχίας και τα είδη θορύβου.

3.6.1 Πολυκαναλική βάση δεδομένων MEMS

Προκειμένου να μελετηθεί η αποτελεσματικότητα διαφόρων γεωμετρικών διατάξεων της συστοιχίας μικροφώνων, έγινε συλλογή μίας πολυκαναλικής βάσης δεδομένων χρησιμοποιώντας μια συστοιχία από μικρόφωνα MEMS (Σχήμα 3.2).

Η τεχνολογία των μικροφώνων MEMS αναπτύχθηκε πρόσφατα. Πρόκειται για μικρόφωνα ιδιαίτερα μικρών διαστάσεων και χαμηλού κόστους. Το ιδιαίτερα μικρό μέγεθος των μικροφώνων καθιστά τη συστοιχία MEMS φορητή και την επαναδιάταξη των μικροφώνων σε διαφορετική γεωμετρία εύκολη. Λεπτομερείς τεχνικές προδιαγραφές για τα μικρόφωνα MEMS μπορούν να βρεθούν στο [37].

Η συλλογή δεδομένων έγινε σε ένα κλειστό δωμάτιο με γραφεία και υπολογιστές. Έγινε συλλογή δεδομένων με μία συστοιχία MEMS με 7 μικρόφωνα. Η



(α)

(β)

(γ)

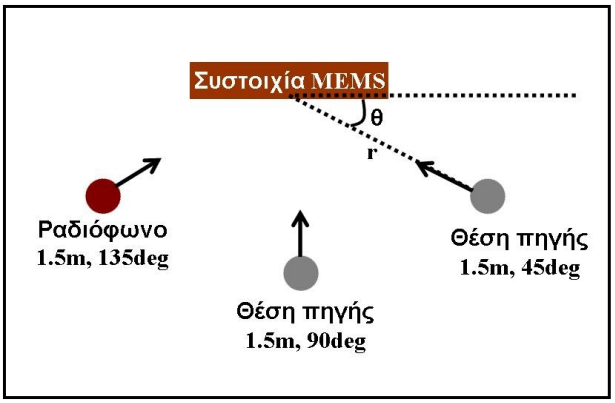
Σχήμα 3.2: Συστοιχία μικροφώνων MEMS: (α) Γραμμική διάταξη, (β) Εξαγωνική διάταξη και (γ) Μικρόφωνο MEMS.

τοποθέτηση των μικροφώνων έγινε σε γραμμική και εξαγωνική διάταξη (Σχήμα 3.2 (α) και (β), αντίστοιχα) πάνω σε ένα επίπεδο τραπέζι. Οι γραμμικές συστοιχίες χρησιμοποιούνται πολύ συχνά στην πράξη. Οι εξαγωνικές συστοιχίες παρουσιάζουν κάποια θεωρητικά πλεονεκτήματα [40], όπως βέλτιστη δειγματοληψία του χωρικού πεδίου [31]. Οι γεωμετρίες που χρησιμοποιήθηκαν είναι οι εξής: γραμμική με ομοιόμορφη απόσταση μεταξύ μικροφώνων $d = 4\text{cm}, 8\text{cm}, 12\text{cm}$ και εξαγωνική με ακτίνα $r = 8\text{cm}, 12\text{cm}, 16\text{cm}$.

Για κάθε γεωμετρική διάταξη και απόσταση μεταξύ μικροφώνων, έγινε συλλογή δεδομένων για δύο είδη πεδίων θορύβου που απαντώνται συχνά στην πράξη, το διάχυτο πεδίο θορύβου και το εντοπισμένο πεδίο θορύβου. Για την παραγωγή διάχυτου πεδίου θορύβου χρησιμοποιήθηκαν ανεμιστήρες και αερόθερμα εντός του δωματίου ηχογραφήσεων. Για την παραγωγή εντοπισμένου πεδίου θορύβου χρησιμοποιήθηκε ένα ηχείο, από το οποίο ακουγόταν ραδιοφωνική εκπομπή. Το ηχείο ήταν τοποθετημένο σε γωνία 135^o και απόσταση 1.5m σε σχέση με το κέντρο της συστοιχίας μικροφώνων.

