ANDREW AND ERNA VITERBI FACULTY OF ELECTRICAL ENGINEERING

SIPL
Signal and Image Processing Lab.

Technion Israel Institute of Technology

דו"ח סיכום פרויקט: מיוחד

יצירת גדר אקוסטית באמצעות הפרדת דוברים עם מערך מיקרופונים

Acoustic Fencing using Multi-Microphone Speaker Separation

מבצעים:

אוראל בן ראובן **Orel Ben Reuven**

תומר פייט **Tomer Fait**

מנחה:

אמיר עברי **Amir Ivry**

סמסטר רישום: חורף תשפ"א

תאריך הגשה: ספטמבר, 2021

בשיתוף עם: Phoenix Audio Technologies

P 6090-1-21

תוכן עניינים

1 מבוא..... 1

2 הצגת הבעיה..... 3

3 סקירת המערכת המוצעת..... 4

3.1 מבנה האלגוריתם..... 4

3.2 ארכיטקטורת U-Net..... 6

4 מדדי ביצועים..... 8

4.1 רקע..... 8

4.2 הפרדת האותות למדדי הביצועים..... 8

4.3 מדדים כמותיים..... 9

4.3.1 מדד Frequency Weighted Segmental SNR..... 9

4.3.2 מדד Output SIR..... 10

4.3.3 מדד SIR Gain..... 10

5 ניסויים..... 11

5.1 מאגר הנתונים TIMIT..... 11

5.2 סימולציה..... 11

5.3 הקלטות אמיתיות..... 14

5.4 תוצאות..... 18

5.4.1 תוצאות הסימולציה..... 18

5.4.2 תוצאות ההקלטות האמיתיות..... 20

6 סיכום ומסקנות..... 23

6.1 דיון בתוצאות – ביצועי הסימולציה..... 23

6.2 דיון בתוצאות – ביצועי ההפרדה בהקלטות אמיתיות..... 23

6.3 כיוונים להמשך..... 24

רשימת מקורות..... 25

רשימת איורים

איור 1: דוגמא לגדר אקוסטית..... 3
איור 2: סכמת מלבנים של אלגוריתם DDESS..... 4
איור 3: ארכיטקטורת U-Net..... 6
איור 4: סימוני האותות בכניסה והיציאה מהמערכת..... 9
איור 5: היסטוגרמת פרמטרי סימולציה..... 13
איור 6: דוגמא לסיטואציה מהסימולציה..... 13
איור 7: תמונה של מערך המיקרופונים STEM TABLE..... 14
איור 8: תמונה של סימולטור הפה..... 15
איור 9: תמונה המתארת את setup ההקלטות ליצירת סט האימון..... 15
איור 10: סכמת החדרים בהם התבצעו ההקלטות..... 17
איור 11: תמונות החדרים בהם נרכש סט הבוחן..... 18
איור 12: ספקטוגרמות של אות הכניסה והאותות המופרדים בסימולציה..... 19
איור 13: ספקטוגרמות של אות הכניסה והאותות המופרדים בהקלטות אמיתיות..... 21
איור 14: מדדי הביצועים בחדרים אמיתיים..... 22

רשימת טבלאות

טבלה 1: תוצאות מדדי הביצועים בסימולציה..... 20
טבלה 2: תוצאות מדדי הביצועים בהקלטות אמיתיות..... 21

תקציר

מטרתו של אלגוריתם ליצירת גדר אקוסטית הינו לבצע הפרדה של אותות דוברים הנמצאים במיקומים שונים בחדר, כך שבכל איזור מוגדר המערכת תעביר ללא עיוות רק את האותות המגיעים מאיזור זה, ותבטל הפרעות ממקומות אחרים בחדר. בפרויקט נבחן אלגוריתם קיים לפתרון הבעיה, נפתח מדדי ביצועים חדשים, ונבחן את המערכת על סימולציה ועל הקלטות אמיתיות בסביבות אקוסטיות שונות, אותם הקלטנו באופן עצמאי עם ציוד אקוסטי מותאם. השיטה שנבחן מבצעת מיסוך במישור הזמן-תדר, על ידי הנחת דומיננטיות של דובר יחיד בכל תא זמן-תדר וסיווגו של כל תא בעזרת רשת נוירונים עמוקה מבוססת קונבולוציה. מדדי ביצועים סטנדרטיים לא מכמתים באופן בלתי תלוי את רמת עיוות האות הרצוי ואת רמת הנחתת אות ההפרעה, ולרוב מכמתים את שתי התופעות בעזרת מדד יחיד. בשל כך, נציג שיטה לאמוד תופעות אלו בנפרד ע"י הפעלת המסכה שהאלגוריתם מייצר על האות הרצוי ועל אות ההפרעה בנפרד ובאופן בלתי תלוי. מסכה זו משתנה בזמן ומייצגת את ההגבר שהרשת מפעילה על הכניסה, כך שהכפלת הדובר הרצוי בלבד במסכה, תאפשר הערכה של פעולת המערכת על אות רצוי בלבד. נמשיך ונבחן את השיטה ואת מדדי הביצועים שהוגדרו על אותות דוברים המגיעים מסימולציה המדמה חדרים בעלי ממדים, מיקומי דוברים, וזמני הדהוד מגוונים. בעקבות הצלחת ההפרדה על אותות הסימולציה, המשכנו בבחינת המערכת על הקלטות אמיתיות שבוצעו במעבדה בעזרת מערך מיקרופונים וסימולטור פה. על מנת לבחון את הכללת המערכת, סט הבוחן הוקלט בחדרים שונים בעלי מאפיינים אקוסטיים מגוונים שלא נכללו בסט האימון, וגם כאן הראנו הפרדה מוצלחת של הקלטות סט הבוחן. לסיכום, מחקר זה מציג מערכת גדר אקוסטית שביצועיה מוערכים על פי מדדים בעלי קורלציה גבוהה לתפיסה אנושית של איכות שמע, ותחתיהם המערכת מראה ביצועים מוצלחים בסביבות אקוסטיות אמיתיות. כמו כן, המשאבים הנמוכים וזמני החישוב הקצרים של המערכת עשויים להצביע על התאמת המערכת להטמעה כיחידה חישובית בפלטפורמות תקשורת פרקטיות הכוללות מערך מיקרופונים.

