בס"ד

הפקולטה להנדסה

המעבדה לעיבוד אותות

Speech Separation and Denoising using LCMV Beamforming

יוסף רשף

פרויקט שנה ד' לקראת תואר ראשון בהנדסה

מנחה: מר אייל שורץ

אחראי אקדמי: פרופ ' שרון גנות

אוקטובר 2021

תוכן עניינים

הצגת הבעיה.....3

RIR Generator.....5

עיצוב אלומה וסינון מרחבי (Beamforming + Spatial Filtering).....9

LCMV.....10

שיטת הפחתה ספקטרלית (Covariance Substruction).....13

שיטת GEVD ו-Covariance Whitening.....15

סימולציות מטלב ותוצאות.....17

רעיונות להמשך המחקר.....33

ביבליוגרפיה.....34

הערה: הספר נכתב בלשון רבים מטעמי נוחות בלבד.

בס"ד

הצגת הבעיה

בפרויקט זה נעסוק בבעיה נפוצה בתחום עיבוד אותות הדיבור – מסיבת הקוקטייל (Cocktail Party). במקרה זה, ישנם מספר אנשים שמדברים בו זמנית בחדר מהדהד ורועש שהצבנו בו מספר מיקרופונים. מטרותינו הן להפריד את הדוברים על מנת שנוכל להתמקד בדובר מסוים בלבד (כלומר להפריד ולהנחית את כל שאר הדוברים הלא-רצויים) ובנוסף גם להנחית את הרעש הנלווה לדוברים.

בעיה זו נפוצה בחיים האמיתיים לכל מקרה בו אנשים נמצאים במקום מסוים ומדברים ביחד.

לשם פשטות, אנו נתעסק במקרה שבו נמצאים בחדר שני דוברים בלבד (ניתן כמובן לבצע הרחבה למספר גדול יותר של דוברים).

אנו נתעסק במקרה שבו נמצאים בחדר 7 מיקרופונים. במקרה הכללי (אליו יתייחסו כמה מהמשוואות שנציג להלן) יש בחדר מערך של M מיקרופונים שהינם קרובים מאוד אחד לשני. המיקרופונים קולטים דיבור משני האנשים בתוספת רעש.

כל קול של כל דובר שמגיע לכל מיקרופון עובר למעשה סינון על ידי פונקציית תמסורת של החדר שמתאימה לדובר ולמיקרופון. לכן, אם יש 2 דוברים ו-7 מיקרופונים, קיימות 14 מערכות אליהן צריך להתייחס.

נסמן:

s^d – אות הדיבור הרצוי.

s^i – אות הדיבור הלא-רצוי (המפריע).

$\boldsymbol{h^d}$ – פונקציות התמסורת האקוסטיות שמקשרות בין הדובר הרצוי לבין המיקרופונים.

$\boldsymbol{h^i}$ – פונקציות התמסורת האקוסטיות שמקשרות בין הדובר הלא-רצוי לבין המיקרופונים.

$\boldsymbol{v}$ – רעשי הרקע במיקרופונים.

$\boldsymbol{z}$ – האותות הנקלטים במיקרופונים. זה יהיה הקלט שהאלגוריתם יקבל.

נשים לב כי וקטורים יסומנו באותיות מודגשות בניגוד לסקלרים. האותות (הקולות) שיוצאים משני הדוברים הינם סקלרים. אלו אותות קבועים שזהים בכל המיקרופונים. השוני בין המיקרופונים נובע מהשוני במערכות. המערכות שונות גם בין המיקרופונים וגם בין הדוברים.

במישור הזמן, אותות הדיבור עוברים קונבולוציה עם כל המערכות, ואל המיקרופונים מגיעות התוצאות בתוספת רעש. לכן, במישור התדר ניתן לתאר את האותות הנקלטים במיקרופונים כך ([1]):

$$\mathbf{z}(l,k) = \mathbf{h}^d(l,k)s^d(l,k) + \mathbf{h}^i(l,k)s^i(l,k) + \mathbf{v}(l,k)$$

כאשר מישור התדר שלנו יהיה מישור *STFT*, שבו l הינו אינדקס של הפריימים ו-k הינו אינדקס התדר. אנו משתמשים במישור *STFT* משום שאותות דיבור אינם אותות סטציונריים ולכן על ידי התמרה לזמן קצר נוכל לזהות מקומות שבהם ההרכבים התדריים שונים. כך נוכל לזהות בדיוק מתי יש דיבור ומתי אין דיבור. לא היינו יכולים להסיק זאת בקלות אם היינו משתמשים בהתמרת פורייה רגילה שהיא למעשה פעולת מיצוע או שיטוח של האות. פעולה זאת מתאימה יותר לאותות סטציונריים אך זה לא מועיל לנו כאן ולכן נשתמש בהתמרת *STFT*.

הנחה חשובה שאנו מניחים הינה שהדוברים הינם סטטיים בחדר ואינם זזים. על כן, פונקציות התמסורת אינן תהליכים אקראיים אלא פונקציות דטרמיניסטיות. התהליך האקראי היחיד הינו הרעש המתווסף.

מתקיים:

$$\mathbf{z}(l,k) = [z_1(l,k), \ldots, z_M(l,k)]^T$$
$$\mathbf{v}(l,k) = [v_1(l,k), \ldots, v_M(l,k)]^T$$
$$\mathbf{h}^d(l,k) = [h_1^d(l,k), \ldots, h_M^d(l,k)]^T$$
$$\mathbf{h}^i(l,k) = [h_1^i(l,k), \ldots, h_M^i(l,k)]^T$$

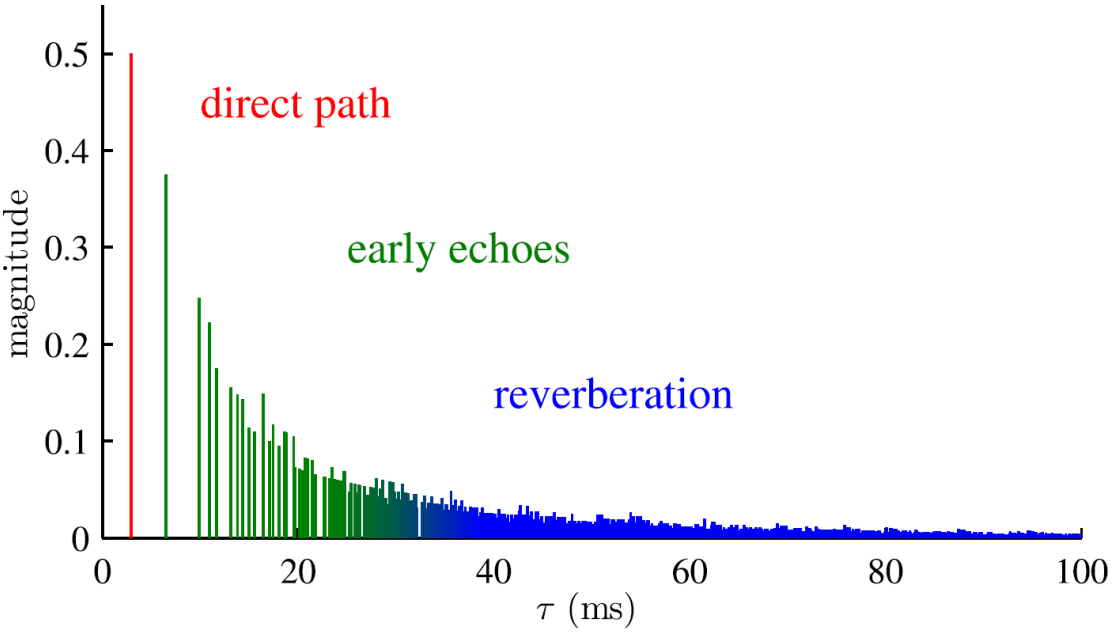
כלומר, עבור כל אינדקס זמן-תדר, יש ערכים שונים במיקרופונים השונים.

הבעיה המרכזית שלנו היא שאנחנו לא יודעים את פונקציות התמסורת. אנחנו רק מקבלים תרכובת של שני אותות שעברו קונבולוציה עם המערכות ובתוספת רעש. איך נוכל מ-$\boldsymbol{z}$ לדעת את המערכות $\boldsymbol{h}$?

RIR (Room Impulse Response) Generator

לפני שנפתור את הבעיה שהצגנו לעיל, נסביר כיצד נראה החדר וכיצד נייצר אותו במטלב.

באופן כללי, תגובה להלם של מיקרופון שנמצא בחדר מהדהד נראית כך ([2]):



כפי שניתן לראות, ניתן לחלק את הגרף לשלושה חלקים:

החלק הראשון הוא הקו האדום המכונה *direct path*. קו זה מייצג את האות שמגיע ישירות מפיו של הדובר אל המיקרופון לפני ההדהודים.

החלק השני, הירוק, מכונה *early echoes*. מכיוון שהחדר מהדהד, הקול פוגע בקירותיו (או ברהיטים שונים הנמצאים בחדר וכו') ומוחזר מהם בתוך החדר למיקרופון. מטבע הדברים, ככל שהקול מתקדם בתוך החדר עוצמתו יורדת וניתן לראות זאת בגרף.

החלק השלישי, הכחול, מכונה *reverberation*. חלק זה מייצג את ההדהודים הקטנים המהווים את ה"זנב אקספוננציאלי" של פונקציית התמסורת.

באופן כללי, ההדהוד הינו אינסופי, אך אנו נגדיר נקודת זמן שהחל ממנה נזניח את ההדהודים האלה. לנקודת זמן זו נקרא t_{60} שתסמל את נקודת הזמן שבה ההנחתה של האות המקורי והישיר הינה *60dB*. מנקודה זו ואילך ניתן להזניח את ההדהודים הבאים ולהניח שההדהוד הסתיים, משום שהאוזן האנושית כבר לא מושפעת כלל מהדהודים זניחים אלו. זמן זה נקרא גם זמן הדהוד = *Reverberation Time*.

בשנת 1979, *Allen* ו-*Berkley* הציעו שיטה למדל חדר אקוסטי תלת מימדי באמצעות שיטה שנקראת *Image Method* (שלא נכנסנו לפרטיה). שיטה זו מסתמכת על משוואת הגלים ומשוואת הלמהולץ על מנת לתאר התפשטות של גלים אקוסטיים דרך תווך (נוזלי או אווירי/גז).

הורדנו מאתר *Audio Labs* פונקציה שנקראת *rir_generator*. פונקציה זו משתמש בשיטה הנ"ל על מנת ליצור חדר אקוסטי על פי פרמטרים שונים. חלק מהפרמטרים שבהם משתמשת הפונקציה לבניית החדר הינם:

$c = 340\frac{m}{s}$ – מהירות הקול.

$f_s = 16000Hz$ – תדר הדגימה.

Mic_position – מיקום המיקרופון/נים בחדר.

Speaker_position – מיקום הדובר בחדר.

L - מימדי החדר.

Beta – מקדם ההדהוד. שווה ל-t_{60} עליו הרחבנו קודם.

N – מספר הדגימות.

Order – מקדם החזרה של הקירות בחדר. כאשר ערך הפרמטר הינו 1-, ההחזרה מהקירות הינה מושלמת ללא כל בליעה של הקול בקירות.

Dim – מספר מימדי החדר (2 או 3). אצלנו, כמובן, נשתמש בחדר תלת מימדי.

שינוי של הפרמטרים הללו יוביל לפונקציית תמסורת שונה.

ניתן מספר דוגמאות כאשר ישנו דובר אחד ומיקרופון אחד בחדר תלת מימדי, ונראה כיצד שינוי זמן ההדהוד ישפיע על התגובה להלם של החדר.

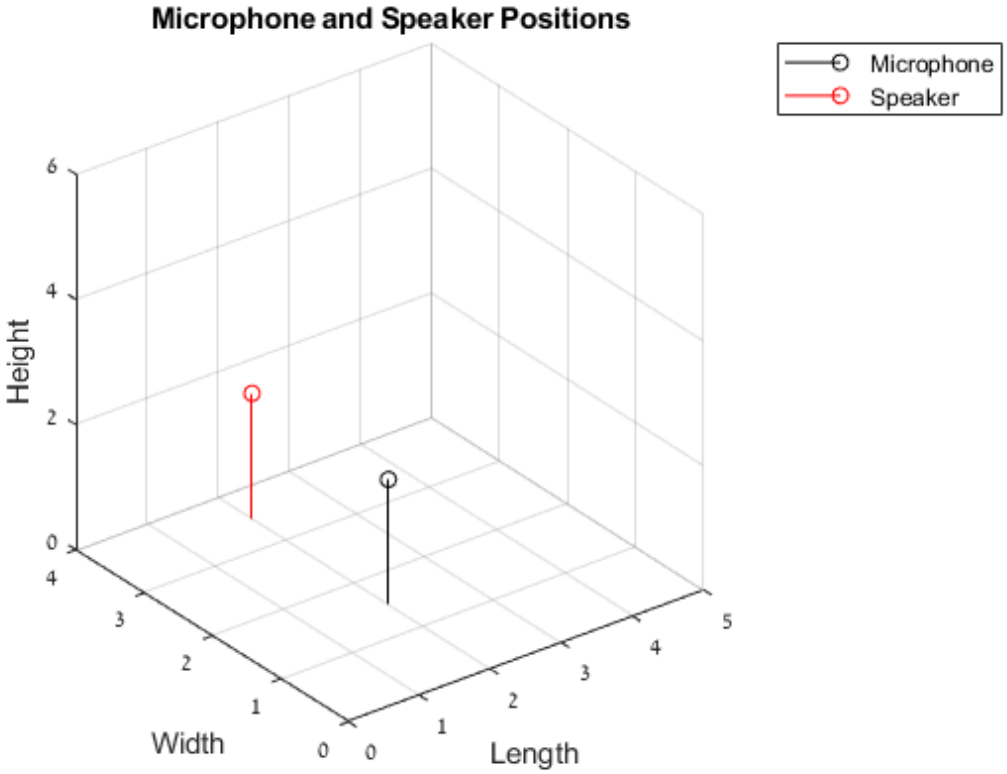
באופן כללי, כאשר נקטין את זמן ההדהוד, האותות שיגיעו למיקרופונים יהיו יותר "מדויקים" כי ההדהודים מעוותים במידה מסוימת את האות המקורי (או אם נרצה, את ה-*direct path*).

