

[Εξηγήστε περιεκτικά και επαρκώς την εργασία σας. Επιτρέπεται προαιρετικά η συνεργασία εντός ομάδων των 2 ατόμων. Κάθε ομάδα 2 ατόμων υποβάλλει μια κοινή αναφορά που αντιπροσωπεύει μόνο την προσωπική εργασία των μελών της. Αν χρησιμοποιήσετε κάποια άλλη πηγή εκτός του βιβλίου και του εκπαιδευτικού υλικού του μαθήματος, πρέπει να το αναφέρετε. Η παράδοση της αναφοράς και του κώδικα της εργασίας θα γίνει ηλεκτρονικά στο mycourses.ntua.gr και επιπλέον η αναφορά της εργασίας θα παραδίδεται τυπωμένη και προσωπικά στην γραμματεία του εργαστηρίου Ρομποτικής (2.1.12, παλαιό ΚΤ.Ηλεκ.), ώρες 09.30-14.00].

Θέμα: Συστοιχίες Μικροφώνων (Microphone Arrays) και Πολυκαναλική Επεξεργασία Σημάτων (Multichannel Signal Processing)

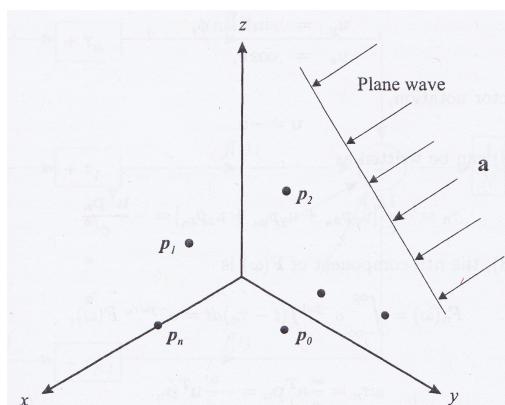
Μέρος 1. Συστοιχίες Μικροφώνων και Χωρικό Φιλτράρισμα (Spatial Filtering)

Η χρήση συστοιχιών μικροφώνων για καταγραφή και επεξεργασία ακουστικών σημάτων γίνεται όλο και πιο διαδεδομένη. Το πλεονέκτημα που παρουσιάζει η χρήση μικροφώνων κατανεμημένων στο χώρο είναι η δυνατότητα καταγραφής και αξιοποίησης όχι μόνο των χρονικών (temporal), αλλά και των χωρικών χαρακτηριστικών (spatial characteristics) των ακουστικών σημάτων. Τα χωρικά χαρακτηριστικά αυτά μπορούν να αξιοποιηθούν σε εφαρμογές όπως εύρεση κατεύθυνσης και εντοπισμός θέσης ακουστικής πηγής (direction-finding και source localization), αποθορυβοποίηση σημάτων ομιλίας (speech enhancement) κ.ά. Η εφαρμογή που θα μελετηθεί στην παρούσα άσκηση είναι το **speech enhancement**. Με τη χρήση συστοιχιών μπορεί να γίνει χωρικό φιλτράρισμα των ακουστικών σημάτων προκειμένου να ενισχυθούν ή να απορριφθούν σήματα που καταφθάνουν στη συστοιχία από συγκεκριμένη κατεύθυνση. Αυτό επιτυγχάνεται με κατάλληλο συνδυασμό των σημάτων που καταγράφονται από τα **διάφορα μικρόφωνα** ώστε το επιιμψητό σήμα που καταφθάνει από συγκεκριμένη κατεύθυνση να ενισχυθεί με ενισχυτική συμβολή, ενώ θόρυβος από τις υπόλοιπες κατευθύνσεις να εξασθενηθεί με αποσβεστική συμβολή.

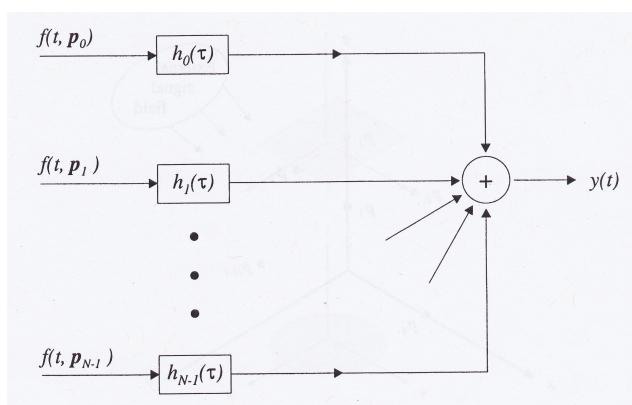
1.1 Beamforming

Έστω μια συστοιχία N μικροφώνων, τα οποία βρίσκονται στα σημεία του χώρου \mathbf{p}_n , $n = 0, 1, \dots, N-1$ (Σχήμα 1). Η συστοιχία αυτή δειγματοληπτεί το ακουστικό πεδίο (acoustic field) καταγράφοντας το σύνολο σημάτων

$$\mathbf{f}(t, \mathbf{p}) = [f(t, \mathbf{p}_0), f(t, \mathbf{p}_1), \dots, f(t, \mathbf{p}_{N-1})]^T. \quad (1)$$



Σχήμα 1: Συστοιχία μικροφώνων



Σχήμα 2: Πολυκαναλική επεξεργασία σημάτων

Κάθε καταγεγραμμένο σήμα φιλτράρεται από ένα γραμμικό, χρονικά αναλλοίωτο φίλτρο με χρονική απόχριση $h_n(t)$ και στη συνέχεια τα σήματα αθροίζονται (Σχήμα 2) δίνοντας την τελική έξοδο

$$y(t) = \sum_{n=0}^{N-1} \int_{-\infty}^{\infty} h_n(t-\tau) f(\tau, \mathbf{p}_n) d\tau = \int_{-\infty}^{\infty} \mathbf{h}^T(t-\tau) \mathbf{f}(\tau, \mathbf{p}_n) d\tau, \quad (2)$$

όπου

$$\mathbf{h}(\tau) = [h_0(\tau), h_1(\tau), \dots, h_{N-1}(\tau)]^T. \quad (3)$$

Στο πεδίο συχνότητας επομένως είναι¹:

$$Y(\omega) = \mathbf{H}^T(\omega) \mathbf{F}(\omega), \quad \mathbf{H}(\omega) = [H_0(\omega), \dots, H_{N-1}(\omega)]^T, \quad \mathbf{F}(\omega) = [F_0(\omega), \dots, F_{N-1}(\omega)]^T, \quad (4)$$

όπου $H_n(\omega) = \mathcal{F}\{h_n(t)\}$ και $F_n(\omega) = \mathcal{F}\{f(t, \mathbf{p}_n)\}$.