Abstract

The goal of an acoustic fencing algorithm is to separate speakers by their physical location in space. In this project, we examine an algorithm which solves this problem, define suitable performance criteria, and test the algorithm in varied environments, both simulated and real. The real recordings were acquired by us with suitable acoustic equipment. We examine a speech separation algorithm based on spectral masking inferred from the speaker's direction. The algorithm assumes the existence of a dominant speaker in each time-frequency (TF) bin and classifies these bins by employing a deep convolutional neural network. Traditional evaluation criteria do not independently quantify the effects of the desired signal distortion and the undesired signal attenuation, and often result in a single numeric value for both effects. In this project, we propose a method for evaluating these phenomena separately by applying the separation mask to the original separated signals. This mask is time-dependent and represents the

network's gain, such that by applying it to the desired signal (for example), we can evaluate the network's effect on the signal. We tested the algorithm and evaluation criteria on a simulated signals with varied room sizes, speaker's locations, and reverberation times. Following the success in the simulation, we continued to test the algorithm on real recordings acquired in the lab employing a microphone array and mouth simulator. To evaluate the generalization of the system, a test set was comprised from recordings acquired in rooms which were not present in the training set. In conclusion, this research describes an acoustic fencing algorithm and evaluation criteria with high correlation human perception, and shows successful performance in a real acoustic environment. Furthermore, the system's low resource consumption and fast response times might indicate that it is suitable as a practical algorithm in a real system.

1 מבוא

הפרדת אותות דיבור הינו תחום מחקר נפוץ שהתפתח רבות בעשורים האחרונים. מספר שיטות להפרדת דוברים מסתמכות על כך שהדוברים נמצאים במיקומים מרחביים שונים ונעזרות במערך מיקרופונים לצורך ביצוע ההפרדה [1]. שיטות שונות מבצעות הפרדה בעזרת מיקרופון יחיד ולא תלויות במיקום המרחבי של הדוברים [2], [3]. בפרויקט זה נתמקד בשיטה המשתמשת במערך מיקרופונים ומפרידה דוברים לפי המיקום שלהם בחדר על מנת ליצור גדר אקוסטית. בבעיה זו, החדר מחולק למספר תחומים, ועל המערכת לבצע הפרדה בין אותות הדוברים שנמצאים בתחום אחד, לבין אותות הדוברים משאר התחומים. מערכת המפרידה דוברים לפי המיקום המרחבי שלהם עשויה להיות שימושית להדגשת דובר הנמצא במיקום קבוע מתוך סביבה רועשת, זיהוי מיקום דוברים המדברים בו-זמנית, והשמעת הדובר הרצוי.

אתגר שעשויים להיתקל בו כאשר מנסים ליצור גדר אקוסטית מתרחש כאשר דוברים מדברים בעוצמות שונות והמטרה הינה לחלץ את אות הדובר החלש. בנוסף, סביבות אקוסטיות מאתגרות הכוללות מגוון חדרים, סוגי הדהודים, ורעשים סטציונריים ולא סטציונריים, עשויים לפגוע ביכולת ההפרדה של המערכת. הדרישה מהמערכת היא להעביר את אות הדובר הרצוי ללא עיוותים, ולבטל את אותות הדוברים הלא רצויים.

בפרויקט זה מימשנו את אלגוריתם Deep Direction Estimation for Speech Separation (DDESS) [4] המבצע הפרדה של הדוברים בעזרת יצירת מסיכות בתחום הזמן-תדר. אלגוריתם זה משתמש ברשת נוירונים עמוקה המבוססת על קונבולוציה על מנת לסווג תאי זמן-תדר למיקומים מרחביים על ידי שימוש בהתמרת ה- Short-time Fourier transform (STFT) [5] של אותות המיקרופונים. בצורה זו מחושבת מסכה שבעזרתה ניתן לשחזר את אותות הדוברים המופרדים.

על מנת להעריך את ביצועי האלגוריתם, הצענו שיטה לחישוב מדדי ביצועים המסוגלת להבחין בין עיוות האות הרצוי לבין הנחתת האותות הלא רצויים. שיטה זו משתמשת במסכה הנלמדת על ידי הרשת בתור הגבר משתנה בזמן התלוי בכניסה. באופן זה, ניתן להפעיל את הגבר רשת הקונבולוציה המאומנת על האות הרצוי בלבד, ולבחון את השפעת המערכת על עיוות האות הרצוי. באותו אופן, ניתן להפעיל את הגבר הרשת על האותות הלא רצויים ולאמוד את רמת ההנחתה שלהם, ללא צימוד למדד העיוות הקודם. כלומר, ניתן לכמת את עיוות האות המשוחזר במוצא ללא התחשבות בהפרעת האות הלא רצוי, וניתן לכמת כמה מתוך האות הלא רצוי "זלג" לתוך שחזור האות הרצוי.