נבנה את החדר לדוגמה על פי קוד המטלב הבא ([3]):

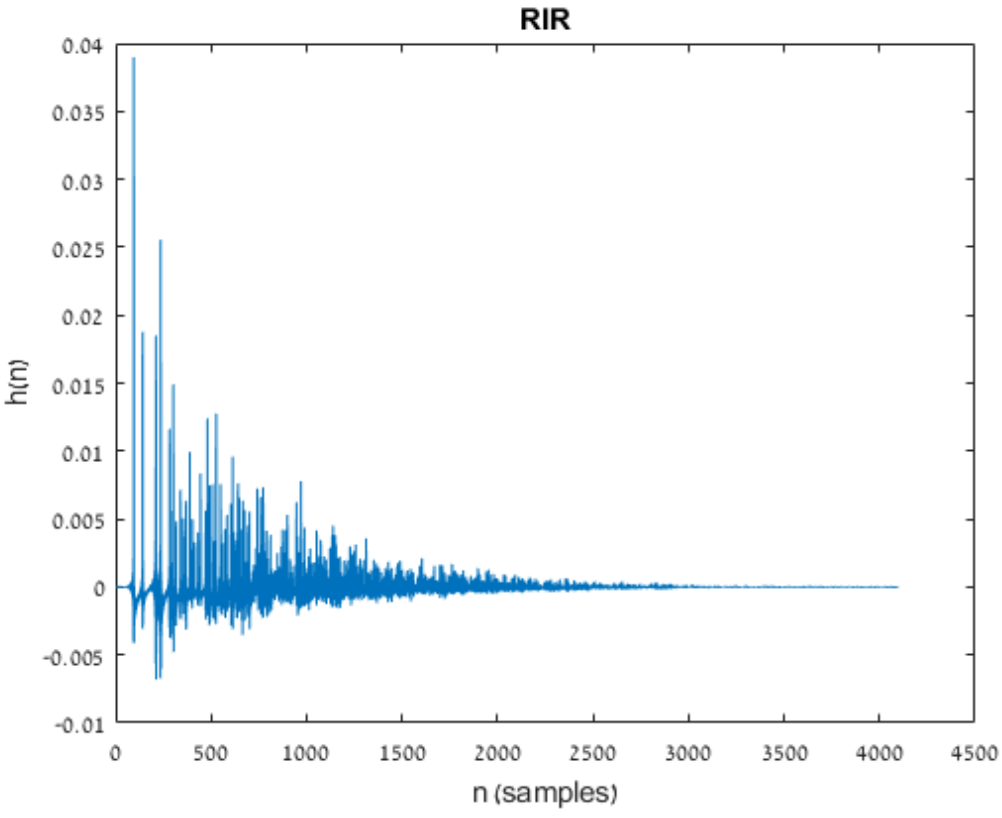
```
c = 340;                    % Sound velocity (m/s)
fs = 16000;                 % Sample frequency (samples/s)
r = [2 1.5 2];              % Receiver position [x y z] (m)
s = [2 3.5 2];              % Source position [x y z] (m)
L = [5 4 6];                % Room dimensions [x y z] (m)
n = 4096;                   % Number of samples
beta = 0.3;                 % Reverberation time (s)
mtype = 'omnidirectional';  % Type of microphone
order = -1;                 % -1 equals maximum reflection order!
Dim = 3;                    % Room dimension
orientation = 0;            % Microphone orientation (rad)
hp_filter = 1;              % Disable high-pass filter

h = rir_generator(c,fs,r,s,L,beta,n,mtype,order,dim,orientation,hp_filter);
```

בס"ד

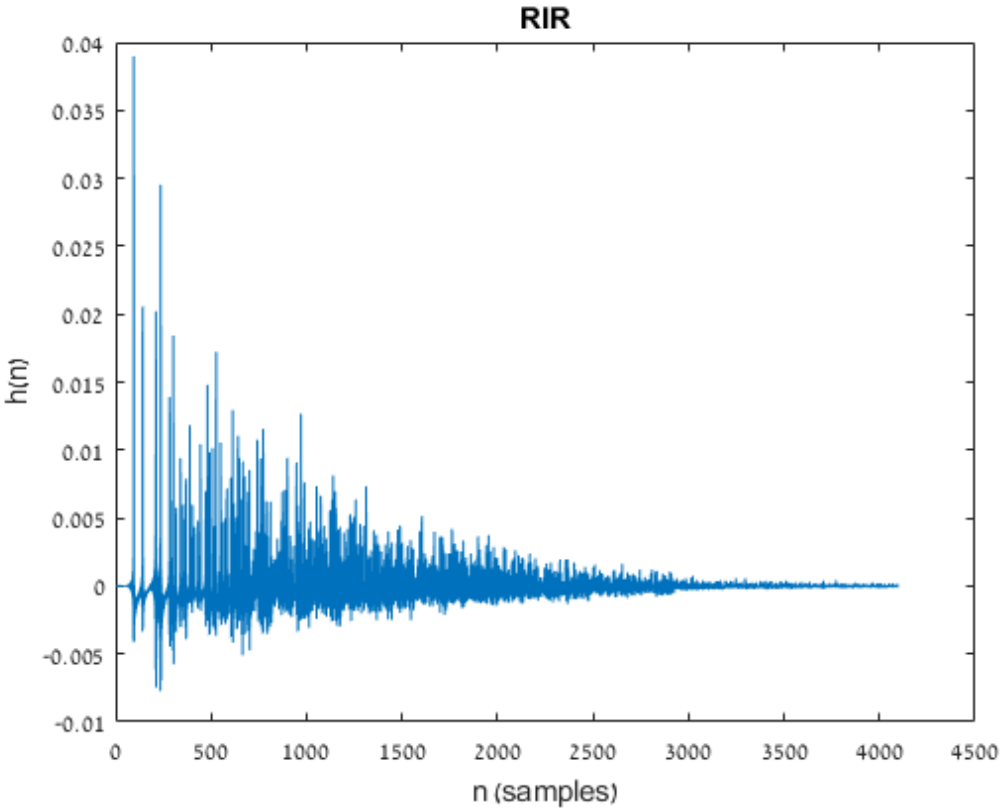
החדר נראה כך:



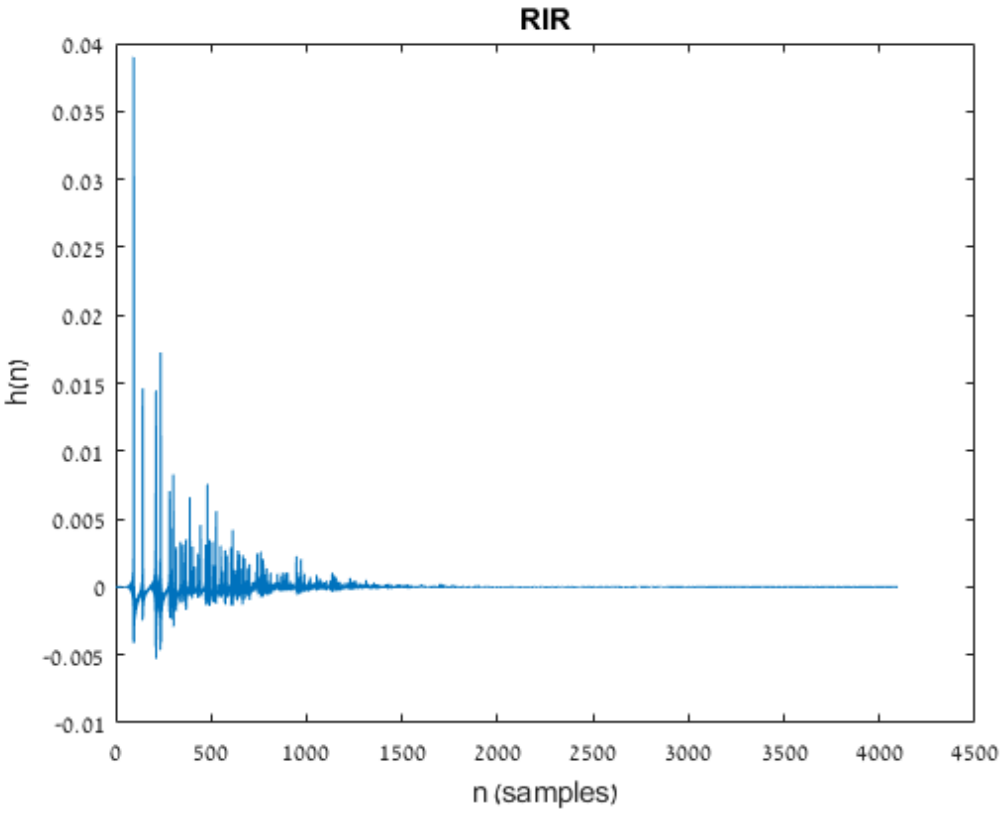
ונקבל את התגובה להלם:



אם נגדיל את זמן ההדהוד ל-*0.4ms* נקבל:



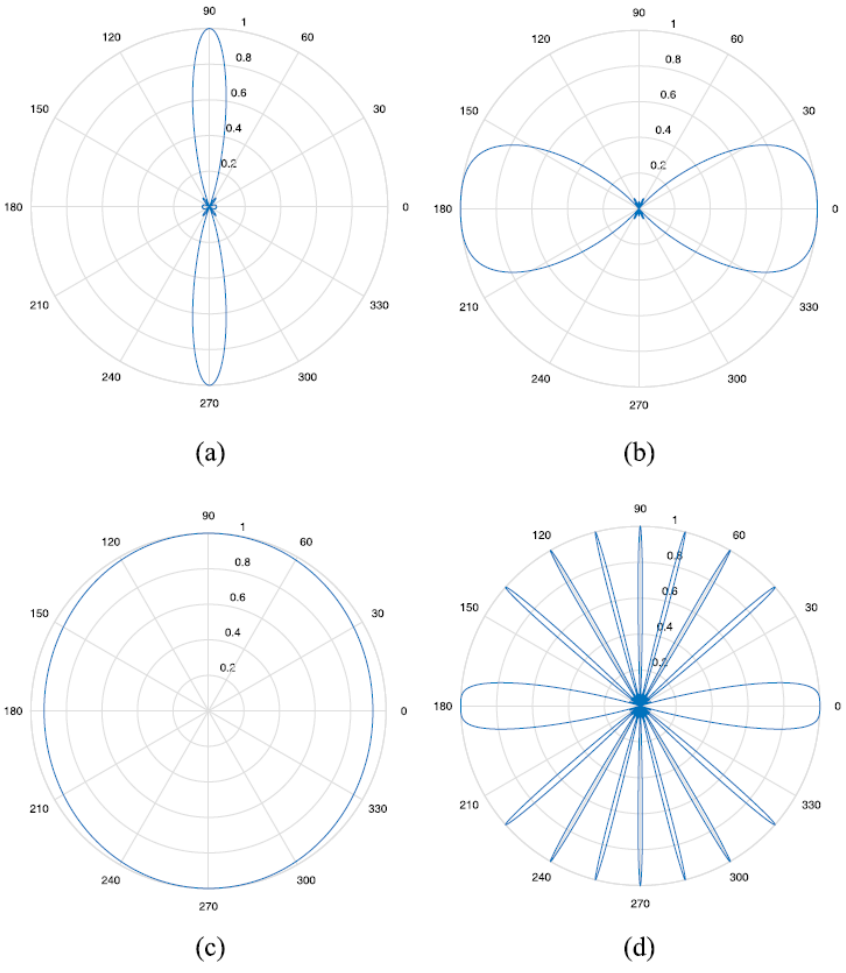
אם נקטין את זמן ההדהוד ל-*0.2ms* נקבל:



Spatial Filtering + Beamforming

הבעיה שאנו מתעסקים בה הינה בעיית עיבוד אותות מרחבי על אותות הנקלטים במערך (*array processing*). בעיבוד אותות מרחבי משתמשים במסננים מרחביים על מנת לסנן אותות מרחביים. מסנן לינארי מרחבי מוגדר על ידי פונקציית משקולות המשתנה בכל תדר, שניתן להפעיל אותה על התמרת *STFT* של אותות הנקלטים במערך (מיקרופונים למשל, אך יכול להיות גם מערך מסוג אחר). לאחר הפעלת המשקולות במישור התדר ניתן לחזור למישור הזמן על ידי התמרת *STFT* הפוכה.

מסנן כזה נקרא **מעצב אלומה** (*Beamformer*). מעצב אלומה נותן משקולות שונים לאותות הנמצאים בכיוונים שונים במרחב, ובכך מתבצע הסינון המרחבי. למשל ([2]):



המספרים מסביב לכל המעגל מייצגים זווית במעלות והמספרים בתוך כל מעגל מייצגים את המשקולת שנותן מעצב האלומה לאות מסוים ש"נפגש" עם מעצב האלומה ברדיוסים שונים ממרכז המעגל.

אנו נשתמש במעצב אלומה על מנת לסנן באופן מרחבי את אותות הדיבור הלא-רצויים וכן את הרעש, על ידי מתן משקולות בהתאם לרצוננו האם להנחית או לסנן.

LCMV – Linear Constrained Minimum Variance

כעת ניגש לפתור את הבעיה שלנו. נעשה זאת באמצעות מעצב אלומה שנקרא *LCMV*.

אנחנו מחפשים פונקציית משקולות $\boldsymbol{w}$ שנכפיל אותה באות שנקלט במיקרופונים $\boldsymbol{z}$ כך שננחית את הדובר הלא-רצוי וגם את הרעש.

במילים אחרות, אנו מחפשים פונקציה $\boldsymbol{w}$ כך שיתקיים:

$$\boldsymbol{w} = argmin\{\boldsymbol{w} * \boldsymbol{z}\ such\ that\ \boldsymbol{w}\left[\boldsymbol{h^d} s^d \ \ \boldsymbol{h^i} s^i\right] = [1 \ \ 0]\}$$

בכל תדר יהיו משקולות אחרות. התדרים בלתי תלויים ביניהם משום שהם אורתוגונליים.

את המשוואה שראינו לעיל שמתארת את המערכת ([1]):

$$\mathbf{z}(l,k) = \mathbf{h}^d(l,k)s^d(l,k) + \mathbf{h}^i(l,k)s^i(l,k) + \mathbf{v}(l,k)$$

ניתן לנרמל ולכתוב כך ([1]):

$$\mathbf{z}(l,k) = \mathbf{c}^d(l,k)\tilde{s}^d(l,k) + \mathbf{c}^i(l,k)\tilde{s}^i(l,k) + \mathbf{v}(l,k)$$

כאשר ([1]):

$$\mathbf{c}^d(l,k) = \left[\frac{h_1^d(l,k)}{h_{\text{ref}}^d(l,k)}, \frac{h_2^d(l,k)}{h_{\text{ref}}^d(l,k)}, \ldots, \frac{h_M^d(l,k)}{h_{\text{ref}}^d(l,k)}\right]^T$$

$$\mathbf{c}^i(l,k) = \left[\frac{h_1^i(l,k)}{h_{\text{ref}}^i(l,k)}, \frac{h_2^i(l,k)}{h_{\text{ref}}^i(l,k)}, \ldots, \frac{h_M^i(l,k)}{h_{\text{ref}}^i(l,k)}\right]^T$$

אלו הן פונקציות התמסורת היחסיות (*RTF – Relative Transfer Function*) של הדובר הרצוי והלא-רצוי בהתאמה. את ה-*RTF* מקבלים על ידי בחירת מערכת מסוימת להוות מערכת ייחוס, שבה מחלקים את כל המערכות לקבלת ה-*RTF*. המשמעות היא שה-*RTF* של המערכת היחסית תהיה תמיד 1.

לכן, אותות הדיבור המנורמלים של הדובר הרצוי והלא-רצוי הם, בהתאמה ([1]):

$$\tilde{s}^d(l,k) = h_{\text{ref}}^d(l,k)s^d(l,k)$$

$$\tilde{s}^i(l,k) = h_{\text{ref}}^i(l,k)s^i(l,k)$$

הסיבה שאנחנו משתמשים ב-*RTF* היא שכאמור איננו יודעים מהן המערכות $\boldsymbol{h}$ אך אנחנו יודעים שיטות שונות לשערוך של *RTF* (נדון בשתיים מהן בהמשך) ולכן נוח לנו לעבור לתצורת *RTF*.