Η διαδικασία φιλτραρίσματος και συνδυασμού των σημάτων από τα πολλαπλά κανάλια για την παραγωγή της εξόδου $y(t)$ είναι γνωστή και ως beamforming, ενώ το σύστημα των N φίλτρων και του αθροιστή καλείται beamformer (Σχήμα 2). Για δεδομένη γεωμετρία της συστοιχίας μικροφώνων, η επιλογή των φίλτρων $H_n(\omega)$ καθορίζει τα χαρακτηριστικά του χωρικού φιλτραρίσματος που επιτυγχάνεται. Οπότε το πρόβλημα σχεδιασμού ενός beamformer, ώστε το χωρικό φίλτρο που προκύπτει να έχει συγκεκριμένα χαρακτηριστικά, έγκειται στην επιλογή των φίλτρων $H_n(\omega)$.

1.2 Beam pattern

Για την περιγραφή της χωρικής απόχρισης ενός beamformer χρησιμοποιείται το beam pattern, το οποίο είναι το χωρικό ανάλογο της απόχρισης συχνότητας ενός χρονικού φίλτρου. Ας υεωρήσουμε ένα επίπεδο ηχητικό κύμα² που φτάνει στη συστοιχία από την κατεύθυνση του μοναδιαίου διανύσματος \mathbf{a} (Σχήμα 1). Κάθε μικρόφωνο καταγράφει το ηχητικό σήμα με μία χρονική μετατόπιση σε σχέση με τα υπόλοιπα μικρόφωνα, λόγω της καθυστέρησης διάδοσης του ηχητικού σήματος. Αν $f(t)$ είναι το ηχητικό σήμα που καταγράφεται στην αρχή των αξόνων, τότε

$$\mathbf{f}(t, \mathbf{p}) = [f(t-\tau_0), f(t-\tau_1), \dots, f(t-\tau_{N-1})]^T, \quad (5)$$

όπου

$$\tau_n = \frac{\mathbf{a}^T \mathbf{p}_n}{c} \quad (6)$$

και c η ταχύτητα του ήχου (περίπου $340 \frac{\text{m}}{\text{s}}$ στον αέρα). Οπότε στο πεδίο συχνότητας είναι:

$$\mathbf{F}(\omega, \mathbf{a}) = \mathbf{d}(\mathbf{k}) F(\omega), \quad \mathbf{d}(\mathbf{k}) = [e^{-j\mathbf{k}^T \mathbf{p}_0}, e^{-j\mathbf{k}^T \mathbf{p}_1}, \dots, e^{-j\mathbf{k}^T \mathbf{p}_{N-1}}]^T, \quad (7)$$

όπου $F(\omega) = \mathcal{F}\{f(t)\}$ και $\mathbf{k} = \frac{\omega}{c} \mathbf{a}$. Το διάνυσμα \mathbf{d} περιέχει όλη την πληροφορία των χωρικών χαρακτηριστικών της συστοιχίας και ονομάζεται array manifold vector. Οπότε η εξίσωση (4) γίνεται:

$$Y(\omega, \mathbf{a}) = \mathbf{H}^T(\omega) \mathbf{F}(\omega, \mathbf{a}) = \mathbf{H}^T(\omega) \mathbf{d}(\mathbf{k}) F(\omega) \quad (8)$$

To beam pattern του beamformer είναι:

$$B(\omega, \mathbf{a}) = \mathbf{H}^T(\omega) \mathbf{d}(\mathbf{k})|_{\mathbf{k}=\frac{\omega}{c} \mathbf{a}} \quad (9)$$

To $B(\omega, \mathbf{a})$ περιγράφει πλήρως τη χωρο-χρονική επεξεργασία που γίνεται από τον beamformer. Όταν το ακουστικό πεδίο (το οποίο είναι η είσοδος του συστήματος συστοιχία-beamformer) είναι ένα επίπεδο «μονοχρωματικό» κύμα $f_{\text{eigen}}(t, \mathbf{p}) = \exp[j(\omega t - \mathbf{k}^T \mathbf{p})]$ συχνότητας ω που οδεύει κατά την κατεύθυνση \mathbf{a} , τότε η έξοδος του beamformer είναι $B(\omega, \mathbf{a}) e^{j\omega t} = |B(\omega, \mathbf{a})| e^{j\omega t + \angle B(\omega, \mathbf{a})}$, δηλαδή ένα μιγαδικό εκθετικό ίδιας χρονικής συχνότητας με το κύμα εισόδου, του οποίου το πλάτος έχει γίνει $|B(\omega, \mathbf{a})|$ και η φάση έχει μετατοπιστεί κατά $\angle B(\omega, \mathbf{a})$. Οπότε το beam pattern είναι το χωρο-χρονικό ανάλογο της απόχρισης συχνότητας των χρονικών φίλτρων. Η διαφορά είναι

¹Στην εργαστηριακή αυτή άσκηση ο συμβολισμός Fourier μετασχηματισμών σημάτων συνεχούς χρόνου χρησιμοποιεί ως σύμβολο συχνότητας το ω όπως και στο μάθημα Σήματα και Συστήματα.

²Τα κύματα που παράγονται από σημειωσές πηγές είναι σφαρικά, ωστόσο αν η πηγή βρίσκεται μακριά από τη συστοιχία (far-field assumption) τότε το μέτωπο του κύματος μπορεί να θεωρηθεί ότι έχει γίνει προσεγγιστικά επίπεδο, δηλαδή φτάσει στη συστοιχία.

ότι η απόκριση εξαρτάται τώρα και από την κατεύθυνση άφιξης \mathbf{a} , επομένως το φιλτράρισμα έχει επεκταθεί και στη διάσταση του χώρου. Τα επίπεδα «μονοχρωματικά» κύματα $f_{\text{eigen}}(t, \mathbf{p})$ είναι ιδιοσυναρτήσεις του beamformer, κατ' αναλογίαν με τα μιγαδικά εκθετικά για τα χρονικά φίλτρα. Ο κυματαριθμός $\|\mathbf{k}\| = \frac{\omega}{c} = \frac{2\pi}{\lambda}$, όπου λ το μήκος κύματος, μπορεί να ερμηνευθεί ως χωρική συχνότητα του κύματος.