בחנו את המערכת על אותות מסימולציה ועל הקלטות אמיתיות. הסימולציה מורכבת מחדרים שונים המגדירים מצבים מגוונים, וההקלטות האמיתיות מכילות נרכשו בעזרת מערך מיקרופונים וסימולטור פה. המערכת שלנו הציגה ביצועים טובים מבחינת שימור האות הרצוי, הורדת הפרעות, יכולת הכללה לסביבות אקוסטיות שונות, ורובסטיות לרמות גבוהות של רעש והדהודים.

כמו כן, המערכת שפיתחנו משתמשת בכמות משאבים וצורכת זמן חישוב המתאימים לשילובה בפלטפורמת תקשורת חומרתית.

2 הצגת הבעיה

הבעיה של יצירת גדר אקוסטית מוגדרת בצורה הבאה: נתון חדר בו ממוקם מערך של M מיקרופונים הקולט אותות דיבור של N דוברים. נסמן את אות הדיבור של הדובר ה-i ב- $s_i(n)$, אות זה מהדהד בחדר ונקלט ע"י המיקרופון ה-m בצורה הבאה:

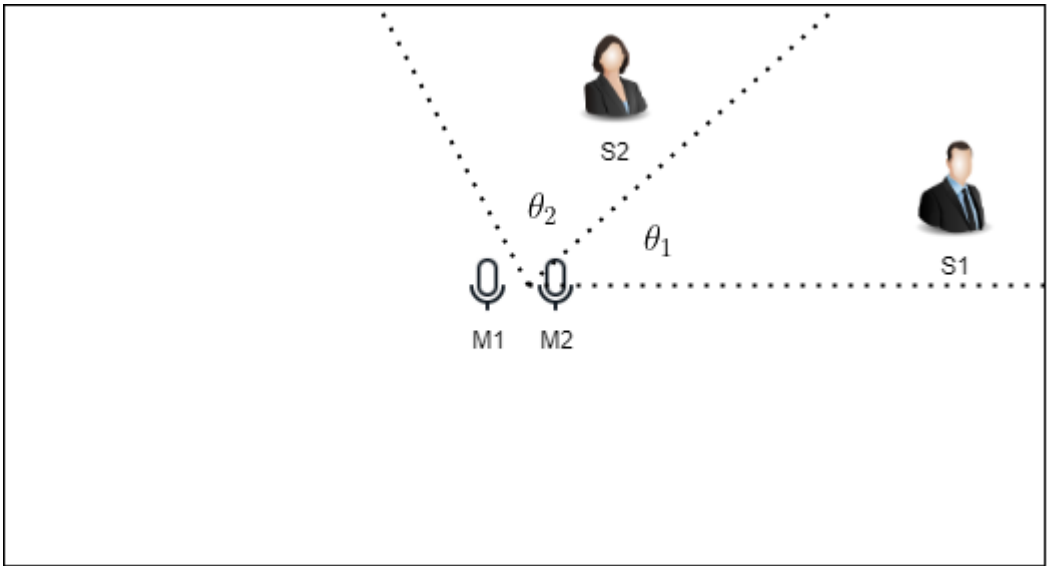
$$z_m(n) = \sum_{i=1}^{N} \left[s_i(n) * h_m^i(n) \right] + W_m(n)\,, \qquad m \in \{1,..,M\} \tag{1}$$

כאשר $h_m^i(n)$ הינה תגובת החדר המתאימה לדובר ה-i ולמיקרופון ה-m, ו- $W_m(n)$ הינו הרעש המתווסף למיקרופון זה. בסימולציה נמדל רעש זה כ Additive White Gaussian Noise (AWGN).

החדר מחולק ל-K תחומי קבלה: $\{\theta_1,..,\theta_K\}$. המטרה היא לקבל את האותות $\{z_m(n)\}_{m=1}^{M}$ ובעזרתם ליצור את K אותות המוצא $\{y_k(n)\}_{k=1}^{K}$, כך שהמוצא $y_k(n)$ מכיל רק את אותות הדוברים שהיו בתחום הקבלה θ_k.

באיור 1: דוגמא לגדר אקוסטית עבור K=M=N=2 ניתן לראות דוגמא לחדר המחולק לשני תחומי קבלה $\{\theta_1,\theta_2\}$, בכל תחום קבלה נמצא דובר יחיד המסומן ע"י s_1 ו- s_2 בהתאמה, ואותות הדוברים נקלטים במערך של שני מיקרופונים המסומנים ע"י $\{M_1,M_2\}$.

בפתרון, כפי שיוצג בהמשך, אין הנחה על מיקום המיקרופונים, מפני שההפרדה נלמדת מהפרשי פאזה בין המיקרופונים במערך. בפרויקט זה נניח כי לא קיימים שני דוברים הנמצאים באותו מקום. בנוסף, נניח כי הדוברים לא משנים את המיקום שלהם באופן משמעותי תוך כדי משפט בודד. נציין כי בבעיה כפי שהוגדרה, המטרה אינה לשערך את המיקום המדויק של הדוברים, אלא רק לקבץ אותם לתחומי קבלה המוגדרים מראש.