סך הכל, נחפש פונקציה $\boldsymbol{w}$ כך שיתקיים:

$$\boldsymbol{w} = argmin\{\boldsymbol{w^H z}\ such\ that\ \boldsymbol{w}\left[\boldsymbol{c^d} s^d\ \ \boldsymbol{c^i} s^i\right] = [1\ \ 0]\}$$

כלומר אנחנו דורשים בעצם שיתקיים:

$$\boldsymbol{w^H C} = \boldsymbol{g}$$

כאשר:

$$\boldsymbol{C} = \left[\boldsymbol{c^d}\ \boldsymbol{c^i}\right], \boldsymbol{g} = [\mathbf{1}\ \mathbf{0}]$$

$\boldsymbol{g}$ קובע איזה מהדוברים אנחנו רוצים לשמוע ואיזה דובר רוצים לסנן. באותה מידה אפשר להחליף את הספרות בוקטור על מנת לשמוע את הדובר השני ולסנן את הדובר הראשון.

נזכיר שמטרתנו היא בנוסף גם להנחית את הרעש. נרצה שכאשר נפעיל את המשקולות על האותות הנקלטים במיקרופונים, נמזער את הרעש ככל שניתן. נמדוד את מזעור הרעש בעזרת שיקולי אנרגיה. נרצה שהאנרגיה של הרעש לאחר שנפעיל את המשקולות תהיה קטנה ככל הניתן. נסתכל על התוחלת:

$$E[|\boldsymbol{w^H v}|^2] = E[\boldsymbol{w^H v}(\boldsymbol{w^H v})^{\boldsymbol{H}}] = E[\boldsymbol{w^H v v^H w}] = \boldsymbol{w^H} E[\boldsymbol{v v^H}]\boldsymbol{w} = \boldsymbol{w^H \Phi_{vv} w}$$

כדי שתוחלת זו תמוזער נדרוש ([4]):

$$\mathbf{w}(l,k) = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \mathbf{w}^H(l,k) \Phi_{vv} \mathbf{w}(l.k) \right\}$$

וקיבלנו בעיית אופטימיזציה: עלינו למצוא $\boldsymbol{w}$ שממזער את הביטוי $\boldsymbol{w^H \Phi_{vv} w}$ תחת האילוץ $\boldsymbol{w^H C} = \boldsymbol{g}$. כזכור, בעיית האופטימיזציה תיפתר עבור כל תדר בנפרד.

כעת נפתור את בעיית האופטימיזציה. נפתור אותה באמצעות שיטת כופלי לגראנז'. נבנה את הלגרנז'יאן:

$$L(\boldsymbol{w}) = \boldsymbol{w^H \Phi_{vv} w} + \boldsymbol{\lambda^H}(\boldsymbol{c^H w} - \boldsymbol{g})$$

נגזור לפי $\boldsymbol{w}$ ולפי $\boldsymbol{\lambda}$ ונמצא מינימום. יש לנו שתי משוואות:

$$\frac{d}{d\boldsymbol{w}} L(\boldsymbol{w}) = \boldsymbol{w^H \Phi_{vv}} + \boldsymbol{\lambda^H c^H} = \mathbf{0}$$

$$\frac{d}{d\boldsymbol{\lambda}} L(\boldsymbol{w}) = \boldsymbol{c^H w} - \boldsymbol{g} = \mathbf{0}$$

מטריצת הקורלציה של הרעש בין המיקרופונים הינה סימטרית (והדבר הגיוני, כי אין זה משנה הסדר שבו אנו שואלים מה הקורלציה בין מיקרופון x למיקרופון y), לכן ניתן למצוא לה מטריצה הופכית.

נכפיל את המשוואה הראשונה בביטוי $\boldsymbol{\Phi_{vv}^{-1}c}$ מימין ונקבל:

$$\boldsymbol{w^H\Phi_{vv}\Phi_{vv}^{-1}c + \lambda^H c^H\Phi_{vv}^{-1}c = w^H c + \lambda^H c^H\Phi_{vv}^{-1}c = 0}$$

נציב את המשוואה השנייה:

$$\boldsymbol{w^H c = g^H \Rightarrow g^H + \lambda^H c^H\Phi_{vv}^{-1}c = 0}$$

נכפיל את שני אגפי המשוואה שקיבלנו בביטוי $\boldsymbol{(c^H\Phi_{vv}^{-1}c)^{-1}}$ מימין ונקבל:

$$\boldsymbol{g^H(c^H\Phi_{vv}^{-1}c)^{-1} + \lambda^H c^H\Phi_{vv}^{-1}c(c^H\Phi_{vv}^{-1}c)^{-1} = g^H(c^H\Phi_{vv}^{-1}c)^{-1} + \lambda^H = 0}$$

$$\boldsymbol{\Rightarrow \lambda^H = -g^H(c^H\Phi_{vv}^{-1}c)^{-1}}$$

וכעת ניתן להציב במשוואה הראשונה לקבלת:

$$\boldsymbol{w^H\Phi_{vv} + \lambda^H c^H = w^H\Phi_{vv} - g^H(c^H\Phi_{vv}^{-1}c)^{-1}c^H = 0}$$

$$\boldsymbol{\Rightarrow w^H\Phi_{vv} = g^H(c^H\Phi_{vv}^{-1}c)^{-1}c^H \Rightarrow w^H = g^H(c^H\Phi_{vv}^{-1}c)^{-1}c^H\Phi_{vv}^{-1}}$$

לכן קיבלנו סה"כ ([1]):

$$\mathbf{w}_{\mathrm{LCMV}}(l,k) = \Phi_{vv}^{-1}(k)\mathbf{C}(l,k)\cdot \left[\mathbf{C}^H(l,k)\Phi_{vv}^{-1}(k)\mathbf{C}(l,k)\right]^{-1}\mathbf{g}(l,k)$$

לאחר שנמצא את המשקולות הללו נוכל לקבל את הדובר הרצוי על ידי ([1]):

$$\hat{\mathbf{s}}^d(l,k) = \mathbf{w}_{\mathrm{LCMV}}^H(l,k)\mathbf{z}(l,k)$$

הביטוי $\boldsymbol{\Phi_{vv}}$ מסמל את מטריצת הקורלציה של הרעש אותה נשערך על ידי ([4]):

$$\hat{\Phi}_{vv}(k) = \frac{1}{l_v^{\mathrm{stop}} - l_v^{\mathrm{start}}} \sum_{l=l_v^{\mathrm{start}}}^{l_v^{\mathrm{stop}}-1} \mathbf{z}(l,k)\mathbf{z}^H(l,k)$$

כעת נדון בשתי דרכים לשערך את וקטורי ה-*RTF*:

הפחתה ספקטרלית – *Spectral Substruction*

נוכל לכתוב את האותות המתקבלים במיקרופונים בצורה הבאה באופן כללי:

$$\boldsymbol{z = c * s + v}$$

כאשר כאן $\boldsymbol{s}$ סומן כוקטור המכיל את שני הקולות משני הדוברים, אך נזכיר כי הוא דטרמיניסטי ואינו משתנה ממיקרופון למיקרופון. להלן, כשנדבר על הערכים במיקרופון הראשון, אין צורך לשים אינדקס מתחת לאות הדיבור משום כך שהוא לא תלוי במיקרופון.

אם נבחר את המיקרופון הראשון להיות מיקרופון הרפרנס, נקבל:

$$\boldsymbol{c_1 = 1}$$

לכן נרשום את האות הנקלט במיקרופון הראשון כך לקבלת:

$$\boldsymbol{z_1 = c_1 * s + v_1 = s + v_1 \Longrightarrow s = z_1 - v_1}$$

נציב במשוואה לעיל:

$$\boldsymbol{z = c * s + v = c(z_1 - v_1) + v}$$

נכפיל את שני האגפים בביטוי $\boldsymbol{z^H}$ מימין ונקבל:

$$\boldsymbol{zz^H = c(z_1 - v_1)z^H + vz^H}$$

נחשב תוחלת על שני האגפים:

$$E[\boldsymbol{zz^H}] = \boldsymbol{c}(E[\boldsymbol{z_1 z^H}] - E[\boldsymbol{v_1 z^H}]) + E[\boldsymbol{vz^H}]$$

אות הדיבור והרעש הינם חסרי קורלציה. בנוסף, הרעשים בין המיקרופונים הינם בלתי תלויים בינם לבין עצמם. לכן מתקיים:

$$E[\boldsymbol{vz^H}] = E[\boldsymbol{vv^H}]$$

$$E[\boldsymbol{v_1 z^H}] = E[\boldsymbol{v_1 v^H}]$$

נציב ונקבל:

$$E[\boldsymbol{zz^H}] = \boldsymbol{c}(E[\boldsymbol{z_1 z^H}] - E[\boldsymbol{v_1 v^H}]) + E[\boldsymbol{vv^H}]$$

$$\Longrightarrow \boldsymbol{\Phi_{zz} = c\left(\Phi_{zz_1} - \Phi_{vv_1}\right) + \Phi_{vv}}$$

$$\Longrightarrow \boldsymbol{\Phi_{zz} - \Phi_{vv} = c\left(\Phi_{zz_1} - \Phi_{vv_1}\right)}$$

כאשר Φ מסמל קורלציה ספקטרלית (משום שאנו מדברים במישור התדר).

כעת איך מהמשוואה נוכל לחלץ את $\boldsymbol{c}$?

נגדיר:

$$A = \boldsymbol{\Phi_{zz}} - \boldsymbol{\Phi_{vv}}$$

נסמן את השורה הראשונה של המטריצה A בסימון $\boldsymbol{a_1}$.

נוכל לקבל את $\boldsymbol{c}$ על ידי:

$$\boldsymbol{c} = A * \boldsymbol{a_1^H}/(\boldsymbol{a_1 a_1^H})$$

הערה: בספרות לעיתים השיטה מכונה שיטת *"Covariance Substruction"* ([5]), זאת משום שאנו מחסירים שתי פונקציות קולרציה במישור התדר, דהיינו מחסירים פונקציות קוואריאנס.

GEVD – Generalized Eigenvalue Decomposition

נציג כעת שיטה נוספת לשערוך וקטורי פונקציות התמסורת היחסיות.

נתבונן בצורת הכתיבה שהצגנו לעיל:

$$\boldsymbol{z} = \boldsymbol{h} * \boldsymbol{s} + \boldsymbol{v}$$

כפי שנראה בסימולציית המטלב בהמשך, המשתנה $\boldsymbol{z}$ הינו תלת מימדי. המימד הראשון שלו מסמל את מספר המיקרופון, המימד השני מסמל את אינדקס התדר k והמימד השלישי מסמל את אינדקס הפריימים l.

כזכור אנו רוצים לשערך את פונקציות ה-*RTF* עבור כל תדר בנפרד. עבור תדר מסויים שנקבע k, המשתנה $\boldsymbol{z}$ הינו מטריצה בגודל $M \times l_{max}$ כאשר M זה מספר המיקרופונים ו-l_{max} זהו מספר הפריימים בהתמרת ה-*STFT*.

נניח כעת שהרעש $\boldsymbol{v}$ הינו רעש לבן.

נתבונן בקורלציה $E[\boldsymbol{z}\boldsymbol{z}^{\boldsymbol{H}}]$ שהינו מטריצה בגודל $M \times M$:

$$\boldsymbol{z}\boldsymbol{z}^{\boldsymbol{H}} = \boldsymbol{h}\boldsymbol{s}\boldsymbol{s}^{\boldsymbol{H}}\boldsymbol{h}^{\boldsymbol{H}} + \boldsymbol{v}\boldsymbol{s}^{\boldsymbol{H}}\boldsymbol{h}^{\boldsymbol{H}} + \boldsymbol{h}\boldsymbol{s}\boldsymbol{v}^{\boldsymbol{H}} + \boldsymbol{v}\boldsymbol{v}^{\boldsymbol{H}}$$

נחשב תוחלת כאשר נתחשב בעובדה שהרעש בלתי תלוי בדיבור ובמערכות:

$$E[\boldsymbol{z}\boldsymbol{z}^{\boldsymbol{H}}] = \boldsymbol{h}E[\boldsymbol{s}\boldsymbol{s}^{\boldsymbol{H}}]\boldsymbol{h}^{\boldsymbol{H}} + E[\boldsymbol{v}\boldsymbol{v}^{\boldsymbol{H}}]$$

הרעש הינו לבן ולכן ניתן לסמן:

$$E[\boldsymbol{v}\boldsymbol{v}^{\boldsymbol{H}}] = \boldsymbol{\sigma}_v I$$

כאשר I מטריצת היחידה בגודל $M \times M$.

לכן:

$$E[\boldsymbol{z}\boldsymbol{z}^{\boldsymbol{H}}] = R_s \boldsymbol{h}\boldsymbol{h}^{\boldsymbol{H}} + \sigma_v I$$

כאשר R_s קבוע.

למטריצה $\boldsymbol{h}\boldsymbol{h}^{\boldsymbol{H}}$ בגודל $M \times M$ קיים פירוק:

$$\boldsymbol{h}\boldsymbol{h}^{\boldsymbol{H}} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-\boldsymbol{1}}$$

כאשר P מטריצה הפיכה ו-D מטריצה אלכסונית. לכן ניתן לרשום:

$$E[\boldsymbol{z}\boldsymbol{z}^{\boldsymbol{H}}] = R_s \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-\boldsymbol{1}} + \sigma_v \boldsymbol{P}\boldsymbol{P}^{-\boldsymbol{1}} = \boldsymbol{P}(R_s \boldsymbol{D} + \sigma_v I)\boldsymbol{P}^{-\boldsymbol{1}}$$

ל-$\boldsymbol{h}$ יש ערך עצמי אחד ואילו ל-I יש M ערכים עצמיים.

כיצד נמצא את הוקטור העצמי של $\boldsymbol{h}$?

נתבונן במטריצה האלכסונית $R_s\boldsymbol{D} + \sigma_v I$. העמודה שבה נמצא הערך הכי גדול של האלכסון, מתאימה לעמודה ב-P שמתארת את הוקטור העצמי שיוצר את $\boldsymbol{h}$.

לאחר שנמצא את כל הוקטורים העצמיים עבור כל תדר, נוכל לחלק בתוצאה עבור מיקרופון הרפרנס לקבלת ה-*RTF*.

לאחר שנבצע זאת עבור שני הדוברים, נקבל שתי פונקציות תמסורת יחסיות עבור כל אחד מהם, שמשתנות בתדר ובמיקרופון. נוכל להציב פונקציות תמסורת יחסיות אלו בנוסחה שמצאנו לעיל למשקולות ה-*LCMV*.

חשוב לשים לב שהאלגוריתם לעיל תקף רק כאשר מסתכלים על סגמנט שבו רק דובר אחד הינו אקטיבי. לכן בסימולציה אנו נשלח לפונקציית ה-*GEVD* את האותות $\boldsymbol{z}$ במקטעים מסוימים שבהם אנו יודעים שרק דובר אחד מדבר. לכן אנו נקרא פעמיים לפונקציה עבור כל דובר. הנחה זו שקיימים מקטעים שבהם רק דובר אחד מדבר הינה נפוצה מאוד ומציאותית, שהרי בשיחה (תרבותית), הדוברים ידברו אחד אחרי השני.