1.2 Delay-and-sum beamformer και array steering

Ένας πολύ απλός beamformer είναι ο **delay-and-sum beamformer** ή **conventional beamformer**. Έστω ότι υπάρχει ένα επιμυητό σήμα που καταφθάνει στη συστοιχία από την κατεύθυνση του μοναδιαίου διανύσματος \mathbf{a}_s . Για το delay-and-sum beamforming, σε κάθε σήμα $f(t, \mathbf{p}_n)$ γίνεται χρονική ολίσθηση ώστε το σήμα που έρχεται από την κατεύθυνση \mathbf{a}_s να ευθυγραμμιστεί χρονικά σε όλα τα κανάλια πριν από την άμθοιση (Σχήμα 3). Για να επιτευχθεί αυτό επιλέγεται

$$\mathbf{H}_{\text{DS}}^T(\omega) = \frac{1}{N} \mathbf{d}^H(\mathbf{k}_s), \quad \mathbf{k}_s = \frac{\omega}{c} \mathbf{a}_s, \quad (10)$$

όπου το $\mathbf{d}^H(\mathbf{k}_s)$ συμβολίζει τον ερμητιανό ανάστροφο του $\mathbf{d}^H(\mathbf{k}_s)$, με αποτέλεσμα

$$B_{\text{DS}}(\omega, \mathbf{a}) = \frac{1}{N} \mathbf{d}^H(\mathbf{k}_s) \mathbf{d}(\mathbf{k}). \quad (11)$$

Συνεπώς $B(\omega, \mathbf{a}_s) = 1$, δηλαδή το επιμυητό σήμα διέρχεται αναλλοίωτο από τον beamformer, ενώ $|B(\omega, \mathbf{a})| \leq 1$ για $\mathbf{a} \neq \mathbf{a}_s$, δηλαδή παρεμβαλλόμενα σήματα από άλλες κατευθύνσεις εξασθενούν. Το \mathbf{a}_s ονομάζεται steering direction ή κύριος άξονας απόκρισης (main response axis), ενώ η διαδικασία εισαγωγής χρονικών ολισθήσεων για ευθυγράμμιση των σημάτων που καταγράφονται από τη συστοιχία καλείται **array steering**. Εν γένει, το array steering, όταν προηγείται του beamforming, μετατοπίζει το beam pattern στο χώρο, καθώς μετά την εισαγωγή των χρονικών ολισθήσεων η εξίσωση (7) γίνεται:

$$\mathbf{F}_{\text{steered}}(\omega, \mathbf{a}) = \mathbf{d}(\mathbf{k} - \mathbf{k}_s) F(\omega) \quad (12)$$

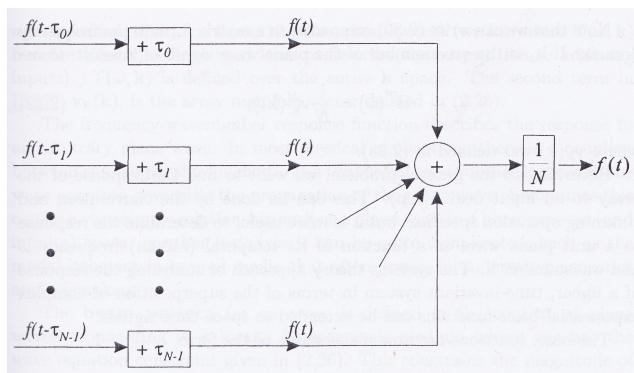
και συνεπώς το beam pattern είναι:

$$B_{\text{steered}}(\omega, \mathbf{a}) = \mathbf{H}^T(\omega) \mathbf{d}(\mathbf{k} - \mathbf{k}_s) = B_{\text{unsteered}}(\omega, \mathbf{a} - \mathbf{a}_s), \quad (13)$$

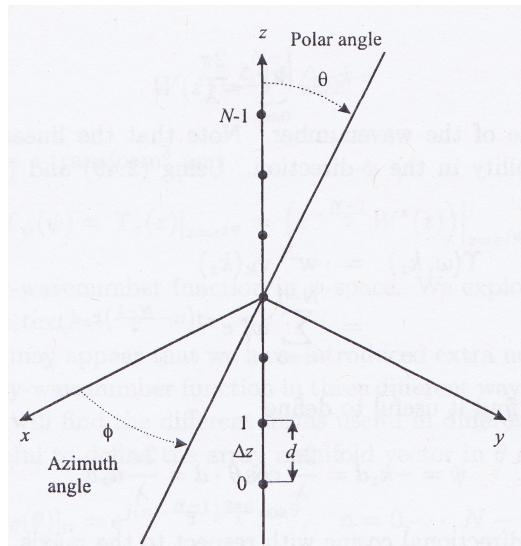
όπου $B_{\text{unsteered}}(\omega, \mathbf{a})$ το unsteered beam pattern που δίνεται από την εξίσωση (9).

Για παράδειγμα, ο delay-and-sum beamformer θα μπορούσε να θεωρηθεί ως ο beamformer με ομοιόμορφα βάρη (uniform weighting) $\mathbf{H}(\omega) = \frac{1}{N} \mathbf{1}$, steered στην κατεύθυνση \mathbf{a}_s :

$$B_{\text{DS}}(\omega, \mathbf{a}) = \frac{1}{N} \mathbf{d}^H(\mathbf{k}_s) \mathbf{d}(\mathbf{k}) = \frac{1}{N} \mathbf{1}^T \mathbf{d}(\mathbf{k} - \mathbf{k}_s). \quad (14)$$



Σχήμα 3: Delay-and-sum beamformer



Σχήμα 4: Ομοιόμορφη γραμμική συστοιχία

To array steering με την εισαγωγή χρονικών ολισθήσεων επιτρέπει μετατόπιση του beam pattern στο χώρο χωρίς να χρειάζεται φυσική μετατόπιση της συστοιχίας. Η μετατόπιση του beam pattern στο χώρο δίνει τη δυνατότητα προσαρμογής του beamformer στη θέση της επιθυμητής πηγής (ώστε το σήμα που προέρχεται από την πηγή αυτή να ενισχύεται, ενώ σήματα από άλλες κατευθύνσεις να εξασθενούν) χωρίς αλλαγή της φυσικής διάταξης της συστοιχίας.