איור 1: דוגמא לגדר אקוסטית עבור *K=M=N=2*

3 סקירת המערכת המוצעת

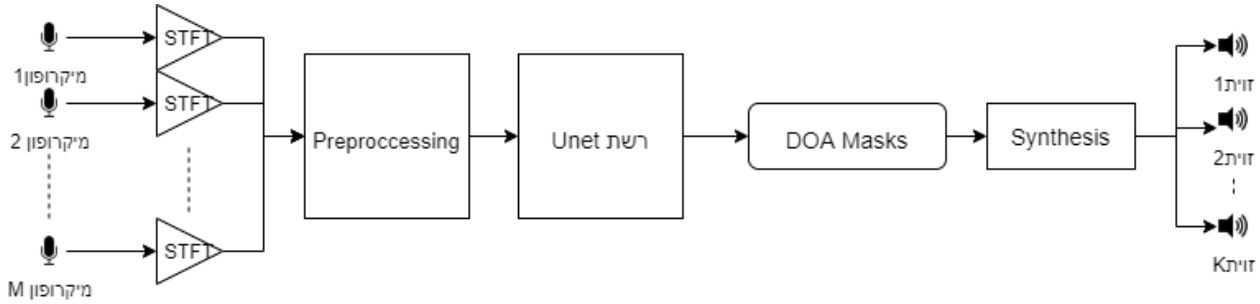
3.1 מבנה האלגוריתם

אלגוריתם ה-DDESS המוצע במאמר [4] יוצר גדר אקוסטית בעזרת מעבר למישור ה-STFT. בשיטה זו יוצרים מסיכה שמוכפלת בהתמרת ה-STFT (ספקטוגרמה) של האות שהוקלט באחד המיקרופונים, הוא מיקרופון הייחוס (reference). בצורה זו מונחתים תאי הזמן-תדר שלא שייכים לתחום הקבלה הרצוי. מיסוך זה מסתמך על עיקרון W Disjoint Orthogonality [7]. לפי עיקרון זה, עבור מספר דוברים מספיק קטן, כל תא זמן-תדר בהתמרת ה-STFT נשלט ע"י דובר יחיד. בהגדרת הבעיה, הנחנו כי מיקומי הדוברים שונים, ולכן נובע כי כל תא זמן-תדר נשלט ע"י מיקום יחיד וכתוצאה מכך גם ע"י אזור קבלה יחיד.

האלגוריתם מבוסס על ארבעה שלבים עיקריים:

1. ביצוע התמרת STFT לאותות הנקלטים במערך המיקרופונים
2. ביצוע עיבוד מקדים על ההתמרה
3. סיווג תאי הזמן-תדר לתחומי קבלה שונים באמצעות רשת קונבולוציה מסוג U-net [8], ויצירת מסיכות
4. מיסוך מיקרופון הרפרנס וסינתזה

באיור 2: סכמת מלבנים של אלגוריתם DDESS ניתן לראות את סכמת הבלוקים של אלגוריתם ה-DDESS המבטאת את השלבים העיקריים שצוינו עבור המקרה הכללי של M מיקרופון ו-K תחומים קבלה.



איור 2: סכמת מלבנים של אלגוריתם DDESS

בשלב הראשון מבצעים התמרת STFT לאותות הנקלטים במיקרופונים. נסמן את התמרת ה-STFT בגודל $L \times W$ של $z_m(n)$, ב- Z_m, ואת צפיפות הרעש הספקטרלית בW_m. כך שעבור כל תא (l,w) מתקיים:

$$Z_m(l,w) = \sum_{i=1}^{N} S_i(l,w) \cdot H_m^i(l,w) + W_m(l,w), \qquad m \in \{1,..,M\} \tag{2}$$

בשלב העיבוד המקדים נחשב $(M-1)$ תמונות חדשות המכילות עבור כל תא (l,w) את הפרש הפאזות בין המספרים המרוכבים: $Z_m(l,w)$ ו $Z_1(l,w)$. כאשר $Z_1 \equiv Z_{\mathrm{ref}}$ הינה התמרת ה-STFT של מיקרופון הרפרנס אותו נמסך . כל אחת מתמונות אלו תהפוך לשתי תמונות חדשות ע"י הפעלת פונקציות טריגונומטריות על כל אחד מהתאים בתמונות. נחבר את התמונות למטריצה תלת ממדית R בגודל $L \times W \times 2 \cdot (M-1)$. נקבל כי כל זוג תאים $(l,w,m_1),(l,w,m_2)$ מוגדר ע"י:

$$r(l,w,m_1) = cos\left(\angle \frac{Z_m(l,w)}{Z_{ref}(l,w)}\right)$$
$$r(l,w,m_2) = sin\left(\angle \frac{Z_m(l,w)}{Z_{ref}(l,w)}\right) \tag{3}$$

נגדיר בעיית סיווג לסט אזורי הקבלה $\Theta = \{\theta_1,..,\theta_K\}$. מטרתנו בבעיית הסיווג הינה לקבל K מסיכות מוצא $\{\widehat{M}_k\}_{k=1}^{K}$ בגודל $L \times W$, כך שכל תא (l,w) במסכת המוצא $\widehat{M}_k$ הינו ההסתברות שתא הזמן-תדר (l,w) שייך לתחום הקבלה k בהינתן הכניסה R. כלומר אם $y(l,w)$ הינו אינדיקטור לתחום הקבלה של התא (l,w) אז:

$$\widehat{M}_k(l,w) = p(y(l,w) = \theta_k | R) \tag{4}$$

את הסיווג נבצע בעזרת רשת קונבולוציה מסוג U-Net עליה נפרט בחלק הבא.