Για κάθε συνδυασμό γεωμετρίας και πεδίου θορύβου έγινε συλλογή δεδομένων για δύο θέσεις ομιλητή: γωνία 45^o και γωνία 90^o σε απόσταση 1.5m σε σχέση με το κέντρο της συστοιχίας μικροφώνων. Ένας χάρτης του δωματίου με τις θέσεις ομιλητή και τη θέση του εντοπισμένου θορύβου απεικονίζεται στο Σχήμα 3.3.

Για κάθε συνδυασμό γεωμετρίας, πεδίου θορύβου και θέσης ομιλητή έγινε συλλογή δεδομένων για 6 ομιλητές, 3 άνδρες και 3 γυναίκες. Κάθε ομιλητής εκφωνούσε στεκόμενος σε κάθε θέση το ίδιο σύνολο από 30 σύντομες προστακτικές προτάσεις, οι οποίες θα μπορούσαν να θεωρηθούν εντολές προς ένα αυτοματοποιημένο σύστημα ελέγχου μίας κατοικίας, όπως το σύστημα υπό ανάπτυξη στα πλαίσια του FP-7



Σχήμα 3.3: Συλλογή δεδομένων με μικρόφωνα MEMS: χάρτης του δωματίου ηχογραφήσεων στον οποίο απεικονίζονται η θέση του ραδιοφώνου και οι δύο θέσεις ομιλητή για τις οποίες έγιναν ηχογραφήσεις (σημείωση: το ραδιόφωνο δεν είναι ενεργό κατά τις ηχογραφήσεις με διάχυτο πεδίο θορύβου).

Ευρωπαϊκού έργου Distant Speech Interaction for Robust Home Applications (DIRHA) [14]. Ο ομιλητής στεκόμενος όρθιος βρίσκεται περίπου στο ίδιο επίπεδο με τα μικρόφωνα της συστοιχίας MEMS, η οποία είναι τοποθετημένη σε επίπεδο τραπέζι.

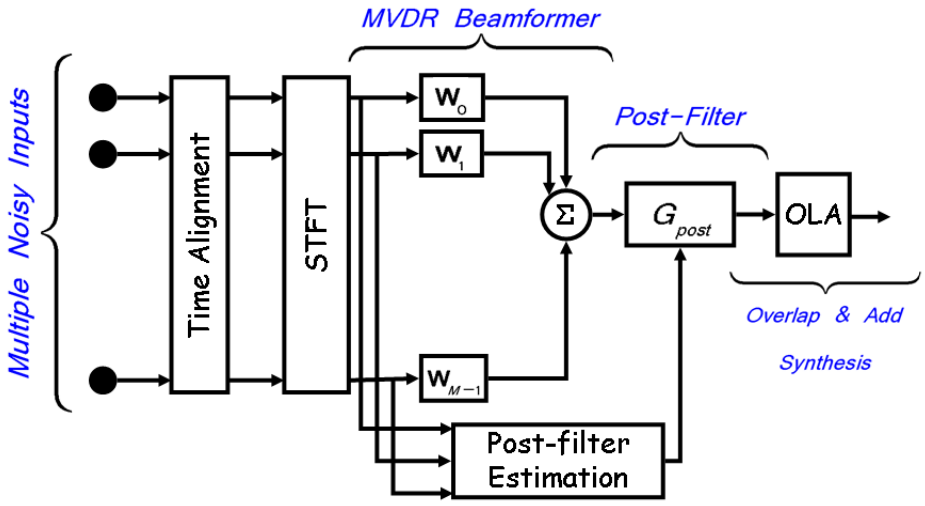
Παράλληλα με την ηχογράφηση των καναλιών της συστοιχίας MEMS, έγινε ηχογράφηση του επιθυμητού σήματος φωνής που εκφέρει ο ομιλητής από close-talk μικρόφωνο κεφαλής. Το σήμα που καταγράφεται από το close-talk μικρόφωνο θεωρείται ως το καθαρό σήμα ομιλίας που είναι επιθυμητό να εκτιμηθεί. Ο ρυθμός δειγματοληψίας όλων των σημάτων είναι 48kHz.

3.6.2 Πολυκαναλικό σύστημα αποθορυβοποίησης

Για την αποθορυβοποίηση των σημάτων της βάσης MEMS χρησιμοποιήθηκε το σύστημα πολυκαναλικής αποθορυβοποίησης που προτάθηκε στο [25]. Το σύστημα υλοποιεί τα MMSE, STSA και log-STSA post-filters. Η μέθοδος εκτίμησης των post-filters που προτείνεται στο [25] αναλύθηκε στην Ενότητα 3.4.3. Η βελτίωση που προτάθηκε στην Ενότητα 3.5 δε χρησιμοποιείται εξαιτίας του πρακτικού προβλήματος που αναλύθηκε στην ίδια ενότητα.