1.3 Ομοιόμορφες γραμμικές συστοιχίες (uniform linear arrays)

Συχνά χρησιμοποιούμενη στην πράξη είναι η γραμμική διάταξη συστοιχιών μικροφώνων. Στη γραμμική διάταξη τα μικρόφωνα τοποθετούνται στην ίδια ευθεία με διάφορες αποστάσεις μεταξύ τους. Η πιο απλή περίπτωση είναι τα μικρόφωνα να ισαπέχουν, οπότε η διάταξη καλείται ομοιόμορφη γραμμική συστοιχία (ΟΓΣ) (uniform linear array, ULA).

Έστω μία ομοιόμορφη γραμμική συστοιχία με N μικρόφωνα που απέχουν κατά d . Χωρίς βλάβη της γενικότητας μπορούμε να υποθέσουμε ότι τα μικρόφωνα είναι τοποθετημένα κατά μήκος του z άξονα ($p_{x_n} = 0, p_{y_n} = 0$) στις εξής θέσεις (Σχήμα 4):

$$p_{z_n} = \left(n - \frac{N-1}{2}\right)d, \quad n = 0, 1, \dots, N-1 \quad (15)$$

Οπότε σε αυτήν την περίπτωση προκύπτουν:

$$\tau_n = \frac{\mathbf{a}^T \mathbf{p}_n}{c} = -\frac{\left(n - \frac{N-1}{2}\right)d \cos \theta}{c} \quad (16)$$

και συνεπώς

$$\mathbf{d}(\mathbf{k}) = e^{-j \frac{N-1}{2} \frac{\omega}{c} d \cos \theta} [1, e^{j \frac{\omega}{c} d \cos \theta}, \dots, e^{j(N-1) \frac{\omega}{c} d \cos \theta}]^T \quad (17)$$

Άρα το delay-and-sum beam pattern είναι:

$$B(\omega, \mathbf{a}) = \frac{1}{N} \mathbf{d}^H(\mathbf{k}_s) \mathbf{d}(\mathbf{k}) = \frac{1}{N} e^{-j \frac{N-1}{2} \frac{\omega}{c} d (\cos \theta - \cos \theta_s)} \sum_{n=0}^{N-1} e^{j \frac{\omega}{c} n d (\cos \theta - \cos \theta_s)} \quad (18)$$

$$B(\omega, \theta) = \frac{1}{N} \frac{\sin[\frac{N}{2} \frac{\omega}{c} d (\cos \theta - \cos \theta_s)]}{\sin[\frac{1}{2} \frac{\omega}{c} d (\cos \theta - \cos \theta_s)]} \quad (19)$$

Παρατηρήστε ότι:

- α. Το beam pattern δεν εξαρτάται από την αζιμουθιακή γωνία ϕ , επομένως αυτή η γεωμετρία δεν έχει δυνατότητα διαχωρισμού σημάτων κατά την αζιμουθιακή γωνία ϕ .
- β. Στην περίπτωση $\theta_s = \frac{\pi}{2}$ είναι $\mathbf{d}(\mathbf{k}_s) = 1$, δηλαδή ο unsteered delay-and-sum beamformer για ΟΓΣ έχει άξονα μέγιστης απόκρισης τον $\theta = \frac{\pi}{2}$.

1.4 Μελέτη χαρακτηριστικών του delay-and-sum beam pattern για ΟΓΣ

Αρχικά θεωρήστε ότι το επιθυμητό σήμα καταφθάνει στη συστοιχία από γωνία $\theta_s = 90^\circ$. Για συχνότητα $f = 2\text{kHz}$, σχεδιάστε το μέτρο του delay-and-sum beam pattern σε λογαριθμική κλίμακα (dB) συναρτήσει της γωνίας $\theta \in [0, 180^\circ]$ και περιγράψτε τί παρατηρείτε³ για τις παρακάτω περιπτώσεις (για καθένα από τα 1 και 2 σχεδιάστε μία γραφική παράσταση που να περιέχει όλες τις καμπύλες για τις διάφορες τιμές της παραμέτρου που μεταβάλλεται):

1. Απόσταση μικροφώνων $d = 4\text{cm}$ αριθμός μικροφώνων $N = 4, 8, 16$.
2. Αριθμός μικροφώνων $N = 8$ και απόσταση μικροφώνων $d = 4\text{cm}, 8\text{cm}, 16\text{cm}$.

Θεωρήστε τώρα ότι η συστοιχία αποτελείται από $N = 8$ μικρόφωνα με απόσταση $d = 4\text{cm}$. Θεωρήστε συχνότητα $f = 2\text{kHz}$.

3. Για $\theta_s = 0^\circ, 45^\circ, 90^\circ$, σχεδιάστε το μέτρο του delay-and-sum beam pattern σε λογαριθμική κλίμακα (dB) συναρτήσει της γωνίας $\theta \in [-180^\circ, 180^\circ]$ σε πολυκό διάγραμμα. Τί παρατηρείτε; Σας δίνεται συνάρτηση `semilogr_polar`, η οποία δέχεται τα ίδια ορίσματα με τη συνάρτηση `polar` του MATLAB, αλλά μετατρέπει τον άξονα της ακτίνας r σε λογαριθμική κλίμακα (dB).

³Πα το 2 δείτε επίσης το παράρτημα A.

Μέρος 2. Εφαρμογή Beamforming για Speech Enhancement

2.1 Beamforming σε προσομοιωμένα σήματα

Θεωρούμε το εξής σενάριο: Μία γραμμική συστοιχία μικροφώνων $N = 7$ στοιχείων με απόσταση $d = 4\text{cm}$ καταγράφει σήματα που παράγονται από δύο σημειακές πηγές. Η μία πηγή παράγει ένα σήμα φωνής και βρίσκεται σε γωνία $\theta = 45^\circ$ σε σχέση με τη συστοιχία. Η δεύτερη πηγή παράγει ένα σήμα θορύβου και βρίσκεται σε γωνία $\theta = 135^\circ$ σε σχέση με τη συστοιχία⁴. Ο θόρυβος είναι ζωνοπερατός και η ενέργειά του είναι συγκεντρωμένη στις συχνότητες $f \in [500\text{Hz}, 2.5\text{kHz}]$. Τα σήματα πηγής και θορύβου είναι ασυσχέτιστα.