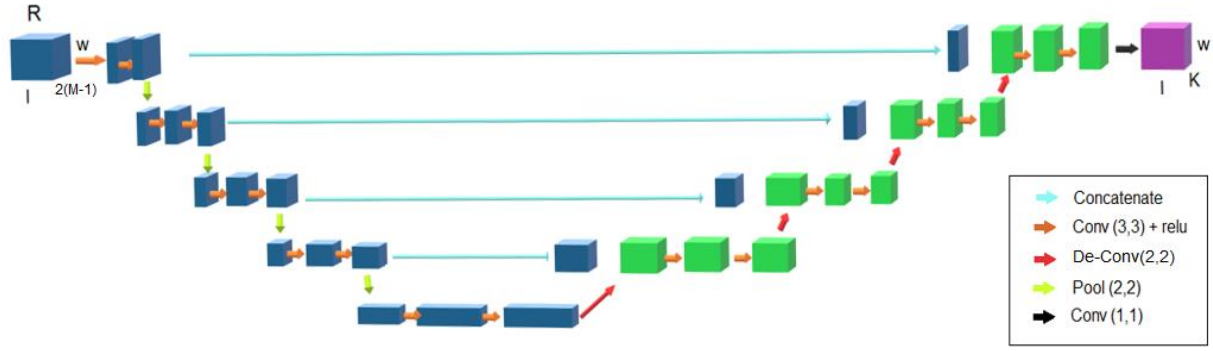
בשלב הסינתזה, ניקח את K המסכות שהתקבלו במוצא הרשת, ונכפיל כל אחת מהמסכות (המתאימות לאזורי הקבלה) התמרת ה-STFT של מיקרופון הרפרנס Z_{ref}. לאחר מכן, נבצע התמרת STFT הפוכה, ונקבל שחזור של האות בזמן לכל אחד מתחומי הקבלה:

$$y_k(n) = iSTFT\left(Z_{\mathrm{ref}} \cdot \widehat{M}_k\right) \tag{5}$$

3.2 ארכיטקטורת U-Net

הכניסה לרשת הינה מטריצת המאפיינים R, שכל איבר בה מוגדר לפי $r(l,w,m_{1,2})$ כפי שנתון במשוואה (3). בארכיטקטורה שלנו גודל הכניסה הינו $L \times W \times 2 \cdot (M-1)$. כאשר $L = 100$ הינו מספר הפריימים בזמן, $W =$ 257 הינו מספר תאי התדר, ו-$M = 9$ הינו מספר המיקרופונים. גודל מוצא הרשת הינו $L \times W \times K$, כאשר K הינו מספר אזורי הקבלה.

ארכיטקטורת הרשת מוצגת באיור 3: ארכיטקטורת U-Net. המלבנים הכחולים שייכים למקודד, והירוקים שייכים למפענח. בשלב המקודד, תמונת הכניסה עוברת הורדת ממדים בעזרת שכבות max pooling, אשר דוגמות את הערך המקסימלי מהממד הגבוה, הנמצא בטווח של גודל הצעד הרצוי. בשלב הפענוח, התמונה חוזרת לגודלה המקורי בעזרת שכבות ConveTranspose2d. בשיטת אינטרפולציה זו, הערכים החסרים בתמונה הגדולה, מחושבים ע"י די-קובולוציה דו ממדית מהערכים של הפיקסלים השכנים. על מנת לא לאבד את המידע המקומי שנמחק בשכבת ה- max pooling, הרשת מחברת בין השכבות במקודד והמפענח וכך מאפשרת למידע לעקוף את צוואר הבקבוק.



איור 3: ארכיטקטורת *U-Net*

נסמן את שכבות הרשת בצורה הבאה:

- $CE_{l,s}$ הינה שכבת קונבולוציה דו-ממדית כאשר l הינו מספר הפילטרים ו- $s \times s$ זהו גודל כל פילטר.
- DE_d הינה שכבת `ConvTranspose2d` כאשר d הינו גודל הגרעין.
- P_s הינה שכבת Max-Pooling.
- RL הינה שכבת אקטיבציה מסוג ReLU.
- BN הינה שכבת Batch Normalization .

המקודד בנוי בצורה הבאה:

$$CE_{16,3} \to BN \to RL \to CE_{16,3} \to BN \to RL \to P_2 \to CE_{32,3} \to BN \to RL \to CE_{32,3} \to BN \to RL \to P_2 \to$$
$$CE_{64,3} \to BN \to RL \to CE_{64,3} \to BN \to RL \to P_2 \to CE_{128,3} \to BN \to RL \to CE_{128,3} \to BN \to RL \to$$
$$P_2 \to CE_{128,3} \to BN \to RL \to CE_{128,3} \to BN \to RL$$

המפענח בנוי בצורה הבאה:

$DE_2 \rightarrow CE_{128,3} \rightarrow BN \rightarrow RL \rightarrow CE_{64,3} \rightarrow BN \rightarrow RL \rightarrow DE_2 \rightarrow CE_{64,3} \rightarrow BN \rightarrow RL \rightarrow CE_{32,3} \rightarrow BN \rightarrow RL \rightarrow DE_2 \rightarrow CE_{32,3} \rightarrow BN \rightarrow RL \rightarrow CE_{16,3} \rightarrow BN \rightarrow RL \rightarrow DE_2 \rightarrow CE_{16,3} \rightarrow BN \rightarrow RL \rightarrow CE_{16,3} \rightarrow BN \rightarrow RL \rightarrow CE_{2,1}$

במוצא הפענח מוסיפים שכבת Softmax על מנת שערכי מוצא הרשת יהיו בתחום $[0,1]$ ויתאימו להסתברויות.