Υπάρχουν μερικά πρακτικά ζητήματα στην υλοποίηση του συστήματος αυτού. Τα πραγματικά μικρόφωνα εισάγουν σφάλματα φάσης και αυτοθόρυβο (self-noise). Για την αποφυγή του προβλήματος αυτού τα βάρη του MVDR beamformer $\mathbf{W}_{\text{MVDR}}$ υπολογίζονται με περιορισμό κέρδους λευκού θορύβου[2] (white noise gain con-

[2]Ο MVDR beamformer ενισχύει τον χωρικά ασυσχέτιστο θόρυβο. Για την αποφυγή αυτού του προβλήματος, πριν τον υπολογισμό του $\mathbf{C}_{vv}^{-1}$, στη διαγώνιο του πίνακα συνοχής $\mathbf{C}_{vv}$ του πεδίου θορύβου προστίθεται μία μικρή σταθερά ϵ.



Σχήμα 3.4: Σύστημα πολυκαναλικής αποθορυβοποίησης σημάτων [25].

straint) [11]. Ο MVDR beamformer δε συμπιέζει αποτελεσματικά το θόρυβο στις χαμηλές συχνότητες. Για το λόγο αυτό, κατά την εκτίμηση των post-filters, αντί του φάσματος ισχύος του θορύβου ϕ_{nn} στην έξοδο του MVDR beamformer χρησιμοποιείται η τροποποιημένη έκφραση [25]:

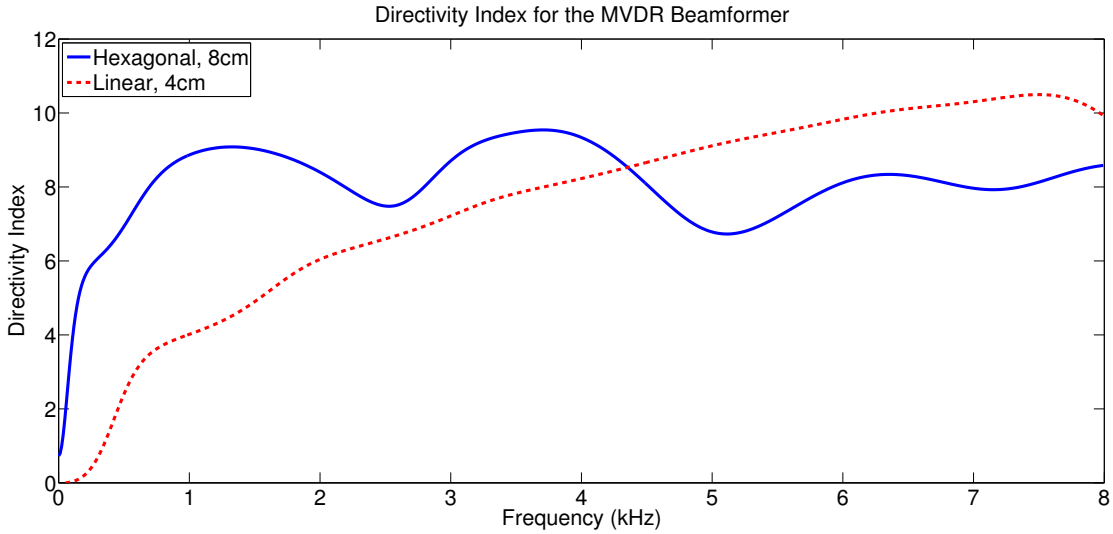
$$\phi_{nn} = \begin{cases} \phi_{vv}, \text{ για } \omega \leq \omega_1 \\ \phi_{nn}, \text{ για } \omega > \omega_1 \end{cases}, \tag{3.55}$$

όπου η παράμετρος ω_1 θέτει το όριο της περιοχής χαμηλών συχνοτήτων. Τέλος τα φάσματα ισχύος και τα ετεροφάσματα $\phi_{x_i x_i}$ και $\phi_{x_i x_j}$ εκτιμώνται χρησιμοποιώντας τη μέθοδο από το [1]:

$$\hat{\phi}_{x_i x_j}(k,\ell) = \alpha \hat{\phi}_{x_i x_j}(k,\ell-1) + (1-\alpha) x_i(k,\ell) x_j^*(k,\ell) \tag{3.56}$$