Στο συμπληρωματικό υλικό της άσκησης, στο φάκελο `MicArraySimulatedSignals` θα βρείτε το σήμα `source.wav`⁵, το οποίο είναι το καθαρό σήμα φωνής όπως καταγράφεται από το κεντρικό μικρόφωνο της συστοιχίας ($n = 3$) και τα σήματα `sensor_{0,...,6}.wav`, τα οποία είναι τα θορυβώδη σήματα όπως καταγράφονται από τα αντίστοιχα μικρόφωνα⁶. Η συχνότητα δειγματοληψίας είναι 48kHz .

A) Delay-and-sum beamforming

Θα επιχειρήσετε να αποθορυβοποιήσετε το σήμα φωνής με delay-and-sum beamforming:

1. Υπολογίστε τα βάρη για τον delay-and-sum-beamformer (εξίσωση (10)) και εφαρμόστε το beamforming ώστε να προκύψει η έξοδος $y(t)$.

Υπόδειξη: Μπορείτε να υλοποιήσετε το delay-and-sum beamforming σε ένα βήμα υπολογίζοντας το διάνυσμα βαρών \mathbf{H} (εξίσωση (10)). Ισοδύναμα, μπορείτε να υλοποιήσετε το beamforming αυτό σε διαδοχικά στάδια ως εξής: Πρώτα, υλοποιήστε μια συνάρτηση που να ολισθάνει χρονικά (time shift) ένα σήμα στο διαχριτό χρόνο κατά δεδομένο αριθμό δειγμάτων (ενδεχομένως μη ακέραιο)⁷. Στη συνέχεια, χρησιμοποιήστε την σε κάθε σήμα μικροφώνου διαδοχικά, προκειμένου να τα ευθυγραμμίσετε στο χρόνο πριν τα αθροίσετε (Σχήμα 3). Για να ελέγξετε ότι η ευθυγράμμιση των σημάτων και το beamforming που υλοποιήσατε είναι σωστά, αφαιρέστε από την έξοδο του beamformer το καθαρό σήμα φωνής που σας δίνεται και αξιολογήστε το αποτέλεσμα ακουστικά: αν η ευθυγράμμιση των σημάτων στο χρόνο είναι σωστή, πρέπει να ακούτε μόνο θόρυβο χωρίς καθόλου σήμα φωνής.

2. Σχεδιάστε και συγκρίνετε τις κυματομορφές και τα σπεκτρογραφήματα (spectrograms) για τα εξής σήματα: (α) το καθαρό σήμα φωνής, (β) το θορυβώδες σήμα στο κεντρικό μικρόφωνο της συστοιχίας και (γ) την έξοδο $y(t)$ του delay-and-sum beamformer.
3. Υπολογίστε το SNR του θορυβώδους σήματος στο κεντρικό μικρόφωνο και το SNR της έξόδου $y(t)$ του delay-and-sum beamformer και συγκρίνετε.

Υπόδειξη: Το σήμα θορύβου στο κεντρικό μικρόφωνο και έξοδο του beamformer μπορεί να βρεθεί αφαιρώντας από το καθαρό σήμα φωνής που σας δίνεται το θορυβώδες σήμα στο κεντρικό μικρόφωνο και το σήμα στην έξοδο του beamformer, αντίστοιχα.

Παραδοτέο: Η έξοδος του beamformer αποθηκευμένη σε αρχείο `wav` με όνομα `sim.ds`.

B) Μονοκαναλικό Wiener φίλτρωμα

Θα συγκρίνετε την πολυκαναλική μέθοδο αποθορυβοποίησης με τη μονοκαναλική μέθοδο Wiener filtering.

Θεωρήστε το πλαίσιο (frame) $f(t, \mathbf{p}_3)$, $t \in [0.36s, 0.39s]$ διάρκειας 30ms, το οποίο περιέχει έναν έμφωνο ήχο καταγεγραμμένο από το κεντρικό μικρόφωνο της συστοιχίας ($n = 3$). Το πλαίσιο αυτό μπορεί να μοντελοποιηθεί ως $x(t) = s(t) + v(t)$, όπου $s(t)$ η συνιστώσα του $x(t)$ που οφείλεται στο επιψυμητό σήμα φωνής, ενώ $v(t)$ είναι η συνιστώσα θορύβου. Θα επιχειρήσετε μονοκαναλική αποθορυβοποίηση του πλαισίου αυτού με Wiener φίλτρωμα.

⁴ Θεωρούμε ότι οι θέσεις των μικροφώνων είναι όπως στο σχήμα 4 (εξίσωση (15)) και οι γωνίες μετρώνται όπως φαίνεται στο ίδιο σχήμα.

⁵ Το σήμα αυτό προέρχεται από πραγματική ηχογράφηση.

⁶ Τα σήματα αυτά έχουν παραχθεί τεχνητά προσομοιώνοντας τις καθυστερήσεις διάδοσης και αυθοίζοντας την πηγή με τεχνητό θόρυβο.

⁷ Για την υλοποίηση φίλτρων κλασματικής καθυστέρησης δείτε επίσης το παρότρημα B

- Τυπολογίστε την απόχριση συχνότητας του IIR Wiener φίλτρου (το οποίο πρακτικά μπορεί να υλοποιηθεί στο πεδίο συχνότητας με DFT), η οποία, εφόσον τα σήματα φωνής και θορύβου είναι ασυσχέτιστα⁸, είναι:

$$H_W(\omega) = \frac{P_s(\omega)}{P_x(\omega)} = 1 - \frac{P_v(\omega)}{P_x(\omega)}, \quad (20)$$

όπου $P_v(\omega)$ το φάσμα ισχύος (power spectrum) του θορύβου και $P_x(\omega)$ το φάσμα ισχύος του συνολικού σήματος $x(t)$. Εκτιμήστε τα φάσματα ισχύος με τη μέθοδο Welch. Σχεδιάστε σε λογαριθμική κλίμακα (dB) το $H_W(\omega)$ για συχνότητες $f \in [0, 8\text{kHz}]$.