כעת, מה קורה במידה והרעש **אינו** לבן?

כעת, מכיוון שהרעש כבר אינו לבן, לא נוכל להסיק מסקנות מהאלכסון של המטריצה. לכן, נשתמש בשיטת הלבנת רעש, שתלבין לנו את הרעש. נמצא מטריצה כלשהי שהפעלתה על האותות הנקלטים במיקרופונים $\boldsymbol{z}$ תלבין את הרעש ואז נוכל לבצע את האלגוריתם הנ"ל על האות המתקבל כתוצאה מההכפלה.

נסמן את קורלציית הרעש ב-Φ_{vv}. קיים פירוק למטריצת קורלציה זו כך ([5]):

$$\boldsymbol{\phi}_{vv} = \boldsymbol{\Phi}_{vv}^{\boldsymbol{H/2}}\boldsymbol{\phi}_{vv}^{\boldsymbol{1/2}}$$

פירוק זה נקרא "פירוק שורש" (*Square-Root Decomposition*) והוא מקרה פרטי של פירוק חולסקי (*Cholesky*).

מתקיים הקשר (נובע משיטות סימון) ([5]):

$$\boldsymbol{\Phi}_{vv}^{\boldsymbol{H/2}} = \left(\boldsymbol{\phi}_{vv}^{\boldsymbol{1/2}}\right)^{\boldsymbol{H}}$$

האות המולבן נתון על ידי $\boldsymbol{\Phi}_{vv}^{\boldsymbol{-H/2}} * \boldsymbol{z}$. כעת, הרעש של התוצאה הינו לבן עם שונות קבועה, ולכן נוכל להפעיל על האות שנקבל את האלגוריתם הנ"ל למציאת פונקציות ה-*RTF* של כל אחד מהדוברים.

הערה: שיטה זו נקראת לעיתים גם *Covariance Whitening*.

סימולציות מטלב ותוצאות

תיאור מהלך התוכנית

קטע הקול שאיתו נעבוד מורכב מדיבור של דובר אחד, דיבור של דובר שני ודיבור מעורבב של שני הדוברים ביחד. בין כל קטע תהיה הפסקה קצרה.

את הדוברים ואת שני קטעי הדיבור לכל דובר נגריל מתוך ספריית ה-*TIMIT*. לאחר מכן ניצור שני קטעי קול של כל דובר בנפרד, כך שכשנחבר ביניהם נקבל את קטע הקול שאותו תיארנו לעיל.

לאחר שהגרלנו את הדוברים ויצרנו את קטעי הקול, ניצור את החדר שבו יהיו הדוברים עם המיקרופונים.

נגריל את אורך ורוחב החדר להיות בין 3 ל-6 מטרים ונקבע את גובה החדר להיות 3 מטרים. מספר הדגימות שנכניס ל-*RIR Generator* יהיה 4096, ומקדם ההדהוד יהיה:

$$\beta = t_{60} = 0.3ms$$

(בהמשך נראה כיצד שינוי פרמטר זה ישפיע על ביצועי ה-*LCMV*).

נמקם את הדוברים בחדר כך שהם יהיו במרחק של לפחות מטר מהמיקרופונים, לפחות חצי מטר מהקיר, לפחות חצי מטר אחד מהשני ובגובה 1.2 מטרים.

נמקם את מערך 7 המיקרופונים במרכז החדר (בקירוב) ובגובה 1.2 מטרים. המרחק בין כל מיקרופון במערך הוא כ-5 סנטימטרים.

לאחר שנפעיל את פונקציות ה-*RIR Generator* נקבל 14 מערכות שונות עבור כל דובר ועבור כל מיקרופון.

לאחר שיצרנו את פונקציות התגובות להלם, נבצע קובולוציות בין כל דובר לבינן ונקבל את הדוברים המסוננים. לאחר שנשלב בין הדוברים נקבל את הקטע שאיתו נעבוד.

נוסיף רעש לבן כך שיחס האות לרעש יהיה $SNR = 20dB$ (בהמשך נראה כיצד שינוי פרמטר זה ישפיע על ביצועי ה-*LCMV*).

אחרי שקיבלנו את האות המורעש נחשב את התמרת ה-*STFT* כאשר אורך כל מקטע יהיה 512 דגימות עם חפיפות של 75% בין המקטעים. בהתמרה נשתמש בחלון *hamming* מחזורי. מכיוון שהאותות שלנו הינם ממשיים בזמן, כל החישובים וההצגות במישור התדר יתבצעו מ-$k = 1$ עד $k = 257$ (אין אינדקס 0 במטלב), זאת משום שהתמרה של אות ממשי הינה סימטרית במישור התדר ולכן ניתן להסתכל ולפעול רק על חצי מהתדרים כדי לקבל את כל המידע הדרוש על האות.

נחשב את קורלציית הרעש. ניקח את 50 הפריימים הראשונים בהם אנו יודעים בוודאות שאין דיבור (משום שיש 20000 דגימות במישור הזמן לפני שמתחיל קטע הדיבור הראשון) לשם חישוב הקורלציה. בנוסף נחשב לקורלציית הרעש פירוק חולסקי בו נוכל להשתמש, אמנם אין צורך לכתחילה להשתמש בפירוק חולסקי אם הרעש הוא לבן, כזכור.

לשם ביצוע פעולת ה-*GEVD* צריך למצוא את הפריימים בהם כל דובר מדבר בנפרד. לשם כך כתבנו פונקציה שמקבלת כל דובר בנפרד ללא רעש ומוצאת את הפריימים בהם הדובר אקטיבי, זאת על ידי קביעת סף מסוים (קטן) על עוצמת האות שנשלח לפונקציה.

לאחר מכן, עבור כל תדר, נשלח את התמרת ה-*STFT* של האות הכולל עם הרעש, את הפריימים שמצאנו לעיל ואת פירוק החולסקי של קורלציית הרעש לפונקציית ה-*GEVD* שכתבנו, בה נמצא את הפירוק למטריצה הפיכה ואלכסונית, ונמצא את הערכים העצמיים המתאימים לוקטורים העצמיים שייצרו את פונקציות ה-*RTF* שלנו.

עבור שיטת הפחתה ספקטרלית נחשב עבור כל דובר את הקורלציה שלו ועבור כל תדר נפחית את הקורלציה של הרעש מהקורלציה של הדובר ונחשב את המטריצה *A* ונבצע את החישוב אותו פירטנו לעיל למציאת וקטורי ה*RTF*.

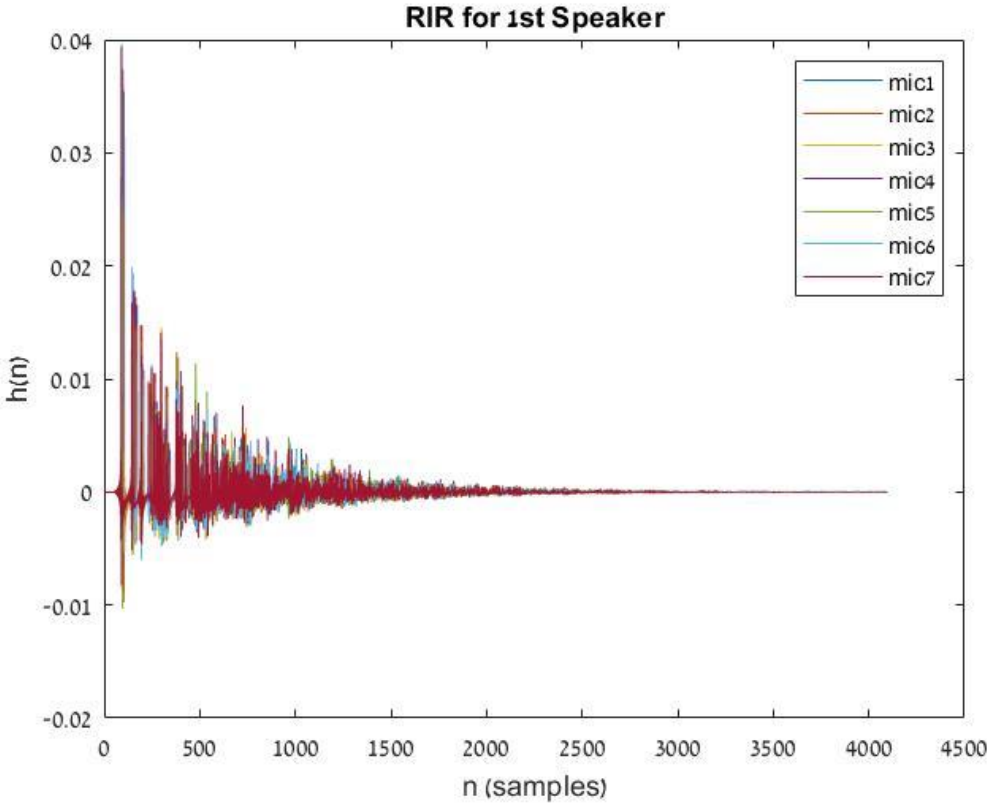
לבסוף, לאחר שקיבלנו את ה-*RTF* נוכל להציב אותו ואת קורלציית הרעש בנוסחה למשקולות ה-*LCMV* עבור כל תדר. לאחר מכן נכפיל את המשקולות בהתמרת ה-*STFT* של האות המשולב של שני הדוברים עם הרעש, ולבסוף נעשה התמרת *STFT* הפוכה על מנת לראות את התוצאות במישור הזמן.

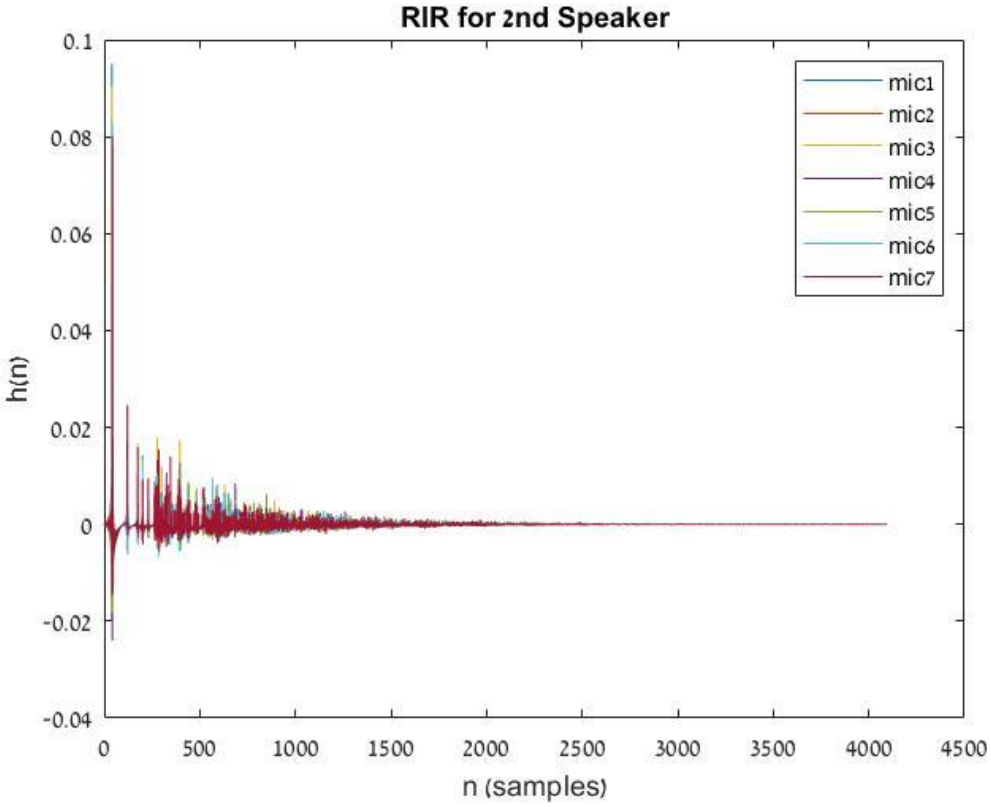
בנוסף, כאשר רצינו להפוך מטריצות הוספנו קצת אנרגיה לאלכסון המטריצה על מנת למנוע בעיות בעת הפיכת המטריצה.

כדי למדוד הנחתה, חישבנו את היחס (הפרש ב-*dB*) בין ממוצעי ההספקים של האותות המתקבלים במיקרופונים ואותות התוצאה של כל דובר. מדדנו את ההנחתה במקומות שבהם מדבר דובר אחד בלבד, וכמו כן מדדנו את הנחתת הרעש במקומות שאין דוברים כלל.