3.6.3 Πειραματικά αποτελέσματα

Για την εξαγωγή των αποθορυβοποιημένων σημάτων χρησιμοποιήθηκε το σύστημα που αναφέρθηκε στην Ενότητα 3.6.2. Η χρονική ευθυγράμμιση των σημάτων επετεύχθη υπολογίζοντας τις καθυστερήσεις διάδοσης του σήματος πηγής προς τα μικρόφωνα με βάση την εκ των προτέρων γνωστή θέση του ομιλητή. Για εφαρμογή STFT, τα σήματα χωρίστηκαν σε παράθυρα Hamming 1200 δειγμάτων (25ms) με επικάλυψη 900 δείγματα ($\approx$ 19ms). Τα βάρη του MVDR beamformer και η εκτίμηση των παραμέτρων των post-filters έγιναν θεωρώντας εκ των προτέρων γνωστή τη συνάρτηση συνοχής του θορύβου, η οποία για το διάχυτο πεδίο θορύβου δίνεται από



Σχήμα 3.5: Δείκτης κατευθυντικότητας για τον MVDR beamformer για εξαγωνική γεωμετρία ακτίνας 8cm και για γραμμική γεωμετρία 4cm.

την εξίσωση (3.6) και για το εντοπισμένο πεδίο θορύβου (ραδιόφωνο) από την εξίσωση (3.7) (για τον υπολογισμό των καθυστερήσεων διάδοσης τ_{v_i} χρησιμοποιήθηκε η γνώση της θέσης του ηχείου που παρήγαγε το ανεπιθύμητο σήμα).

Για την αξιολόγηση της ποιότητας του αποθορυβοποιημένου σήματος χρησιμοποιήθηκε η μετρική Segmental SNR (SSNR) [21]. Το SSNR προκύπτει ως το μέσο SNR των πλαισίων του σήματος και έχει βρεθεί ότι είναι πιο ενδεικτικό της υποκειμενικής αντίληψης που έχει ο άνθρωπος για την ποιότητα του σήματος από το ολικό SNR. Πλαίσια με SNR άνω των 35dB δεν συνεισφέρουν σημαντικά στην υποκειμενική αντίληψη για την ποιότητα του σήματος και αντικαθίστανται στο μέσο όρο από 35dB. Αντίστοιχα ορίζεται ένα κάτω όριο στο SNR των πλαισίων ώστε πλαίσια σιωπής να μην επηρεάσουν αρνητικά το μέσο όρο. Για τα πειράματα που παρουσιάζονται εδώ το κάτω όριο τέθηκε στα -15dB.

Τα αποτελέσματα των πειραμάτων για την εξαγωνική γεωμετρία παρουσιάζονται στον Πίνακα 3.1 και για τη γραμμική γεωμετρία στον Πίνακα 3.2. Τα αποτελέσματα δίνονται με τη μετρική SSNR Enhancement που υπολογίζεται ως η διαφορά μεταξύ του SSNR της εξόδου του post-filter και του μέσου SSNR των θορυβωδών πολυκαναλικών σημάτων εισόδου.

3.6.4 Συμπεράσματα

Στην περίπτωση του διάχυτου πεδίου θορύβου, η εξαγωνική γεωμετρία υπερέχει έναντι της γραμμικής. Συγκεκριμένα, για θέση ομιλητή στις 90^o, η χειρότερη επίδοση της εξαγωνικής γεωμετρίας (ακτίνα 16cm) υπερβαίνει τα 5dB SSNRE, ενώ η καλύτερη

επίδοση της γραμμικής γεωμετρίας (απόσταση μικροφώνων 4cm) προκύπτει περίπου 4.9dB SSNRE. Η υπεροχή της εξαγωνικής γεωμετρίας για το διάχυτο πεδίο θορύβου μπορεί να εξηγηθεί θεωρητικά από το γεγονός ότι η γραμμική γεωμετρία έχει αξονική συμμετρία με αποτέλεσμα να αδυνατεί να διαχωρίσει σήματα που προέρχονται από τον ίδιο κώνο κατευθύνσεων (σήματα που διαδίδονται από οποιαδήποτε από αυτές τις κατευθύνσεις καταλήγουν στα μικρόφωνα με τις ίδιες καθυστερήσεις τ_m) [40]. Σε ένα πεδίο διάχυτου θορύβου, σήματα θορύβου καταφθάνουν στα μικρόφωνα από όλες τις κατευθύνσεις, συνεπώς, η μεγαλύτερη διακριτική ικανότητα της εξαγωνικής γεωμετρίας οδηγεί σε αυξημένη συμπίεση θορύβου. Μία μετρική της ικανότητας ενός beamformer να συμπιέσει διάχυτο θόρυβο είναι ο δείκτης κατευθυντικότητας (directivity index), ο οποίος ορίζεται ως [4]:

$$\mathrm{DI}(\omega) = 10 \log \left(\frac{\mathbf{W}^H \mathbf{D}}{\mathbf{W}^H \mathbf{C}_{vv}^{\mathrm{dif}} \mathbf{W}} \right) \tag{3.57}$$

Στο Σχήμα 3.5 παρουσιάζονται οι δείκτες κατευθυντικότητας για την εξαγωνική γεωμετρία ακτίνας 8cm και τη γραμμική γεωμετρία με απόσταση 4cm μεταξύ μικροφώνων που έδωσαν τα καλύτερα αποτελέσματα για το διάχυτο πεδίο θορύβου.

Για συχνότητες ως 4kHz (στις οποίες είναι συγκεντρωμένο το μεγαλύτερο ποσοστό της ενέργειας σημάτων φωνής) παρατηρείται μεγαλύτερη ικανότητα συμπίεσης του διάχυτου πεδίου θορύβου για την εξαγωνική γεωμετρία.

Για το εντοπισμένο πεδίο θορύβου, παρατηρείται ως προς την καλύτερη επίδοση που επιτεύχθηκε υπεροχή της γραμμικής γεωμετρίας με 6dB SSNRE για απόσταση μικροφώνων 4cm για τη θέση ομιλητή στις 45^o. Όμως, για τις εξαγωνικές γεωμετρίες με ακτίνα 12cm και 16cm παρατηρείται παρόμοια ή ελαφρά βελτιωμένη επίδοση σε σχέση με τις γραμμικές γεωμετρίες με απόσταση μικροφώνων 8cm και 12cm. Δηλαδή η εξαγωνική γεωμετρία πετυχαίνει παρόμοια επίδοση με πιο αραιή δειγματοληψία του ακουστικού πεδίου στο χώρο.

Και για τα δύο πεδία θορύβου, για δεδομένη γεωμετρία, παρατηρείται αύξηση της επίδοσης, καθώς μειώνεται η απόσταση μεταξύ μικροφώνων. Η μείωση της απόστασης των μικροφώνων πυκνώνει τη δειγματοληψία του ακουστικού πεδίου (acoustic field) οδηγώντας σε αυξημένες επιδόσεις. Σύμφωνα με το θεώρημα χωρικής δειγματοληψίας που παρουσιάζεται στο [40] η μείωση της απόστασης των μικροφώνων αυξάνει τη μέγιστη συχνότητα για την οποία είναι δυνατός ο χωρικός διαχωρισμός σημάτων.

Συμπερασματικά, η εξαγωνική γεωμετρία παρουσιάζει βελτιωμένη επίδοση σε σχέση με τη γραμμική με αραιότερη δειγματοληψία του ακουστικού πεδίου. Το συμπέρασμα αυτό συνάδει με θεωρητικά αποτελέσματα για τη βελτιστότητα του εξαγωνικού πλέγματος για δειγματοληψία ακουστικών πεδίων [31]. Για δεδομένη γεωμετρία η μείωση της απόστασης των μικροφώνων ευνοεί την πολυκαναλική αποθορυβοποίηση, καθώς πυκνώνει τη δειγματοληψία του ακουστικού πεδίου.