Χρήσιμη συνάρτηση MATLAB: `pwelch`

Το Wiener φίλτρο προκαλεί παραμόρφωση (distortion) στο σήμα φωνής $s(t)$, η οποία ισούται με $s(t) - h_W(t) * s(t)$. Ένας τρόπος να μετρηθεί η παραμόρφωση αυτή είναι το speech distortion index, το οποίο ορίζεται ως ο λόγος του φάσματος ισχύος της παραμόρφωσης προς το φάσμα ισχύος του σήματος φωνής:

$$n_{sd}(\omega) = \frac{\mathbb{E}[|S(\omega) - H_W(\omega)S(\omega)|^2]}{P_x(\omega)} = |1 - H_W(\omega)|^2 \quad (21)$$

- Τυπολογίστε και σχεδιάστε σε λογαριθμική κλίμακα (dB) το $n_{sd}(\omega)$ για συχνότητες $f \in [0, 8\text{kHz}]$. Τί παρατηρείτε; Εξηγείστε.
- Εφαρμόστε το Wiener φίλτράρισμα. Σχεδιάστε στην ίδια γραφική παράσταση και συγκρίνετε τα φάσματα ισχύος για τα εξής σήματα: (α) το καθαρό σήμα φωνής στην είσοδο του Wiener φίλτρου $s(t)$, (β) τη θορυβώδη είσοδο του Wiener φίλτρου $x(t)$, (γ) την έξοδο του Wiener φίλτρου και (δ) το σήμα θορύβου $v(t)$ στην είσοδο του Wiener φίλτρου. Σχεδιάστε τα φάσματα ισχύος σε λογαριθμική κλίμακα (dB) και για συχνότητες $f \in [0, 8\text{kHz}]$.
- Τυπολογίστε το SNR στην έξοδο του Wiener φίλτρου. Συγκρίνετε με το SNR της εισόδου $x(t)$. Τυπολογίστε τη βελτίωση στο SNR του συγκεκριμένου πλαισίου που επιτεύχθηκε με την πολυκαναλική μέθοδο. Σχεδιάστε στην ίδια γραφική παράσταση και συγκρίνετε τα φάσματα ισχύος για τα εξής σήματα: (α) το καθαρό σήμα φωνής για το πλαίσιο υπό μελέτη $s(t)$, (β) το θορυβώδες πλαίσιο $x(t)$, (γ) την έξοδο του Wiener φίλτρου και (δ) την έξοδο του delay-and-sum beamformer για το πλαίσιο υπό μελέτη. Σχεδιάστε τα φάσματα ισχύος σε λογαριθμική κλίμακα (dB) και για συχνότητες $f \in [0, 8\text{kHz}]$. Συγκρίνετε την πολυκαναλική μέθοδο με το μονοκαναλικό Wiener φίλτρο.

2.2 Beamforming σε πραγματικά σήματα

Θεωρήστε τώρα το εξής σενάριο: Μία γραμμική συστοιχία μικροφώνων $N = 7$ στοιχείων με απόσταση $d = 4\text{cm}$ βρίσκεται σε ένα θορυβώδες δωμάτιο και καταγράφει το σήμα φωνής που εκφωνεί ένας άνθρωπος σε γωνία $\theta = 45^\circ$. Ο θόρυβος δεν προέρχεται από σημειακή πηγή αλλά από διάφορες πηγές, όπως ανεμιστήρες κ.ά., οι οποίες δημουργούν ένα ισοτροπικό και ομογενές πεδίο θορύβου που ονομάζεται diffuse noise field. Ο θόρυβος μπορεί να θεωρηθεί ότι είναι στάσιμος (stationary random process).

Στο συμπληρωματικό υλικό της άσκησης, στο φάκελο `MicArrayRealSignals` θα βρείτε το σήμα `source.wav`, το οποίο είναι το καθαρό σήμα φωνής όπως καταγράφεται στη θέση της πηγής και τα σήματα `sensor_{0, ..., 6}.wav`, τα οποία είναι τα θορυβώδη σήματα όπως καταγράφονται από τα αντίστοιχα μικρόφωνα⁹. Η συχνότητα δειγματοληψίας είναι 48kHz .

A) Delay-and-sum beamforming

- Τυπολογίστε τα βάρη για τον delay-and-sum beamformer όπως προηγουμένως και εφαρμόστε το beamforming.

⁸Συνεπώς $P_x(\omega) = P_s(\omega) + P_v(\omega)$, διότι $P_{xv}(\omega) = 0$

⁹Όλα τα σήματα έχουν ηχογραφηθεί σε πραγματικές συνθήκες.

2. Σχεδιάστε και συγχρίνετε τις χυματομορφές και τα σπεκτρογραφήματα (spectrograms) για τα εξής σήματα: (α) το καθαρό σήμα φωνής, (β) το θορυβώδες σήμα στο κεντρικό μικρόφωνο της συστοιχίας και (γ) την έξοδο του delay-and-sum beamformer.

Για σήματα φωνής, έχει βρεθεί ότι το ολικό SNR ως μετρική ποιότητας δε συμβαδίζει με την υποκειμενική αντίληψη ποιότητας που έχει ο άνθρωπος για το σήμα (perceptual evaluation of speech quality). Μία πιο κατάλληλη μετρική είναι το segmental SNR (SSNR) το οποίο ορίζεται ως το μέσο SNR των πλαισίων βραχέος χρόνου του σήματος φωνής:

$$\text{SSNR} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \frac{\sum_{n=Lm}^{Lm+L-1} s^2(n)}{\sum_{n=Lm}^{Lm+L-1} v^2(n)}, \quad (22)$$

όπου L το μήκος των πλαισίων, $s(n)$ το σήμα φωνής και $v(n)$ ο θόρυβος. Πλαίσια με SNR μεγαλύτερο των 35dB δεν έχουν σημαντικές διαφορές στην ποιότητα σήματος και τίθενται στα 35dB για την εξαγωγή του μέσου όρου. Σε πλαίσια σιωπής το SNR είναι έντονα αρνητικό, οπότε πλαίσια με SNR μικρότερο μίας τιμής κατωφλίου που θέτει το όριο διαχωρισμού φωνής από σιωπή αγνοούνται κατά τον υπολογισμό του μέσου όρου. Συνήθως, για το κατώφλι επιλέγεται μία τιμή στο διάστημα [-20dB, 0dB].