על מנת לאמן את הרשת השתמשנו באותות דוברים אקוסטיים המגיעים מסימולציה (יפורט בהמשך). בעזרת קולות הדוברים המופרדים המגיעים מהסימולציה ניתן למצוא את תחום הקבלה הדומיננטי השייך לכל תא זמן-תדר וכך לקבל את תמונת התיוגים בגודל $L \times W$ המכילה בכל תא (l, w) את האינדקס של התחום הדומיננטי השייך לתא זה.

פונקציית המטרה של הרשת הינה פונקציית ה- Cross Entropy [9] בין הסיווג האמיתי לסיווג המשוערך ע"י הרשת:

$$\text{loss}(x, \text{class}) = -\log\left(\frac{exp(x[\text{class}])}{\sum_j exp(x[j])}\right) \tag{6}$$

$$\text{loss} = \sum_{i=1}^{N} \text{loss}(i, \text{class}[i]) \tag{7}$$

כאשר x הינו מוצא הרשת עבור תחום קבלה יחיד, ו-$class$ הינו האינדקס מהתיוג המצביע על תחום הקבלה הרצוי. בהמשך, נבחן אפשרות של שינוי פונקציית המטרה והתיוג. הרשת ממומשת ע"י PyTorch ומאומנת בעזרת ADAM optimizer [10]. גודל ה-minibatch נבחר להיות 8, ופרמטרי הרשת נבחרים לפי נקודת המינימום של השגיאה על סט הוולידציה.

4 מדדי ביצועים

4.1 רקע

כאשר נבחן גדר אקוסטית, נתבונן בשני מטרות בסיסיות של האלגוריתם שאת הצלחתן נרצה לכמת. המטרה הראשונה, הינה העברת האותות השייכים לאזור הקבלה ללא עיוותים, והשנייה הינה הנחתה של האותות שאינם שייכים לאזור הקבלה.

עבור אלגוריתמים הפועלים בעזרת מיסוך של ייצוג כלשהו של האות, כאשר המסכה חוסמת חלקים מהייצוג של האות הרצוי, נוצר עיוות בצורה של "אכילות" של האות. לדוגמא, באלגוריתם DDESS, אם ערך המסכה בתא זמן-תדר ששייך לדובר הרצוי הוא נמוך, נקבל "אכילה" בספקטוגרמה של האות הרצוי. באלגוריתם מסוג זה, הנחתה של אותות לא רצויים מתבצעת ע"י חסימה של תאים בהתמרה שאינם שייכים לדובר הרצוי. כאשר חלק מהתאים של האותות הלא רצויים לא נחסמים, נקבל "זליגה" מהאותות הלא רצויים לשחזור של האות הרצוי. ייתכן כי במרחב ההתמרה תא כלשהו יכיל מידע השייך גם לאותות רצויים וגם לאותות לא רצויים, במצב זה לא ניתן להימנע מאכילה או זליגה.

נרצה למדוד את הביצועים של שני המטרות בנפרד. מדידה זו עשויה להיות מאתגרת, מפני שהשחזור במוצא האלגוריתם מכיל גם את האות הרצוי וגם את הזליגה מהאות הלא רצוי. שימוש ישיר באות המשוחזר על מנת להעריך את הביצועים ייתן לנו רק מדד כמותי יחיד ולא יבטא את הביצועים של כל אחת מהמטרות בנפרד. ניסיון נאיבי להפרדת המדדים של שתי המטרות, יהיה להכניס לאלגוריתם רק אותות רצויים או לא רצויים וכך לקבל במוצא רק עיוותים (המגיעים מהאותות הרצויים) או זליגות (המגיעות מהאותות הלא רצויים). שיטה זו אכן מודדת עיוות וזליגה בנפרד, אך כיוון שהמקרה בו הכניסה מורכבת רק מאותות רצויים או לא רצויים הוא פשוט להפרדה, הביצועים שנקבל יהיו טובים בהרבה מהביצועים בפעולה רגילה של האלגוריתם, כלומר כניסה מעורבת (סכום של אותות רצויים ולא רצויים).

4.2 הפרדת האותות למדדי הביצועים

בפרויקט זה, נציע שיטה למדידת עיוות האות הרצוי והנחתת האות הלא רצוי בנפרד: בשלב הראשון נכניס לאלגוריתם את הקלט המקורי שיסומן ב-$r(n)$ (המכיל שילוב של האותות הרצויים והאותות הלא רצויים), ונקבל במוצא את האות המשוחזר שיסומן ב- $r'(n)$. בשלב השני, נחלץ את המסכה שהאלגוריתם ייצר בשלב בראשון $\widehat{M}_k$ המעבירה את תחום הקבלה θ_k. בשלב השלישי, ניצור בעזרת המסכה $\widehat{M}_k$ שני אותות חדשים:

- ניקח את האות הרצוי (שנמצא בתחום θ_k) שיסומן ב- $p(n)$, נכפיל אותו במסכה ונשלים את השחזור. האות המשוחזר שנקבל $p'(n)$ הינו האות הרצוי לאחר המעבר במערכת – אות המעבר. בעזרת אות המעבר נוכל למדוד את העיוות.

- ניקח את האות הלא הרצוי (כל הדוברים בתחומים $\theta_{j\neq k}$) שיסומן ב- $b(n)$, נכפיל אותו במסכה ונשלים את השחזור. האות המשוחזר שנקבל $b'(n)$ הינו האות הלא רצוי לאחר המעבר במערכת – אות החסום. בעזרת האות החסום נוכל למדוד את ההנחתה.