Εξαγωνική Γεωμετρία				
Απόσταση Μικροφώνων (cm)	Πεδίο Θορύβου	Θέση Ομιλητή (m, deg)	Post-filter	SSNR Enhancement (dB)
8	Διάχυτο	(1.5, 90)	MMSE	+ 7.4902
			STSA	+ 7.2042
			log-STSA	+ 7.3580
		(1.5, 45)	MMSE	+ 4.1258
			STSA	+ 4.0012
			log-STSA	+ 4.0698
	Εντοπισμένο	(1.5, 90)	MMSE	+ 3.3573
			STSA	+ 2.8262
			log-STSA	+ 2.9850
		(1.5, 45)	MMSE	+ 4.5840
			STSA	+ 3.4388
			log-STSA	+ 3.7515
12	Διάχυτο	(1.5, 90)	MMSE	+ 6.9925
			STSA	+ 6.7263
			log-STSA	+ 6.8744
		(1.5, 45)	MMSE	+ 3.7492
			STSA	+ 3.6326
			log-STSA	+ 3.6995
	Εντοπισμένο	(1.5, 90)	MMSE	+ 3.1764
			STSA	+ 2.8208
			log-STSA	+ 2.9598
		(1.5, 45)	MMSE	+ 3.7659
			STSA	+ 3.1373
			log-STSA	+ 3.3622
16	Διάχυτο	(1.5, 90)	MMSE	+ 5.6580
			STSA	+ 5.4100
			log-STSA	+ 5.5583
		(1.5, 45)	MMSE	+ 3.1164
			STSA	+ 3.0062
			log-STSA	+ 3.0701
	Εντοπισμένο	(1.5, 90)	MMSE	+ 2.7294
			STSA	+ 2.3460
			log-STSA	+ 2.5045
		(1.5, 45)	MMSE	+ 3.9384
			STSA	+ 3.0375
			log-STSA	+ 3.2937

Πίνακας 3.1: Αποτελέσματα πολυκαναλικής αποθορυβοποίησης για την εξαγωνική γεωμετρία συστοιχίας μικροφώνων.

Γραμμική Γεωμετρία				
Απόσταση Μικροφώνων (cm)	Πεδίο Θορύβου	Θέση Ομιλητή (m, deg)	Post-filter	SSNR Enhancement (dB)
4	Διάχυτο	(1.5, 90)	MMSE	+ 4.9047
			STSA	+ 4.8500
			log-STSA	+ 4.8953
		(1.5, 45)	MMSE	+ 3.4843
			STSA	+ 3.4438
			log-STSA	+ 3.4765
	Εντοπισμένο	(1.5, 90)	MMSE	+ 1.1961
			STSA	+ 1.1754
			log-STSA	+ 1.1867
		(1.5, 45)	MMSE	+ 6.6108
			STSA	+ 6.2796
			log-STSA	+ 6.4529
8	Διάχυτο	(1.5, 90)	MMSE	+ 4.5229
			STSA	+ 4.4562
			log-STSA	+ 4.5076
		(1.5, 45)	MMSE	+ 2.8745
			STSA	+ 2.8003
			log-STSA	+ 2.8477
	Εντοπισμένο	(1.5, 90)	MMSE	+ 1.0018
			STSA	+ 0.9810
			log-STSA	+ 0.9938
		(1.5, 45)	MMSE	+ 3.8644
			STSA	+ 3.4405
			log-STSA	+ 3.6067
12	Διάχυτο	(1.5, 90)	MMSE	+ 4.1857
			STSA	+ 4.1158
			log-STSA	+ 4.1730
		(1.5, 45)	MMSE	+ 1.9936
			STSA	+ 1.9469
			log-STSA	+ 1.9835
	Εντοπισμένο	(1.5, 90)	MMSE	+ 0.8820
			STSA	+ 0.8302
			log-STSA	+ 0.8562
		(1.5, 45)	MMSE	+ 2.8489
			STSA	+ 2.4679
			log-STSA	+ 2.6151

Πίνακας 3.2: Αποτελέσματα πολυκαναλικής αποθορυβοποίησης για τη γραμμική γεωμετρία συστοιχίας μικροφώνων.

Κεφάλαιο 4

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Μελετήθηκαν τα προβλήματα εντοπισμού θέσης πηγής και αποθορυβοποίησης σημάτων ομιλίας από πολυκαναλικά δεδομένα.