3. Υπολογίστε το SSNR στο κεντρικό μικρόφωνο της συστοιχίας και στην έξοδο του beamformer και συγχρίνετε. Είναι ικανοποιητική η βελτίωση;

Τυπόδειξη: Την ισχύ του θορύβου σ_v^2 μπορείτε να την υπολογίσετε από ένα κομμάτι του σήματος που περιέχει μόνο θόρυβο. Θεωρώντας ότι ο θόρυβος είναι στάσιμος, το σ_v^2 δε μεταβάλλεται με το χρόνο (είναι το ίδιο για όλα τα πλαίσια). Σε κάθε πλαίσιο $x(t) = s(t) + v(t)$ μπορείτε να υπολογίσετε την ισχύ του σήματος φωνής $s(t)$ ως $\sigma_s^2 = \sigma_x^2 - \sigma_v^2$, θεωρώντας ότι ο θόρυβος και το σήμα φωνής είναι ασυσχέτιστα.

Παραδοτέο: Η έξοδος του beamformer αποθηκευμένη σε αρχείο wav με όνομα real.ds.

B) Post-filtering με Wiener φίλτρο (προαιρετικό: bonus¹⁰ 20%).

Στην περίπτωση του diffuse noise field, ο delay-and-sum beamformer δεν έχει καλή απόδοση, διότι τα σήματα θορύβου εμφανίζουν μεγάλη συσχέτιση μεταξύ μικροφώνων, ειδικά στις χαμηλές συχνότητες. Είναι συνήθης πρακτική να εφαρμόζεται και μονοκαναλικό φιλτράρισμα μετά το beamforming για περαιτέρω βελτίωση της ποιότητας του σήματος. Η διαδικασία αυτή ονομάζεται post-filtering.

1. Εφαρμόστε μονοκαναλικό IIR Wiener φιλτράρισμα στην έξοδο του beamformer. Το φάσμα ισχύος του θορύβου μπορείτε να το εκτιμήστε με τη μέθοδο Welch από ένα κομμάτι του αρχικού σήματος που περιέχει μόνο θόρυβο και να θεωρήσετε ότι δε μεταβάλλεται με το χρόνο, καθώς ο θόρυβος είναι στάσιμος. Επειδή το σήμα φωνής δεν είναι στάσιμο πρέπει να ακολουθήσετε ανάλυση βραχέος χρόνου (short-time analysis) χωρίζοντας το σήμα σε επικαλυπτόμενα πλαίσια διάρκειας 25 – 30ms. Χρησιμοποιήστε Hamming παράθυρο. Για κάθε πλαίσιο υπολογίστε τη συνάρτηση μεταφοράς του Wiener φίλτρου εκτιμώντας το φάσμα ισχύος για το κάθε πλαίσιο του θορυβώδους σήματος με τη μέθοδο Welch. Φιλτράρετε κάθε πλαίσιο και τέλος ανασυνθέστε το σήμα στην έξοδο με overlap-add σύνθεση (για να είναι δυνατή η ανακατασκευή του σήματος με overlap-add σύνθεση πρέπει να έχετε επιλέξει κατάλληλα το ποσοστό επικάλυψης μεταξύ πλαισίων).
2. Σχεδιάστε και συγχρίνετε τις χυματομορφές και τα σπεκτρογραφήματα για τα εξής σήματα: (α) το καθαρό σήμα φωνής, (β) το θορυβώδες σήμα στο κεντρικό μικρόφωνο της συστοιχίας, (γ) την είσοδο του Wiener φίλτρου και (δ) την έξοδο του Wiener φίλτρου.
3. Υπολογίστε το SSNR στην είσοδο και στην έξοδο του Wiener φίλτρου και συγχρίνετε.
4. Υπολογίστε το μέσο όρο των SSNRs των σημάτων εισόδου στο σύστημα delay-and-sum-beamformer + Wiener post-filter και συγχρίνετε με το SSNR στην τελική έξοδο του συστήματος. Πόση βελτίωση επιτεύχθη;

¹⁰Επί του βαθμού της παρούσας δισκησης

Παραδοτέο: Η έξοδος του Wiener φίλτρου αποθηκευμένη σε αρχείο wav με όνομα `real_mmse`.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] H. L. Van Trees, *Optimum Array Processing*. Wiley, 2002.
- [2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer, 2008.
- [3] 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.

ΠΑΡΑΡΤΗΜΑ

A. Χωρική Δειγματοληψία (Spatial Sampling) και το Φαινόμενο των Grating Lobes

To delay-and-sum beam pattern για ομοιόμορφη γραμμική συστοιχία δίνεται από την εξίσωση (19):

$$B(\omega, \theta) = \frac{1}{N} \frac{\sin[\frac{N}{2} \frac{\omega}{c} d(\cos \theta - \cos \theta_s)]}{\sin[\frac{1}{2} \frac{\omega}{c} d(\cos \theta - \cos \theta_s)]}$$

Το μέτρο του beam pattern $|B(\omega, \theta)|$ γίνεται 1 όταν:

$$\sin[\frac{1}{2} \frac{\omega}{c} d(\cos \theta - \cos \theta_s)] = 0, \quad (23)$$

ισοδύναμα όταν:

$$\cos \theta = \cos \theta_s + m \frac{c}{d} \frac{2\pi}{\omega} = \cos \theta_s + m \frac{\lambda}{d}, \quad m = \dots, -1, 0, 1, \dots, \quad (24)$$

όπου $\lambda = \frac{c}{f}$ το μήκος κύματος.

Για $m = 0$, η λύση είναι $\theta = \pm \theta_s$, δηλαδή το μέτρο του beam pattern είναι 1 στο steering direction, όπως αναμένεται.