באיור 4: סימוני אותות מסוכמים הסימונים שישמשו אותנו לצורך מדדי הביצועים.

$$\begin{array}{c} \widehat{M} \\ r(n) \rightarrow \ r'(n) \\ p(n) \rightarrow \ p'(n) \\ b(n) \rightarrow \ b'(n) \end{array}$$

איור 4: סימוני אותות בכניסה והיציאה מהמערכת. $b(n)$ הינו האות הלא רצוי, $p(n)$ הינו האות הרצוי ו- $r(n)$ הינו האות המעורב.

מליניאריות פעולת המיסוך מתקיים כי:

$$\begin{aligned} r(n) &= p(n) + b(n) \\ r'(n) &= p'(n) + b'(n) \end{aligned} \tag{8}$$

בעזרת האותות שחולצו נוכל להפעיל מדדי ביצועיים כמותיים על האותות ולמדוד את העיוות וההנחתה בנפרד.

4.3 מדדים כמותיים

4.3.1 מדד Frequency Weighted Segmental SNR

המדד הכמותי שישמש אותנו למדידת העיוות של האות הרצוי הינו מדד ה- Frequency Weighted Segmental Signal To Noise Ratio (fwSNRseg) [11] . מדד זה מחשב את יחס האנרגיות (ב-dB) בין האות הרצוי $p(n)$ להפרש בין האות הרצוי לאות המעבר $p(n) - p'(n)$ לאחר משקול תדרי על מנת לדמות שמיעה אנושית. את יחס האנרגיות ממצעים על פני M חלונות בגודל L עם חפיפה של 75%:

$$fwSNRseg = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)\log_{10}\frac{P(j,m)^2}{\left(P(j,m)-P'(j,m)\right)^2}}{\sum_{j=1}^{K} W(j,m)} \tag{10}$$

כאשר P, P' הינם התמרות STFT של האותות p, p'. $W(j,m)$ הינו המשקל המתאים לתא הזמן-תדר (j,m).

גודל החלון הינו 30[ms], כלומר- $L = 30 \cdot F_s$ (fs הינו תדר הדגימה), ומספר החלונות M מתקבל על ידי: $M = \left\lceil \frac{length(p(n))}{0.25 \cdot L} \right\rceil$. ככל שמדד ה- $fwSNRseg$ גדול יותר, כך אות המעבר קרוב יותר לאות הרצוי והעיוות קטן יותר.

4.3.2 מדד Output SIR

המדד הכמותי שישמש אותנו למדידת טיב ההפרדה הינו המדד oSIR (Output Signal to interference Ratio). מדד זה מחשב את יחס האנרגיות בין האות הרצוי במוצא $p'(n)$ לאות הלא רצוי במוצא $b'(n)$:

$$oSIR = 20 \cdot \log_{10} \frac{RMS(p')}{RMS(b')} \; [dB] \tag{11}$$

כאשר RMS הינו Root Mean Squered.

ככל שמדד טיב ההפרדה גדול יותר, כך האות החסום קטן יותר ביחס לאות הרצוי במוצא, כלומר ההפרדה טובה יותר.

4.3.3 מדד SIR Gain

מדד ה- Signal to Interference Ratio Gain (SIR Gain) מודד את שיפור המערכת בהפרדת הדוברים. כלומר, בכמה dB השתפר היחס בין עוצמת הדובר הרצוי לעוצמת הדובר הלא רצוי ביציאה (Output SIR) לעומת היחס בכניסה (Input SIR).

$$\begin{gathered} iSIR = \frac{RMS(p)}{RMS(b)} \; [dB] \\ SIR\,Gain = oSIR - iSIR \\ RMS(\vec{x}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \quad | \; \vec{x} = (x_1, x_2, \dots, x_N) \end{gathered} \tag{12}$$

מדד זה גדל ככל שהמערכת משפרת את יחס האנרגיות לעומת הכניסה. בעזרת מדד זה ניתן להבין כמה המערכת תרמה להפרדה ובשילוב עם המדד הקודם ניתן להסיק את $iSIR$.

5 ניסויים

5.1 מאגר הנתונים TIMIT

אותות הדוברים לקוחים ממאגר הנתונים [12] (TIMIT) DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. מאגר זה מכיל 6300 משפטים באנגלית הנאמרים ע"י 630 דוברים בעלי 8 מבטאים שונים. יותר מ-2000 משפטים שונים טקסטואלית. המשפטים נבחרו בצורה כזו שידגישו את ההבדלים בין המבטאים השונים של הדוברים, יכילו משפטים מאתגרים מבחינת ההגייה, ובעלי סגנונות הגייה מגוונים. מאגר הנתונים מחולק לסט אימון וסט מבחן. סט המבחן הינו 27% אחוז מסך המאגר (בערך 40 דקות של שמע) וסט האימון הינו 73% אחוז (בערך 100 דקות). סט המבחן נבחר בצורה כזו שהדוברים והמשפטים שונים מסט האימון. מאגר זה הינו פופולרי בספרות.

בפרויקט זה, כל סיטואציה מכילה שני דוברים המדברים במקביל ולכן דורשת שתי הקלטות לכל דוגמא. על מנת להגדיל את סט האימון, ניתן לחזור על הקלטה מספר פעמים בצרוף הקלטות שונות מסט האימון. בצורה זו, ניתן להגדיל את סט האימון בצורה משמעותית מבלי לחזור על אותה דוגמא בדיוק.
כך, בעזרת 4620 הקלטות מקוריות של המאגר ניתן ליצור מספר גדול של זוגות:

$$\binom{4620}{2} = 21{,}344{,}400$$

בניסויים שלהלן, השתמשנו בשיטה זו על מנת להגדיל את מספר הדוגמאות לסט האימון, ע"י בחירת חלק מהזוגות האפשריים.