Για το πρόβλημα εντοπισμού θέσης πηγής, παρουσιάστηκε μία νέα μέθοδος ελαχίστων τετραγώνων που οδηγεί σε λύση κλειστής μορφής και είναι συνεπώς υπολογιστικά αποδοτική και κατάλληλη για εφαρμογές πραγματικού χρόνου. Η μέθοδος βασίζεται σε εκτίμηση διαφοράς χρόνου άφιξης (Time Difference of Arrival) μεταξύ ζευγών μικροφώνων. Για την εκτίμηση TDOA χρησιμοποιήθηκε μία βελτιωμένη έκδοση της μεθόδου μέτρου συνοχής φάσης ετεροφάσματος (Crosspower-spectrum Phase Coherence Measure, CSP-CM). Οι βελτιώσεις στη μέθοδο CSP-CM αποδείχτηκε πειραματικά σε προσομοιώσεις ότι αυξάνουν την ακρίβεια της κλασικής μεθόδου CSP-CM. Η προτεινόμενη μέθοδος εντοπισμού θέσης αποδείχτηκε με πειραματισμό σε βάσεις δεδομένων με πολυκαναλικές ηχογραφήσεις σε πραγματικές συνθήκες ότι είναι ακριβής, με μέσο σφάλμα εντοπισμού θέσης μικρότερο των 50cm, ενώ, ταυτόχρονα, ο χρόνος εκτέλεσης του αλγορίθμου δεν υπερβαίνει το 0.75RT καθιστώντας τη μέθοδο κατάλληλη για εφαρμογές πραγματικού χρόνου. Τέλος, έγινε χρήση της μεθόδου εντοπισμού ως πρώτο στάδιο (front-end) πολυκαναλικού συστήματος αποθορυβοποίησης και προέκυψε πειραματικά ότι η ποιότητα του αποθορυβοποιημένου σήματος είναι εφάμιλλη με αυτήν που προκύπτει όταν χρησιμοποιείται γνώση της πραγματικής θέσης πηγής, συνεπώς η προτεινόμενη μέθοδος εντοπισμού θέσης πηγής είναι αρκετά ακριβής για χρήση σε πολυκαναλικά συστήματα αποθορυβοποίησης.

Για το πρόβλημα της πολυκαναλικής αποθορυβοποίησης σημάτων, έγινε μελέτη της αποτελεσματικότητας πολυκαναλικού συστήματος αποθορυβοποίησης με beamforming και post-filtering για γραμμικές και εξαγωνικές διατάξεις της συστοιχίας μικροφώνων και διάφορα είδη θορύβου. Η μελέτη έγινε χρησιμοποιώντας πολυκα-

ναλικά δεδομένα που συλλέχθηκαν με μία συστοιχία μικροφώνων MEMS, τα οποία αποτελούν μια καινούρια τεχνολογία φορητών αισθητήρων πολύ μικρών διαστάσεων. Η συλλογή των πολυκαναλικών δεδομένων έγινε σε πραγματικές συνθήκες, σε περιβάλλοντα διάχυτου και εντοπισμένου θορύβου για εξαγωνική και γραμμική διάταξη της συστοιχίας με διάφορες αποστάσεις μεταξύ μικροφώνων. Πειραματικά, προέκυψε ότι η εξαγωνική συστοιχία υπερέχει της γραμμικής σε δυνατότητα συμπίεσης του θορύβου. Το αποτέλεσμα αυτό συνάδει με υπάρχοντα θεωρητικά αποτελέσματα για τη βελτιστότητα του εξαγωνικού πλέγματος για χωρική δειγματοληψία ακουστικών πεδίων [21]. Πέραν της μελέτης για την αποτελεσματικότητα της γεωμετρίας της συστοιχίας μικροφώνων, προτάθηκε θεωρητική βελτίωση στο υπάρχον σύστημα πολυκαναλικής επεξεργασίας σημάτων, η οποία, ωστόσο, στην πράξη, δε βελτιώνει το αποτέλεσμα της αποθορυβοποίησης για λόγους που εξηγήθηκαν θεωρητικά.

Μελλοντικά, η έρευνα μπορεί να επεκταθεί στις εξής κατευθύνσεις της περιοχής πολυκαναλικής επεξεργασίας ομιλίας μακρινού πεδίου (far-field multichannel speech processing):

- Συνδυασμός εντοπισμού θέσης με σύστημα αναγνώρισης ακουστικών γεγονότων (acoustic event detection and classification), ώστε να μελετηθεί η δυνατότητα αξιοποίησης της πληροφορίας της θέσης ενός ακουστικού γεγονότος για την αναγνώρισή του.

- Συνδυασμός πολυκαναλικής αποθορυβοποίησης με σύστημα ανίχνευσης και αναγνώρισης ακουστικών γεγονότων, ώστε να μελετηθεί πιθανή βελτίωση που μπορεί να επιτευχθεί σε θορυβώδη περιβάλλοντα.