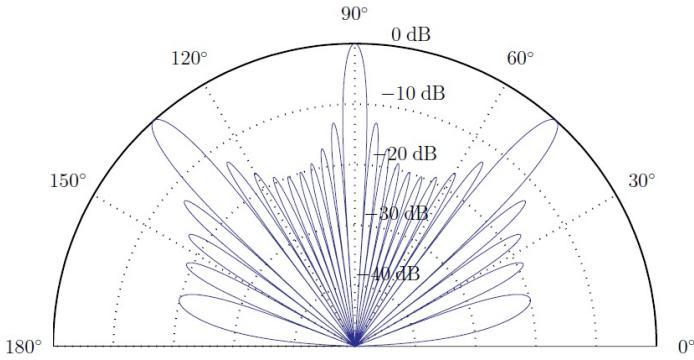
Αν όμως $|\cos \theta_s + \frac{\lambda}{d}| \leq 1$ ή $|\cos \theta_s - \frac{\lambda}{d}| \leq 1$, τότε θα υπάρχει και $\theta \neq \theta_s$ για το οποίο $|B(\omega, \theta)| = 1$, δηλαδή ο beamformer δε θα μπορεί να διαχωρίσει το σήμα που προέρχεται από την κατεύθυνση θ_s από σήματα που προέρχονται από άλλες κατευθύνσεις για τις οποίες $|B(\omega, \theta)| = 1$, αφού και αυτά τα σήματα θα παραμένουν αναλλοίωτα μετά το beamforming. Για αποφυγή του φαινομένου αυτού απαιτείται:

$$d < \frac{\lambda}{1 + |\cos \theta_s|} \quad (25)$$

Για να ικανοποιείται αυτό για κάθε θ_s απαιτείται:

$$d < \frac{\lambda}{2} \quad (26)$$

Η ανίσωση αυτή μπορεί να ερμηνευθεί ως χωρικό θεώρημα δειγματοληψίας, το οποίο δείχνει πόσο πυκνή πρέπει να είναι η δειγματοληψία που κάνει η συστοιχία στο χώρο συναρτήσει του μήκους κύματος του ακουστικού σήματος που διαδίδεται σε αυτόν. Αν η ανίσωση αυτή δεν ικανοποιείται τότε δημιουργείται spatial aliasing: στο beam pattern εμφανίζονται grating lobes, δηλαδή πλευρικοί λοβοί μοναδιαίου πλάτους (Σχήμα 5) και συνεπώς ο beamformer αδυνατεί να διαχωρίσει το σήμα που προέρχεται από το steering direction από σήματα που προέρχονται από τις κατευθύνσεις των grating lobes.



Σχήμα 5: Το φαινόμενο των grating lobes σε delay-and-sum beam pattern συστοιχίας με $N = 10$, $d = \frac{3}{2}\lambda$, $\theta_s = 90^\circ$

B. Φίλτρο Κλασματικής Καθυστέρησης (Fractional Delay Filter)

Για την υλοποίηση delay-and-sum beamforming στο διαχριτό χρόνο χρειάζεται χρονική ολίσθηση (time shift) των σημάτων που καταγράφονται στα μικρόφωνα. Η απαιτούμενη χρονική ολίσθηση ενδέχεται να μην είναι ακέραιος αριθμός δειγμάτων. Για την επίτευξη χρονικής ολίσθησης κατά μη ακέραιο αριθμό δειγμάτων χρειάζεται παρεμβολή (interpolation).

Έστω ένα ζωνοπερατό (band-limited) σήμα συνεχούς χρόνου $x(t)$, με (μονόπλευρο) εύρος ζώνης B . Έστω $x[n]$ η δειγματοληπτημένη έκδοση του $x(t)$ με συχνότητα δειγματοληψίας F_s :

$$x[m] = x(mT_s), \quad (27)$$

όπου $T_s = \frac{1}{F_s}$ η περίοδος δειγματοληψίας.

Αν $F_s > 2B$, τότε είναι δυνατή η ανακατασκευή του $x(t)$ από τα δείγματα $x[n]$ με sinc interpolation¹¹:

$$x(t) = \sum_{m=-\infty}^{\infty} x[m] \text{sinc}\left(\frac{t - mT_s}{T_s}\right) \quad (28)$$

$$= \sum_{m=-\infty}^{\infty} x[m] \frac{\sin(\pi(t - mT_s)/T_s)}{\pi(t - mT_s)/T_s} \quad (29)$$

Συνεπώς, αν εισαχθεί μία καθυστέρηση t_0 (στο συνεχή χρόνο):

$$x(t - t_0) = \sum_{m=-\infty}^{\infty} x[m] \text{sinc}\left(\frac{t - t_0 - mT_s}{T_s}\right) \quad (30)$$

Δειγματοληπτώντας το $x(t - t_0)$, προκύπτουν τα δείγματα:

$$x_D[n] = x(nT_s - t_0) \quad (31)$$

$$= \sum_{m=-\infty}^{\infty} x[m] \text{sinc}\left(\frac{nT_s - t_0 - mT_s}{T_s}\right) \quad (32)$$

$$= \sum_{m=-\infty}^{\infty} x[m] \text{sinc}\left[n - m - \frac{t_0}{T_s}\right] \quad (33)$$

$$= x[n] * \text{sinc}[n - D], \quad (34)$$

όπου $D = \frac{t_0}{T_s}$ η καθυστέρηση σε αριθμό δειγμάτων (όχι απαιραίτητα ακέραιος) και $*$ συμβολίζει συνέλιξη.

Συνεπώς, στο διαχριτό χρόνο, η εισαγωγή κλασματικής καθυστέρησης D μπορεί να επιτευχθεί με sinc interpolation, δηλαδή με χρήση του φίλτρου:

$$h_D[n] = \text{sinc}[n - D] \quad (35)$$

¹¹Το χρονικό ισοδύναμο του ιδανικού βαθυπερατού φίλτραρίσματος με συχνότητα αποκοπής $\frac{F_s}{2}$.

Στο πεδίο συχνότητας:

$$H_D(e^{j\omega}) = e^{-j\omega D} \quad (36)$$

Το φίλτρο H_D είναι IIR. Η υλοποίηση του μπορεί να γίνει με οποιαδήποτε μέθοδο υλοποίησης IIR φίλτρων.

Μία δυνατή υλοποίηση του φίλτρου κλασματικής καθυστέρησης H_D είναι στο πεδίο συχνότητας με χρήση DFT:

$$H_D[k] = e^{-j\omega_k D}, \quad k = 0, \dots, N - 1 \quad (37)$$

όπου N το μήκος του DFT. Λόγω του άπειρου μήκους της χρονιστικής απόκρισης $h_d[n]$, με την προσέγγιση αυτή δημιουργείται aliasing στο χρόνο. Ωστόσο, επειδή η συνάρτηση sinc αποσβέννυται αυξανομένου του $|n|$, με χρήση επαρκώς μεγάλου μήκους DFT το aliasing θα είναι σχετικά μικρό.