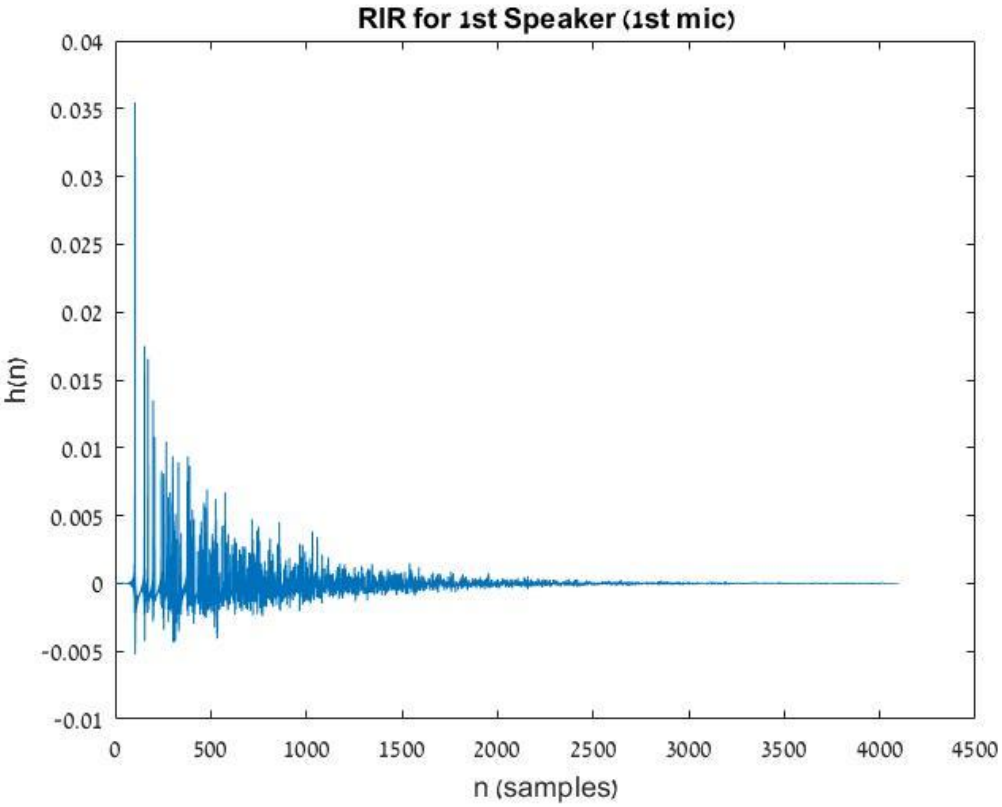
<u>תוצאות</u>

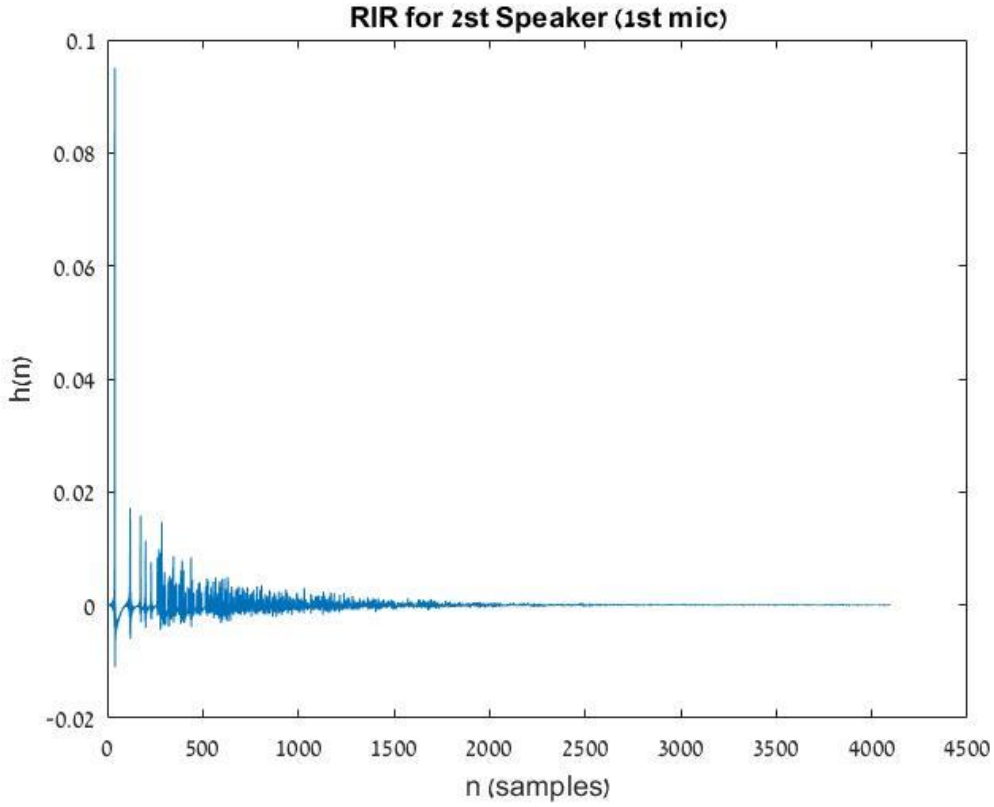
הגרלנו דוברים מה-*TIMIT* ויצרנו את המערכות:



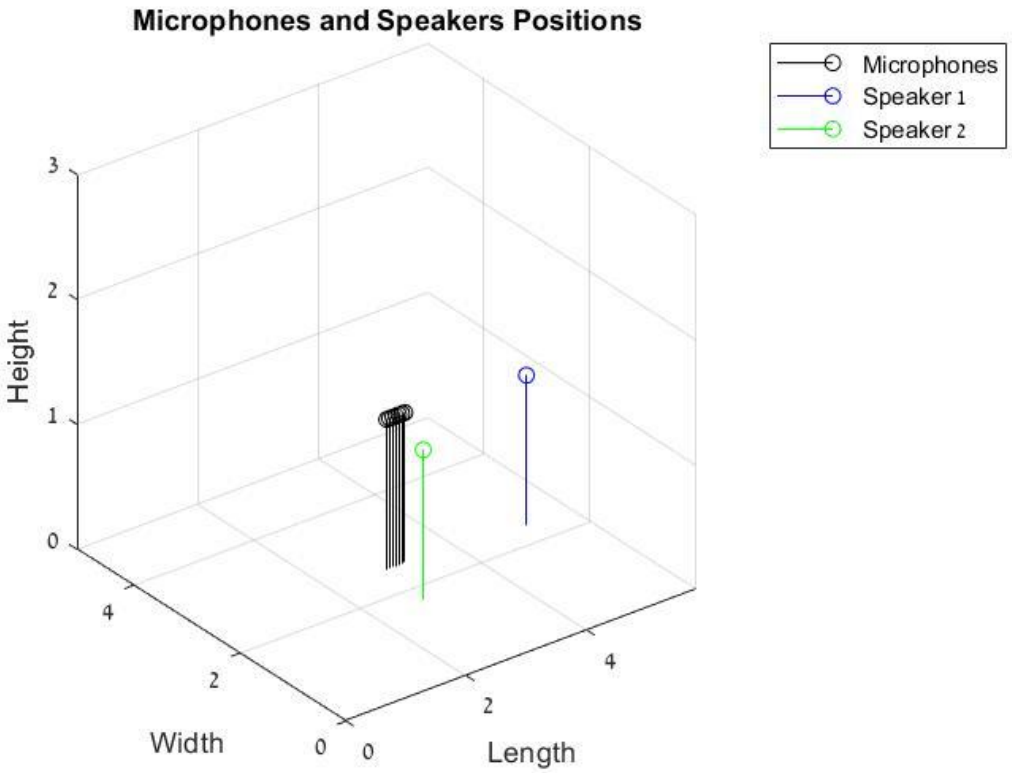


עבור מיקרופון אחד בלבד (הראשון), התגובות להלם נראות כך:





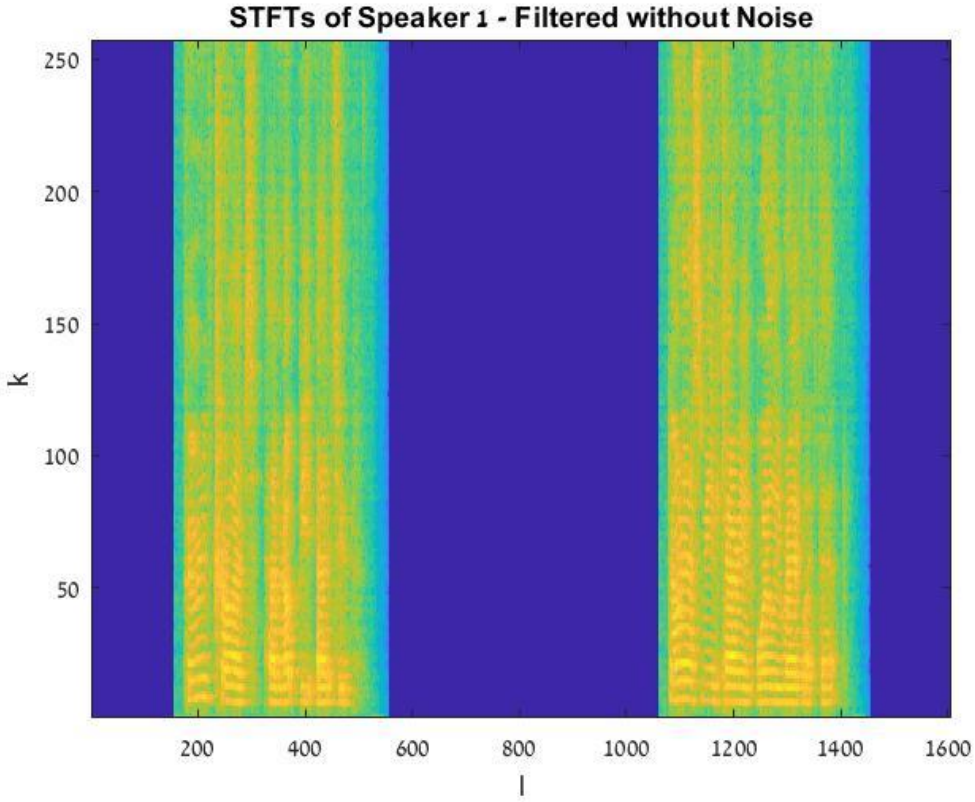
החדר נראה כך:

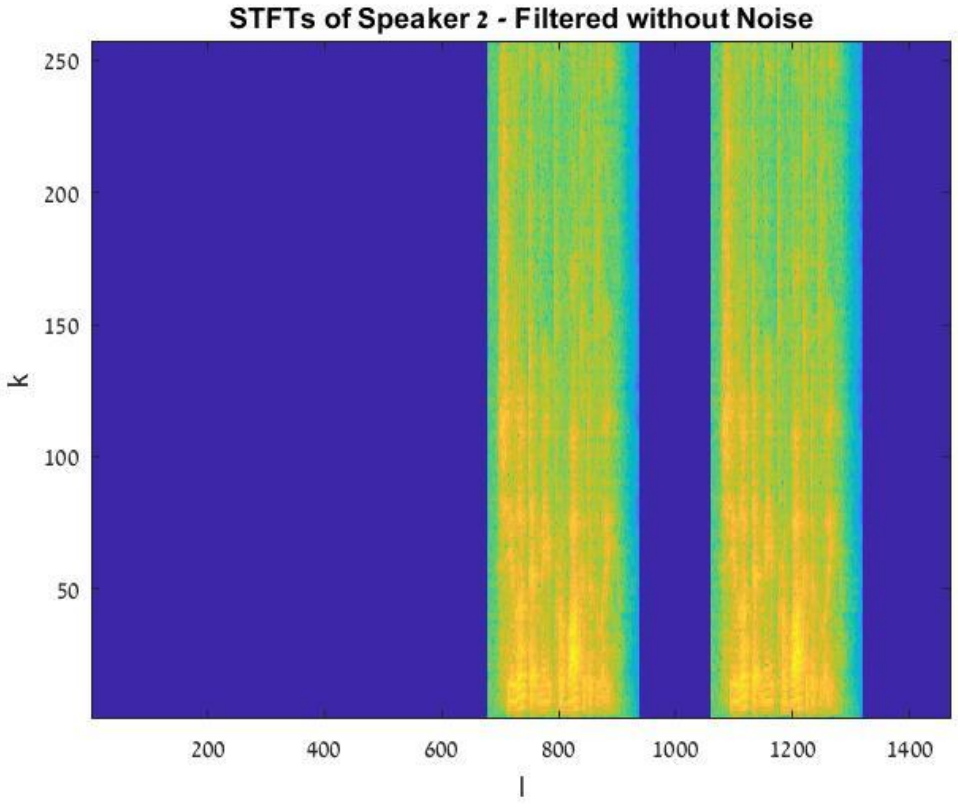


ניתן לראות את הדוברים ואת מערך המיקרופונים מסומנים בתוך החדר. כזכור, הדוברים הינם סטטיים בתוך החדר ואינם זזים.

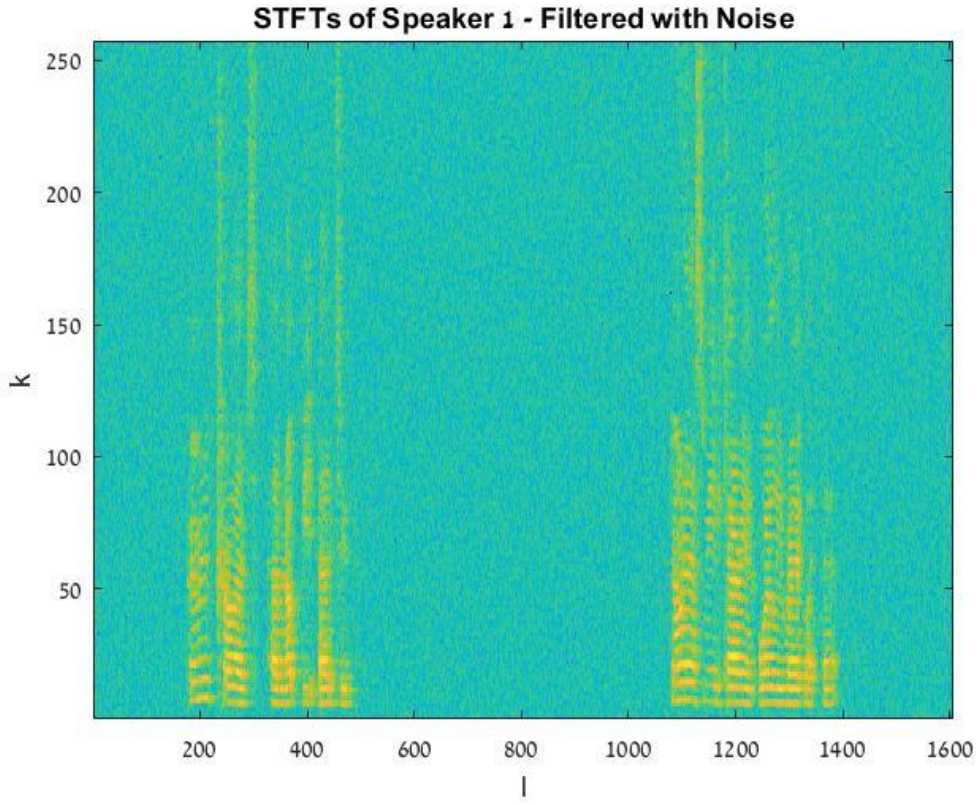
ניתן לראות בגרפים לעיל כי התגובות להלם של הדובר הראשון יותר "מהדהדות" מהתגובות להלם של הדובר השני. הדבר נובע מהעובדה שהדובר השני נמצא קרוב יותר למיקרופונים. משום כך, גם ה-*direct path* של התגובות להלם של הדובר השני חזק יותר, וגם יש פחות הדהוד. הקול של הדובר השני מגיע למערך המיקרופונים הרבה לפני שהוא מגיע אל הקירות שיחזירו את הקול וייצרו הדהודים בחדר, זאת בניגוד לדובר הראשון שהינו קרוב יותר אל הקיר ולכן בהתחלה יהיו הרבה יותר הדהודים.