- Συνδυασμός πολυκαναλικής αποθορυβοποίησης με σύστημα αυτόματης αναγνώρισης ομιλίας (Automatic Speech Recognition, ASR), ώστε να μελετηθεί η βελτίωση στην αναγνώριση που μπορεί να επιτευχθεί.

Βιβλιογραφία

[1] J. Allen, D. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 912–915, 1977.

[2] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2002, pp. 209–213.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer, 2008, vol. 1.

[4] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 2, pp. 19–38.

[5] M. Brandstein, J. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer Speech and Language*, vol. 9, no. 2, pp. 153–169, 1995.

[6] ——, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech and Audio Process.*, vol. 5, no. 1, pp. 45–50, 1997.

[7] A. Brutti, "Speech Acoustic Scene Analysis and Interpretation: Data Collections," FBK. [Online]. Available: http://shine.fbk.eu/people/brutti/database

[8] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *Proc. Interspeech*, 2005, pp. 2337–2340.

[9] M. Chan, E. Campo, D. Estève, and J.-Y. Fourniols, "Smart homes – current features and future perspectives," *Maturitas*, vol. 64, no. 2, pp. 90–97, 2009.

[10] Y. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, 1994.

[11] H. Cox, R. M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans. Speech and Audio Process.*, vol. 34, no. 3, pp. 393–398, 1986.

[12] J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[13] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer, 2001.

[14] "The DIRHA (Distance-speech Interaction for Robust Home Applications) FP-7 EU project." [Online]. Available: http://dirha.fbk.eu/

[15] S. Doclo, "Multi-microphone noise reduction and dereverberation techniques for speech applications," Ph.D. dissertation, Katholieke Universiteit Leuven, 2003.

[16] S. Doclo and M. Moonen, "Design of far-field and near-field broadband beamformers using eigenfilters," *Speech Communication*, vol. 83, pp. 2641–2672, 2003.

[17] G. W. Elko, "Spatial coherence function for differential microphones in isotropic noise fields," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 4, pp. 61–85.

[18] Y. Ephraim and D. Mallah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[19] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech and Sig. Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[20] M. Gillette and H. Silverman, "A linear closed-form algorithm for source localization from time-differences of arrival," *IEEE Signal Processing Letters*, vol. 15, pp. 1–4, 2008.

[21] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 1998, pp. 2819–2822.

[22] Y. Huang, J. Benesty, G. Elko, and R. Mersereau, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 8, pp. 943–956, 2001.

[23] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Prentice Hall, 1993.

[24] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[25] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech communication*, vol. 49, no. 7, pp. 657–666, 2007.

[26] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 709–716, 2003.

[27] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. ICASSP*, 1997.

[28] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. ICASSP*, 1994.

[29] ——, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech and Audio Process.*, vol. 5, no. 3, pp. 288–292, 1997.

[30] M. Omologo, P. Svaizer, A. Brutti, and L. Cristoforetti, "Speaker localization in CHIL lectures: Evaluation criteria and results," in *Machine Learning for Multimodal Interaction*, S. Renals and S. Bengio, Eds. Springer, 2006.

[31] D. P. Petersen and D. Middleton, "Sampling and reconstruction of wavenumber-limited functions in n-dimensional Euclidean spaces," *Information and control*, vol. 5, no. 4, pp. 279–323, 1962.

[32] I. Rodomagoulakis, P. Giannoulis, Z.-I. Skordilis, P. Maragos, and G. Potamianos, "Experiments on far-field multichannel speech processing in smart homes," in *Proc. Int. Conf. Digital Signal Processing*, Santorini, July 2013.

[33] H. Schau and A. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 8, pp. 1223–1225, 1987.

[34] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[35] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 3, pp. 39–60.

[36] J. Smith and J. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 12, pp. 1661–1669, 1987.

[37] *MEMS audio sensor omnidirectional digital microphone*, MP34DT01, STMicroelectronics, 2013. [Online]. Available: http://www.st.com/st-web-ui/static/active/en/resource/technical/document/datasheet/DM00039779.pdf

[38] T. Sullivan, "CMU microphone array database," 1996. [Online]. Available: http://www.speech.cs.cmu.edu/databases/micarray

[39] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. ICASSP*, 1997.

[40] H. L. Van Trees, *Optimum Array Processing.* Wiley, 2002.

[41] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, pp. 4–24, 1988.

[42] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. ICASSP*, 1988.