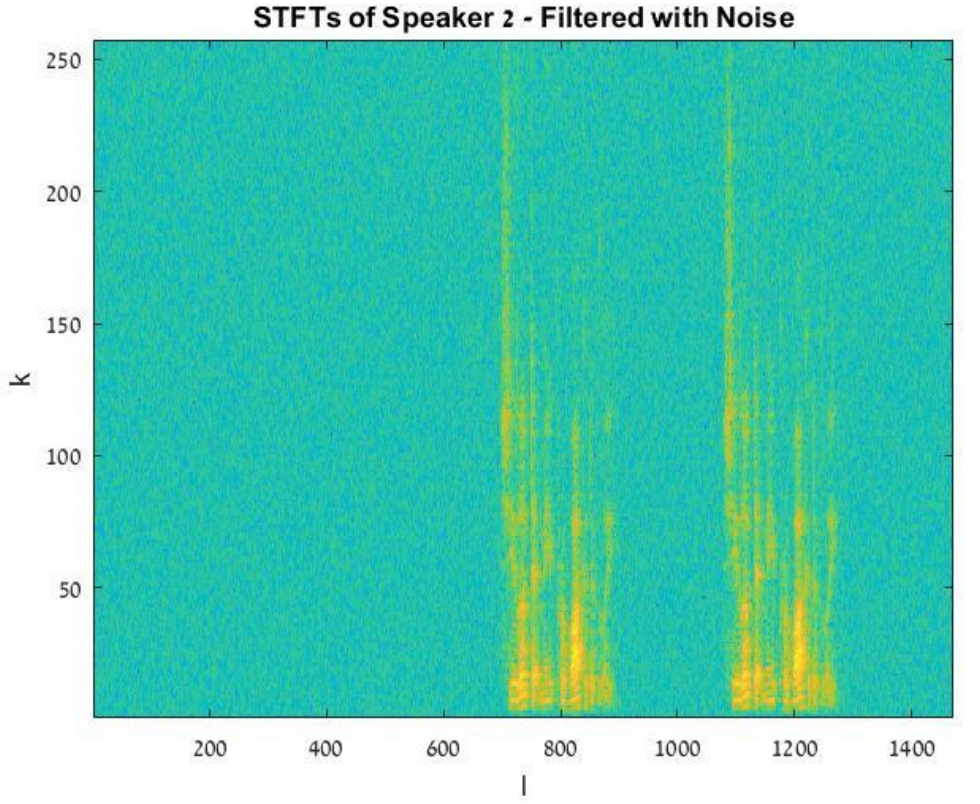
התמרות ה-*STFT* של הדוברים במיקרופון הראשון בנפרד לפני שנוסף הרעש:



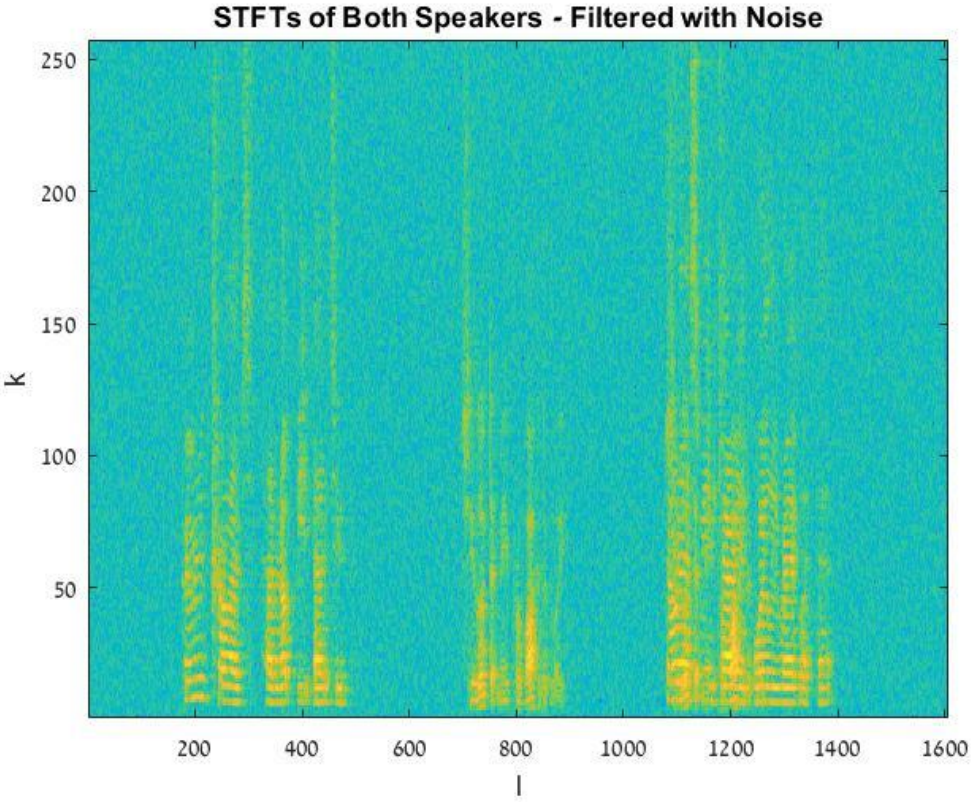


לאחר שנוסיף רעש לבן גאוסי כך שיתקיים $SNR = 20dB$, נקבל את ההתמרות (במיקרופון הראשון):





ולאחר שנאחד את שני הדוברים עם הרעש נקבל את קטע הקול שנקלט במערך המיקרופונים:

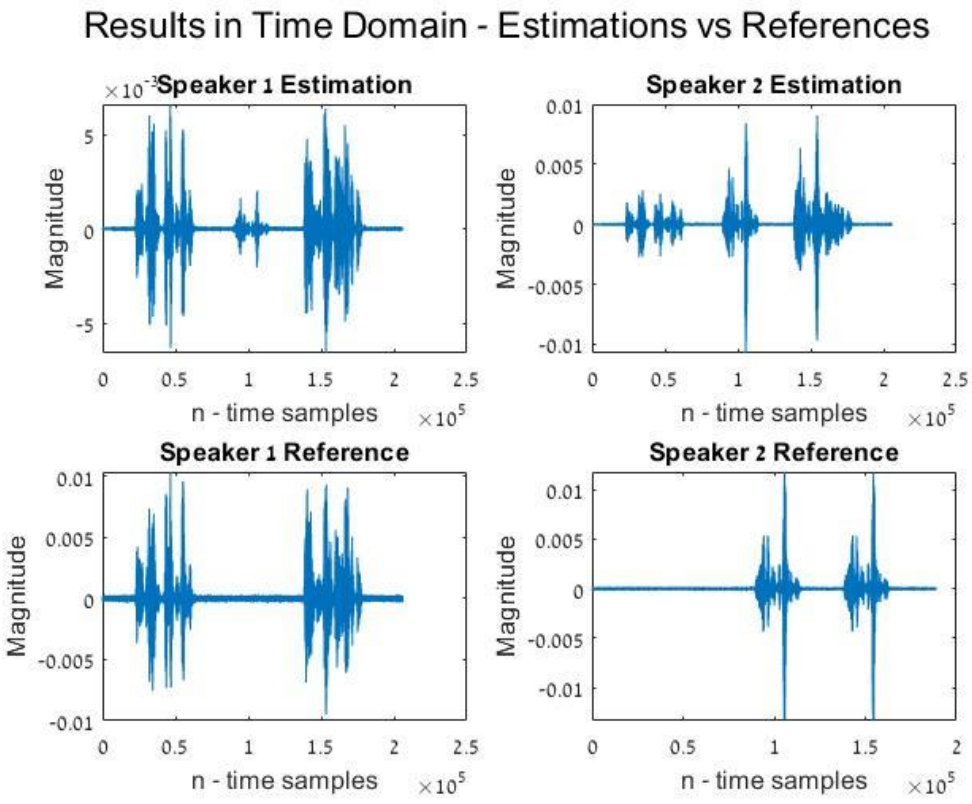


בקטע הראשון מדבר הדובר הראשון בלבד, בקטע השני מדבר הדובר השני בלבד ובקטע השלישי שני הדוברים מדברים בו-זמנית.

כעת נשערך את פונקציות ה-*RTF* בשיטת *GEVD*. לאחר מכן נשערך בעזרת הפחתה ספקטרלית.

הגרפים הבאים שנציג מתארים את תוצאות ה-*LCMV*. נציג כמה גרפים שונים. הגרף הראשון יראה את תוצאות ה-*LCMV* עבור כל דובר בהשוואה לגרף של הדובר כאשר הוא מדבר לבד ומסונן במערכת עבור אחד המיקרופונים (הראשון). לאחר מכן נציג גרפים של כל דובר בהשוואה לאות הנקלט במיקרופון, "נלביש" את הגרפים אחד על השני ונראה את ההנחתה שהתבצעה בזכות האלגוריתם.

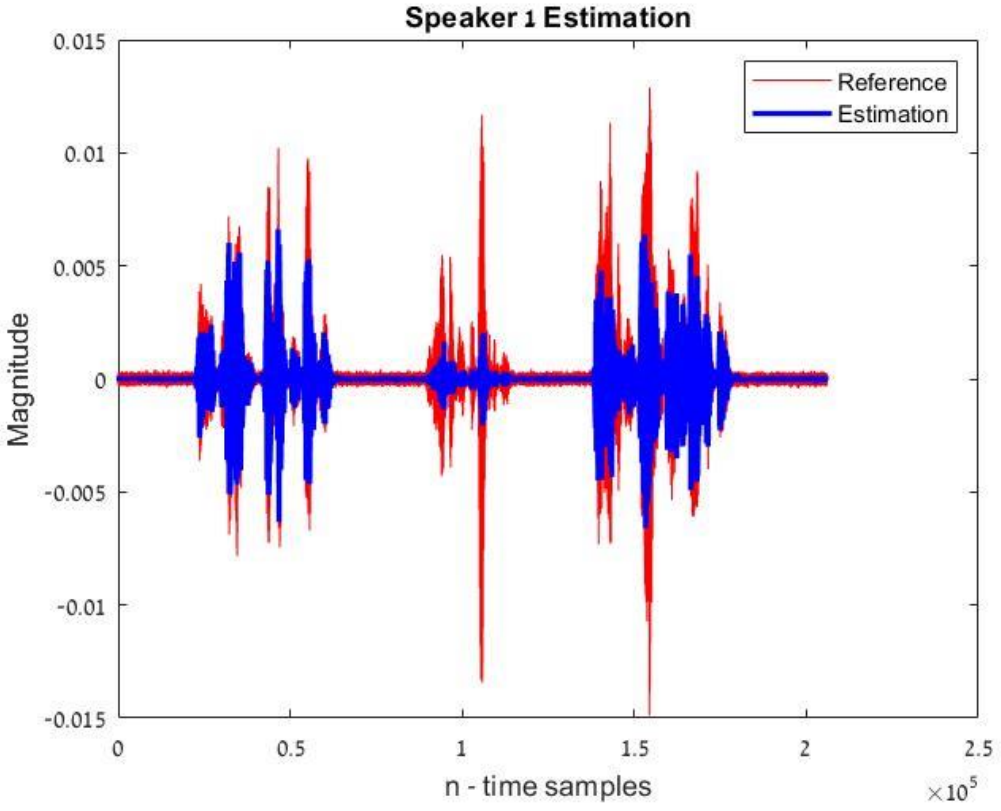
לאחר חישוב משקולות ה-*LCMV* והכפלתן באות z נקבל:

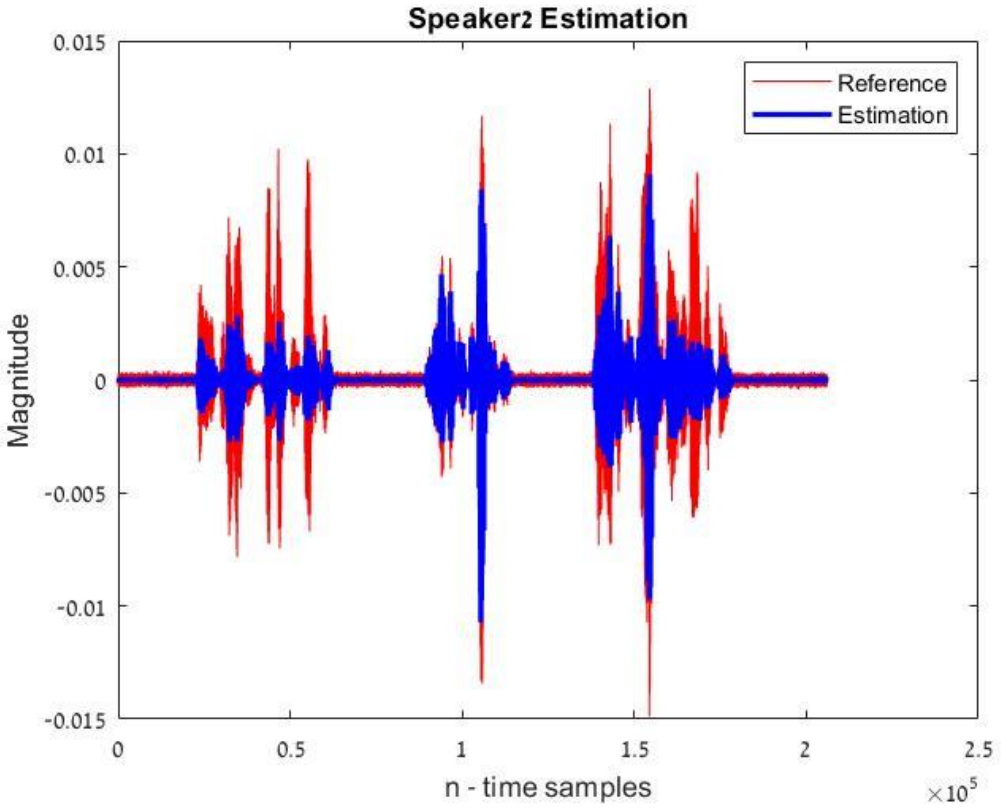


ניתן לראות כי התבצעה הנחתה של הדובר הלא-רצוי בכל אחד מהמקרים. בנוסף, ניתן לראות שהרעש גם הונחת. "עובי" הרעש (שמייצג בעצם את השונות שלו) לאחר ה-*LCMV* קטן יותר.

ניתן גם לראות (ולשמוע) כי שערוך הדובר הראשון התבצעה בצורה טובה יותר משערוך הדובר השני. ניתן לראות זאת גם לפי השוואת עוצמות הקול במקטעים שבהם הדוברים מדברים לבד, וגם לפי זה שחלק מהדיבור השני של הדובר הראשון (כלומר קטע קול של הדובר הראשון שהוא מדבר ביחד עם הדובר השני) נשאר בתוך שערוך הדובר השני בסוף קטע הקול של שערוך הדובר השני. דבר זה מראה שהאלגוריתם לא עובד בצורה מושלמת, אך הוא כן מספק לנו הנחתה שאפשר לתת עליה את הדעת.

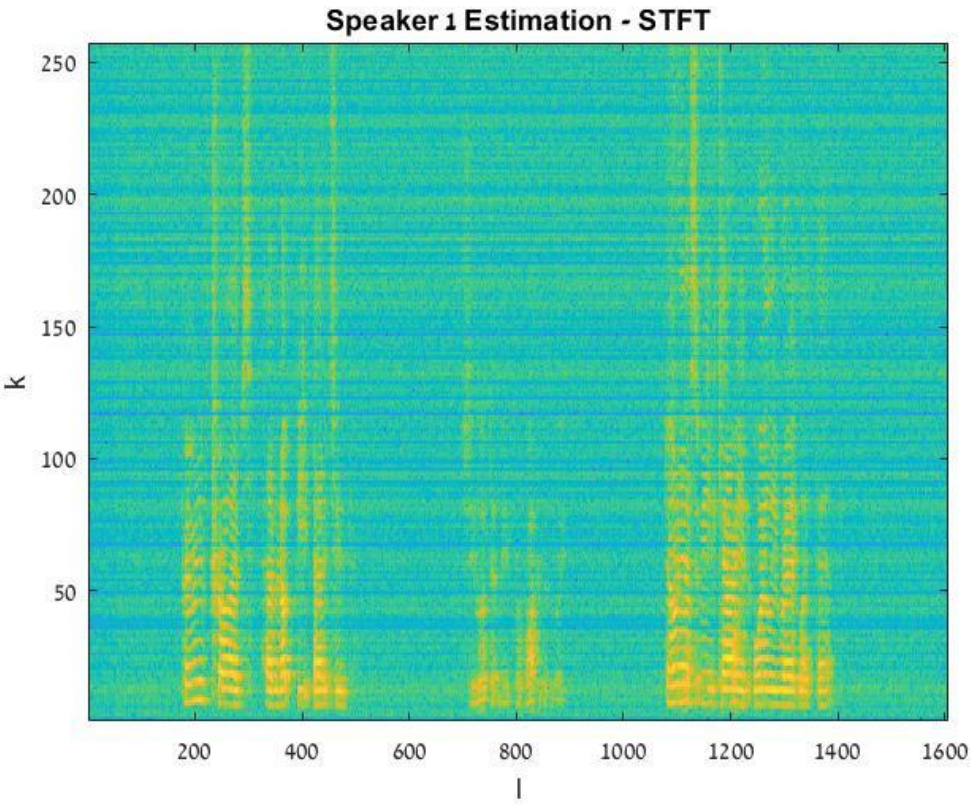
תוצאות ה-*LCMV* ביחס לאותות שנקלטו במיקרופונים:

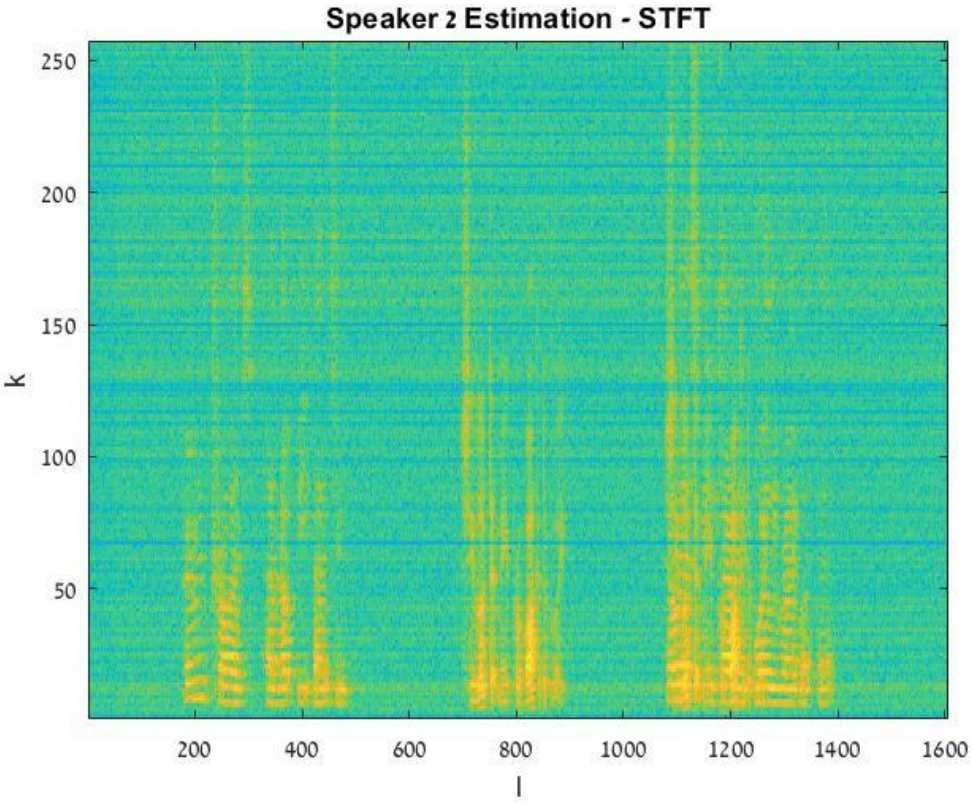




ניתן לראות שאכן התבצעה הנחתה, בעיקר במקומות כאשר הדוברים מדברים בנפרד.

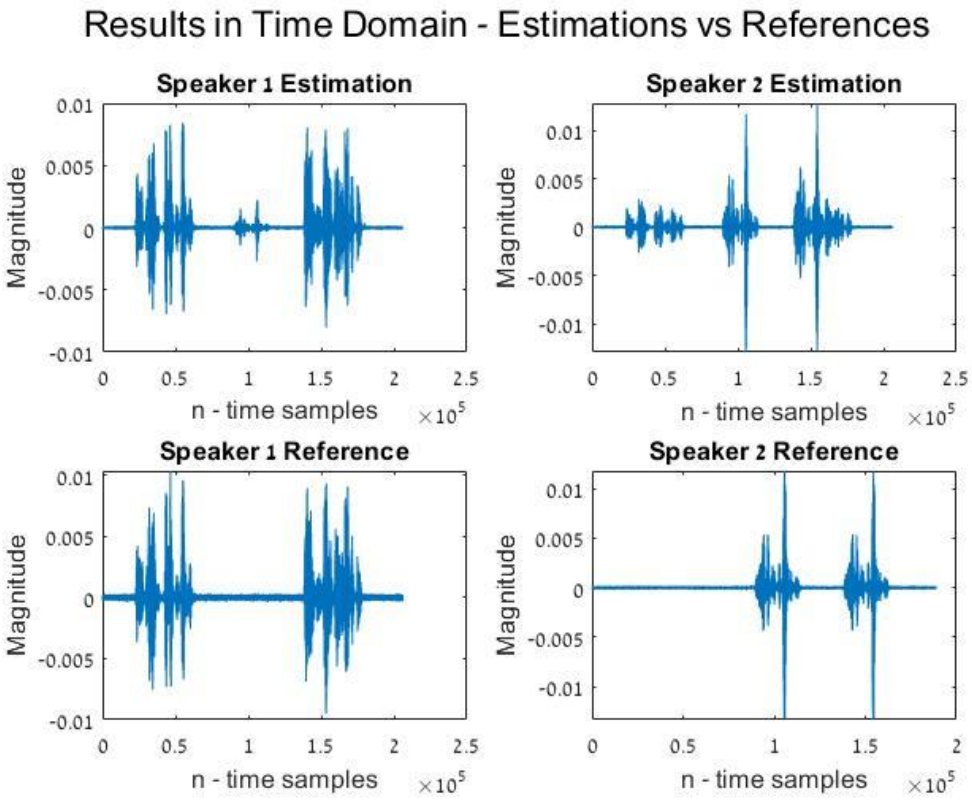
נראה את תוצאות ה-*LCMV* במישור *STFT* עבור כל דובר:

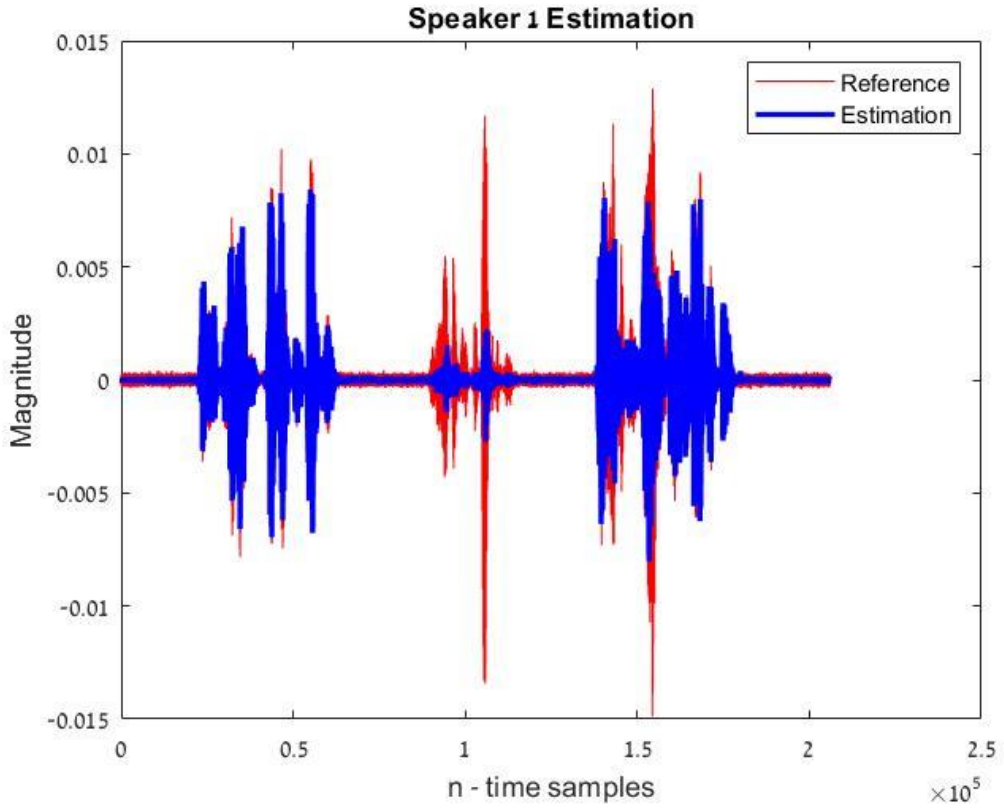


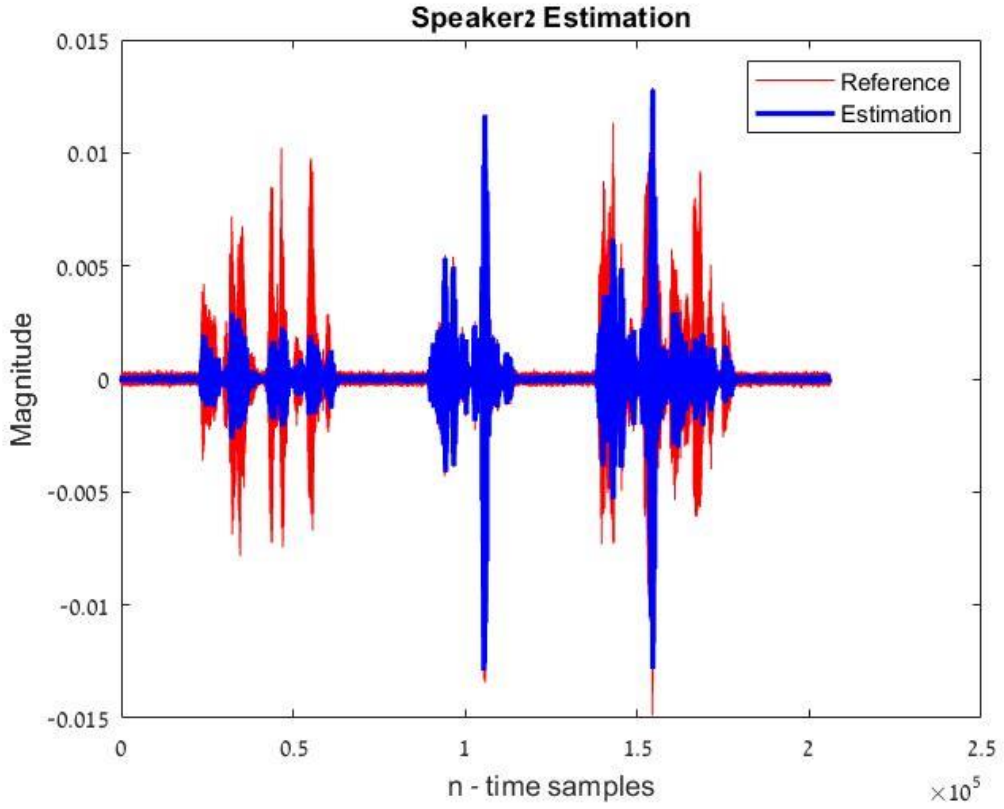


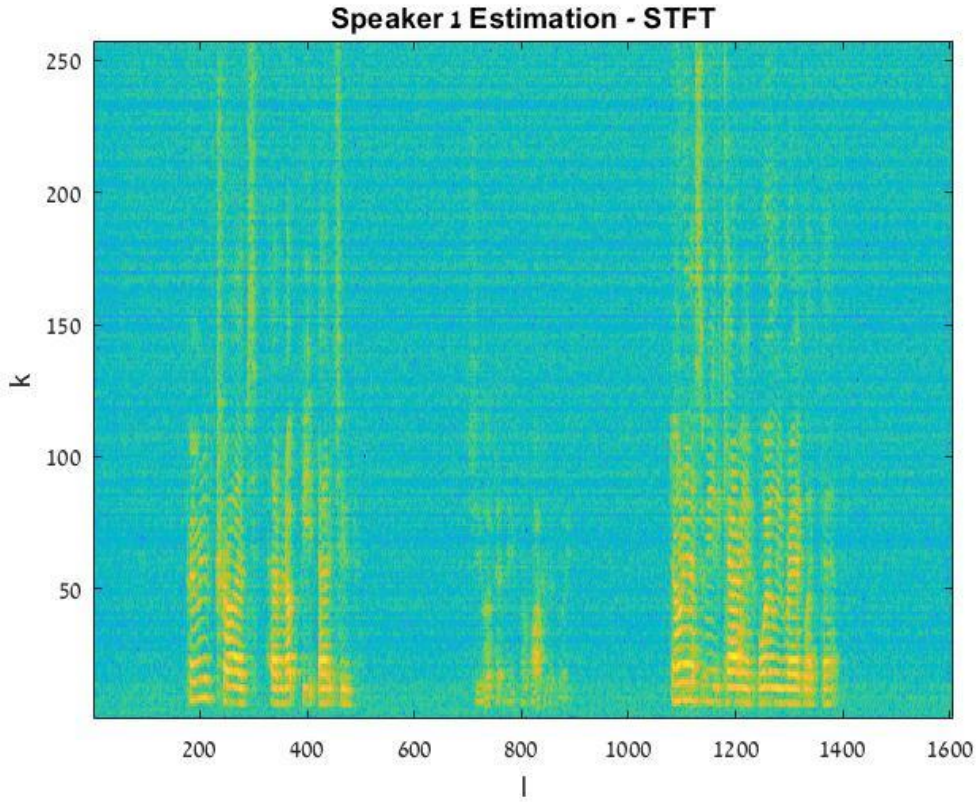
במישור התדר קצת קשה יותר לראות את הנחתת הדוברים הלא-רצויים מאשר במישור הזמן, אך עדיין אפשרי לראות את ההנחתה. ניתן גם אכן לראות ששערוך הדובר הראשון התבצע טוב יותר מאשר שערוך הדובר השני. הדובר השני מונחת יותר בגרף הראשון מאשר ההנחתה של הדובר הראשון בגרף השני ובנוסף גם הרעש הונחת יותר בשערוך הדובר הראשון.

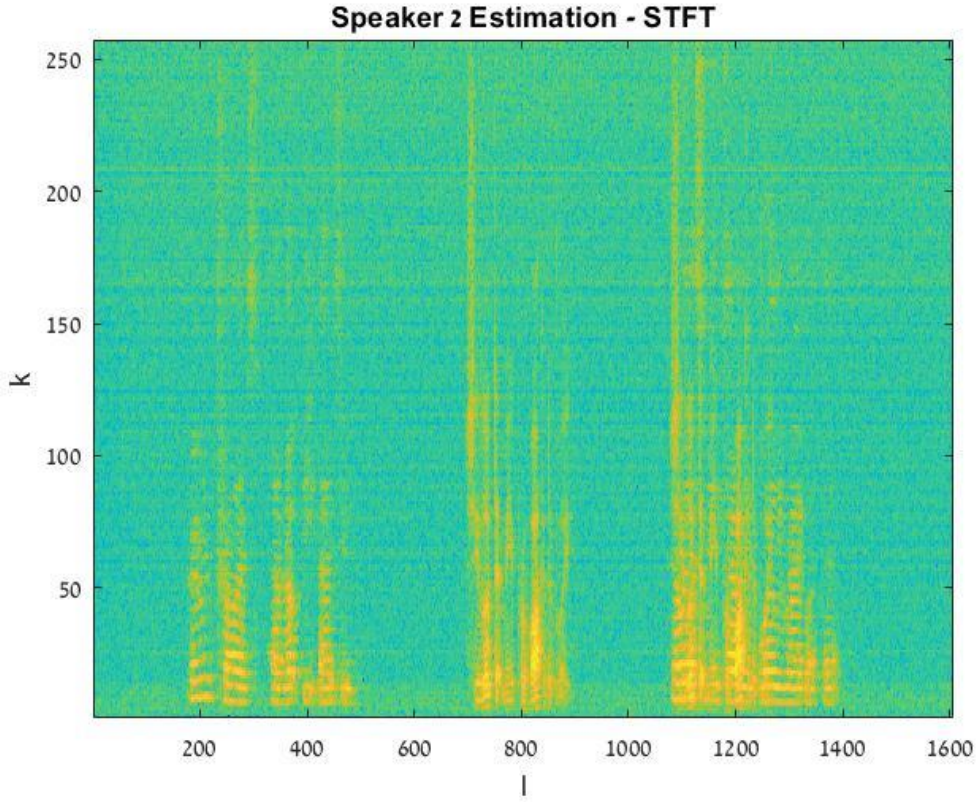
נציג כעת את תוצאות ה-*LCMV* כאשר שערוך ה-*RTF* נעשה בשיטת הפחתה ספקטרלית:




Speaker 1 Estimation
Reference
Estimation
Magnitude
n - time samples
×10^5


Speaker2 Estimation
Reference
Estimation
Magnitude
n - time samples
×10^5





ניתן לראות שהביצועים אמנם דומים מאוד, אך ה-*GEVD* עבד יותר טוב. בפרט ניתן לראות זאת מהרעש שפחות הונחת מה-*GEVD*.

נראה בטבלה להלן את תוצאות ההנחתות של הדוברים ושל הרעשים:

שיטה / אות מונחת	רעש בדובר הראשון	רעש בדובר השני	הנחתת הדובר הראשון	הנחתת הדובר השני
GEVD (CW)	9.4873	9.0037	9.6662	13.3703
הפחתה ספקטרלית *(CS)*	7.177	6.3747	9.217	13.4963

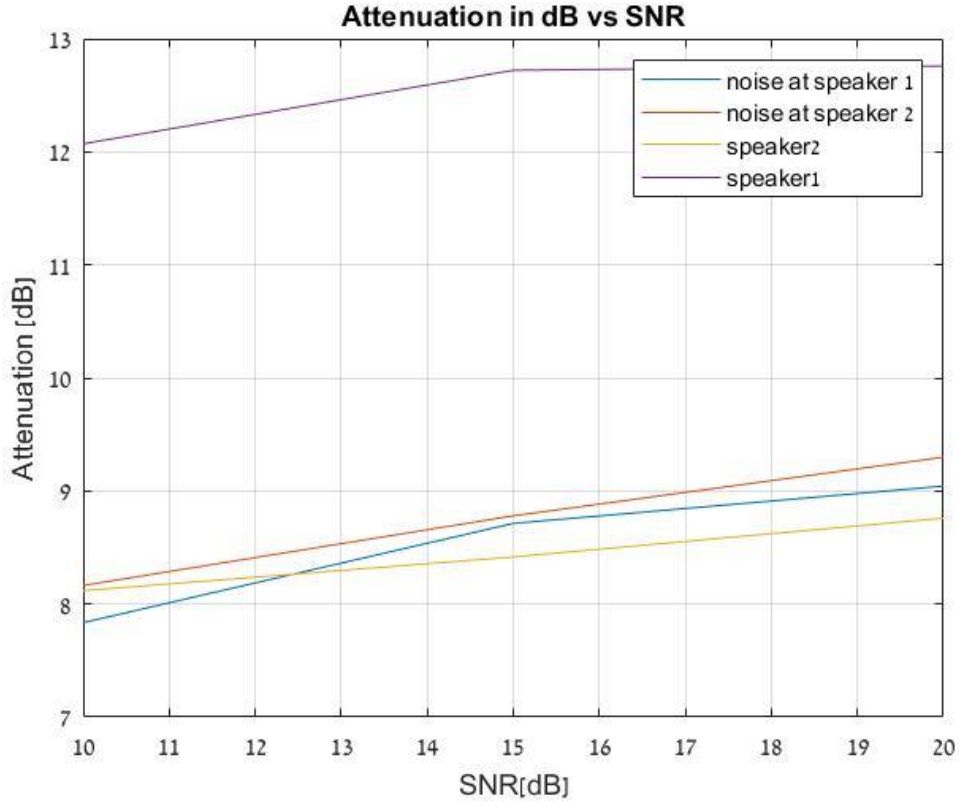
כל המספרים המופיעים בטבלה מסמנים ניחות ב-*dB*.

ניתן לראות גם מכאן כי שיטת ה-*GEVD* טובה יותר משיטת ההפחתה הספקטרלית. תוצאה זו מסתדרת עם תוצאה מהספרות ([5]) שמראה כי במקרה של רעש לבן שיטת *CW* (שקולה ל-*GEVD*) טובה יותר משיטת *CS* (שקולה להפחתה ספקטרלית).

כעת נבחן כיצד שינויים ב-*SNR* ובמקדם ההדהוד T_{60} ישפיעו על ביצועי ה-*LCMV*. נבחן את ההשפעה כאשר ה-*RTF* ישוערך בשיטת *GEVD*.

ראשית נשנה את ה-*SNR* כאשר $\beta = 0.3ms$. נבחן את ביצועי ה-*LCMV* עבור $SNR = 10,15,20dB$.

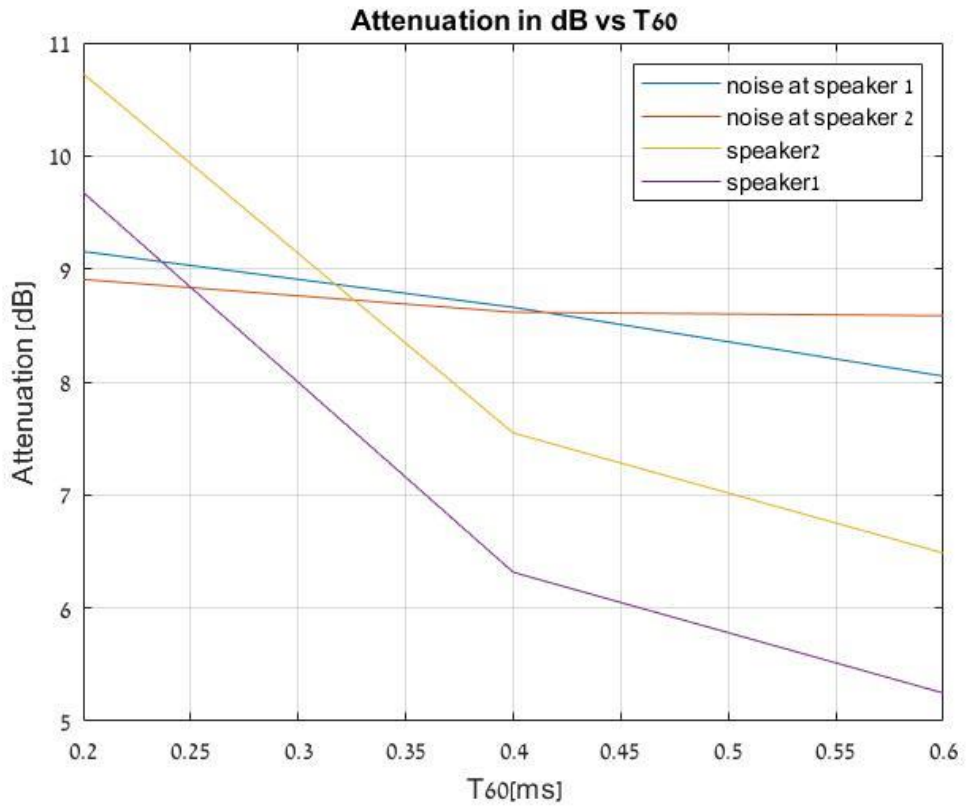
עבור כל ערך של *SNR* הגרלנו מספר קטעי קול וראינו איך ההנחתה מתנהגת כפונקציה של ה-*SNR*. נזכיר שאלגוריתם ה-*LCMV* לא עובד באופן אופטימלי שהיינו רוצים ולפעמים קיבלנו תוצאות לא טובות. סך הכל, עבור קטעי קול שבהם האלגוריתם עבד בצורה שמניחה את הדעת (אלו הם רוב המקרים), קיבלנו את הגרף הבא עבור ההנחתות:



ניתן לראות שככל שה-*SNR* יותר גבוה ההנחתה הינה יותר גדולה ויש פחות משקל לרעש.

כעת נשנה את $T_{60} = \beta$ כאשר $SNR = 20dB$. נבחן את ביצועי ה-*LCMV* עבור $T_{60} = 0.2,0.4,0.6ms$.

גם כאן, עבור קטעי קול שבהם האלגוריתם עבד בצורה שמניחה את הדעת (אלו הם רוב המקרים), קיבלנו את הגרף הבא עבור ההנחתות:



ניתן לראות שככל שמקדם ההדהוד (זמן ההדהוד בחדר) הינו גדול יותר, כך ההנחתה קטנה יותר כלומר ביצועי ה-*LCMV* גרועים יותר.

כאשר זמן ההדהוד גדול יותר, התגובה להלם של מערכת נמשכת לאורך יותר זמן, וגורמת לאות הדיבור "להימרח" יותר בהגעה אל המיקרופונים. ככל שיש פחות הדהוד, כך אות הדיבור מגיע חד יותר ומדויק יותר למיקרופונים. אבל אם הוא עובר קונבולוציה עם הרבה דלתאות, כאשר ההדהוד ארוך יותר, האות נמרח ופחות מדויק. לכן קל יותר לשערך אות שמגיע אלינו יותר "חד ומדויק".

נשים לב שכאן הנחתת הרעש הושפעה פחות מהנחתת הדוברים, זאת מכיוון שהרעש אינו תלוי בחדר. הפרמטר היחיד של הרעש הינו ה-*SNR* (כאן קבענו אותו להיות גבוה - $20dB$). אמנם אנו רואים שיש השפעה על מקדם ההדהוד בחדר על הנחתת הרעש אך היא קטנה יותר מההשפעה על איכות הנחתת הדוברים.

רעיונות להמשך המחקר

ל-*LCMV*, על כל יתרונותיו, יש גם חסרונות. אחד החסרונות העיקריים שלו הוא שהוא מניח הנחות שלא ניתן להניח בזמן אמת. הוא מניח שאנחנו יודעים מתי אין דיבור, מתי דובר אחד בלבד מדבר (ומי מהם) ומתי שני הדוברים מדברים ביחד. בזמן אמת איננו יודעים נתונים אלו כלל.

לכן יש צורך לתת מענה לבעיה זו. עלינו לבנות מערכת שתדע לסווג את הפריימים לדוברים השונים (או לרעש).

בשנים האחרונות מתבצעים מחקרים רבים בנושא שימוש בלמידת מכונה ו/או בלמידה עמוקה לפתירת הבעיה.

ב-[1] מוצע לאמן גלאי דיבור רב ערוצי מקבילי (*MCCSD*) שמבוסס על רשת נוירונים על מנת לבצע קלסיפיקציה לפריימים השונים. פריימים שישויכו לרעש בלבד ישמשו לשערוך מטריצת הקורלציה של הרעש, ופריימים שישויכו לדיבור יחיד (לא שני הדוברים בו-זמנית) ישמשו לשערוך ה-*RTF*.

ב-[4] מוצע לאמן גלאי קול (*VAD – Voice Activity Detector*) למציאת פריימים המכילים רעש בלבד על מנת לעדכן את מטריצת הקורלציה של הרעש בכל תדר, ובנוסף גם לאמן מזהה מקום דיבור (*SPI – Speakre Position Identifier*) על מנת לשערך את פונקציות ה-*RTF*.

בכיוון קצת אחר: אלגוריתם ידוע לקלאסטרינג הינו אלגוריתם *GMM*. לפי [6], יעיל להשתמש במודל *cGMM (complex Gaussian Mixture Model)* כאשר האותות המדוברים הינם בעלי הספק משמעותי רק בנקודות זמן-תדר מעטות. הנחה זו יוצאת מתוקפה כאשר אנו באים לדבר על אותות דיבור שנוסף אליהם רעש רקע, שמוריד מאיכותו של המודל. כותבי המאמר מציעים מודל *noisy cGMM* שניתן ליישם על אותות שלא בהכרח מקיימים את ההנחה לעיל. בנוסף, מציעים כותבי המאמר אלגוריתם מהיר יותר מאלגוריתם ה-*EM* הרגיל לשערוך פרמטרי המודל. המאמר מצא שהמודל ה"רועש" אפקטיבי להפרדת מקורות דיבור ולסינון רעש יותר מהמודל הרגיל. כיוון מעניין לחקור הוא להשתמש במודל זה כדי ליצור גלאי *MCCSD* שיסווג את הפריימים השונים לקבוצות, ונשתמש בקבוצות אלו על מנת לשערך את מטריצת הקורלציה של הרעש ואת פונקציות ה-*RTF*, אותן נכניס ל-*LCMV*.

ביבליוגרפיה

[1] S. E. Chazan, J. Goldberger and S. Gannot, "LCMV Beamformer with DNN-Based Multichannel Concurrent Speakers Detector," 2018 26th European Signal Processing .Conference (EUSIPCO), 2018, pp. 1562-1566, doi: 10.23919/EUSIPCO.2018.8553564

[2] S. Gannot, E. Vincent, S. Markovich-Golan and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 4, .pp. 692-730, April 2017, doi: 10.1109/TASLP.2016.2647702

[3] Habets, Emanuel AP. "Room impulse response generator." Technische Universiteit Eindhoven, Tech. Rep 2.2.4 (2006): 1.

[4] A. Malek et al., "Speaker extraction using LCMV beamformer with DNN-based SPP and RTF identification scheme," 2017 25th European Signal Processing Conference (EUSIPCO), 2017, pp. 2274-2278, doi: 10.23919/EUSIPCO.2017.8081615.

[5] S. Markovich-Golan, S. Gannot and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," 2018 26th European Signal Processing Conference (EUSIPCO), 2018, pp. 2499-2503, doi: 10.23919/EUSIPCO.2018.8553007.

[6] N. Ito, C. Schymura, S. Araki and T. Nakatani, "Noisy cGMM: Complex Gaussian Mixture Model with Non-Sparse Noise Model for Joint Source Separation and Denoising," 2018 26th European Signal Processing Conference (EUSIPCO), 2018, pp. 1662-1666, doi: 10.23919/EUSIPCO.2018.8553410.