5.2 סימולציה

כפי שהוצג בפרק 2, באפליקציה שלנו קיים חדר מהדהד עם מספר דוברים בתחומים מוגדרים. אותות הדוברים עוברים דרך הסביבה המהדהדת והרועשת ונקלטים במערך מיקרופונים. בפרויקט זה נצטמצם לבעיה בה יהיה מערך של תשעה מיקרופונים ושני אזורי קבלה כאשר בכל אזור קבלה נמצא דובר יחיד. בחלק זה, נמדל את סביבת החדר, ההדהוד, הרעש, אותות הדוברים ומערך המיקרופונים בעזרת סימולציה.

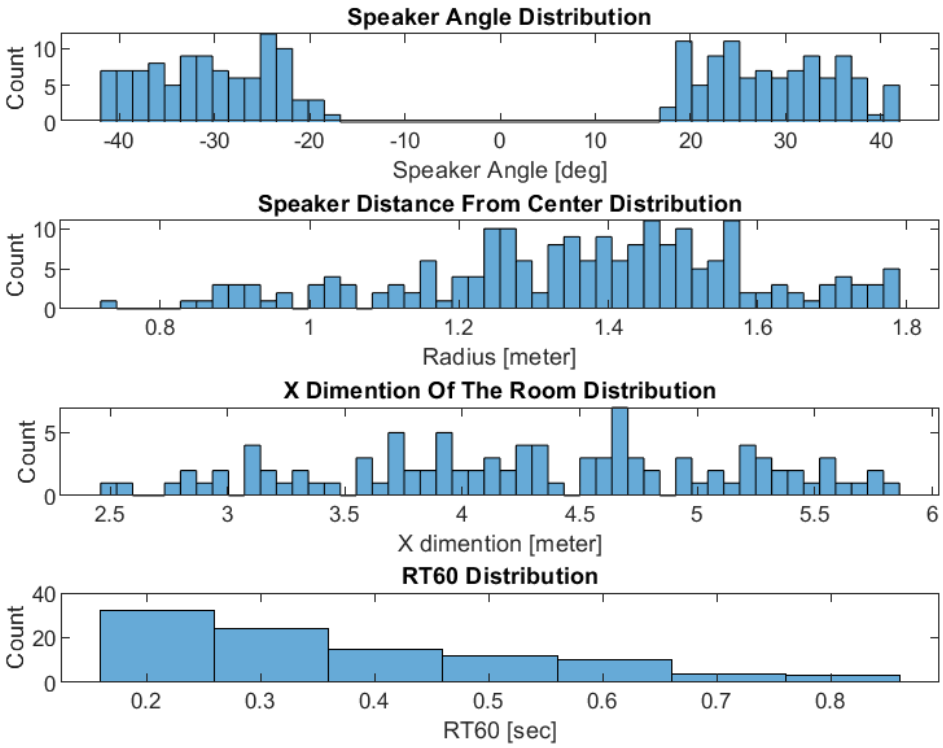
הסימולציה בנויה מהרבה "סיטואציות" המתארות מצבים אפשריים של חדר עם דוברים ומערך מיקרופונים. כל סיטואציה מוגדרת ע"י סט של 18 תגובות חדר – תגובה אחת לכל זוג מיקרופון ודובר. על מנת למדל את הדהוד אותות הדוברים, השתמשנו במודל הדהוד מסוג Image Method [16] המייצר תגובת חדר המתאימה למיקום המקור האקוסטי והמיקרופון בו הוא נקלט. בצורה זו ניתן לחשב את הגדלים שהוצגו במשוואה (1): $s_i(n)$ – אות הדובר הנקי שהגיע ממאגר TIMIT, $h_m^i(n)$ – תגובת החדר המתאימה לדובר זה ולמיקרופון ה-m.
לצורך סט האימון וסט הבוחן יצרנו 125 ו 50 סיטואציות בהתאמה. כל סיטואציה מורכבת משלל פרמטרים המגדירים אותה:

- ממדי החדר – כל אחד ממדי החדר (X,Y,Z) מוגרלים מההתפלגות הנורמלית $N(4.25,\ 1)$ הקטומה לתחום $(2.5,6)[m]$, על מנת לייצג חדרים בגדלים סבירים.
- מיקום מערך המיקרופונים – מרכז מערך המיקרופונים לא תמיד ממוקם במרכז החדר. הסטייה שלו ממרכז החדר (בכל ציר) הינה אקראית ונדגמת מההתפלגות האחידה $Uniform[-0.5,0.5]$.
- מיקום הדוברים – מיקום הדוברים מיוצג בעזרת קואורדינטות פולריות סביב מרכז החדר. רדיוס מיקום הדוברים נדגם מההתפלגות $Uniform[1.2,1.8]$, והזווית של מיקום הדוברים נדגמת מההתפלגות $Uniform[18°, 42°]$ *כאשר הזווית* $0°$ מגדירה את תחילת אזור הקבלה הרצוי.
- זמן הדהוד – זמן ההדהוד $RT60$ מוגרל מההתפלגות הפואסונית $\frac{Poisson(3)}{10}$ הקטומה לתחום $(0.2,0.8)[sec]$.
- רעש – את הרעש המתווסף לכל מיקרופון מידלנו בעזרת AWGN. עוצמת הרעש מוגרלת כך שה SNR עבור האות שנקלט במיקרופון מפולג נורמלי $N(35,\ 25)[dB]$.

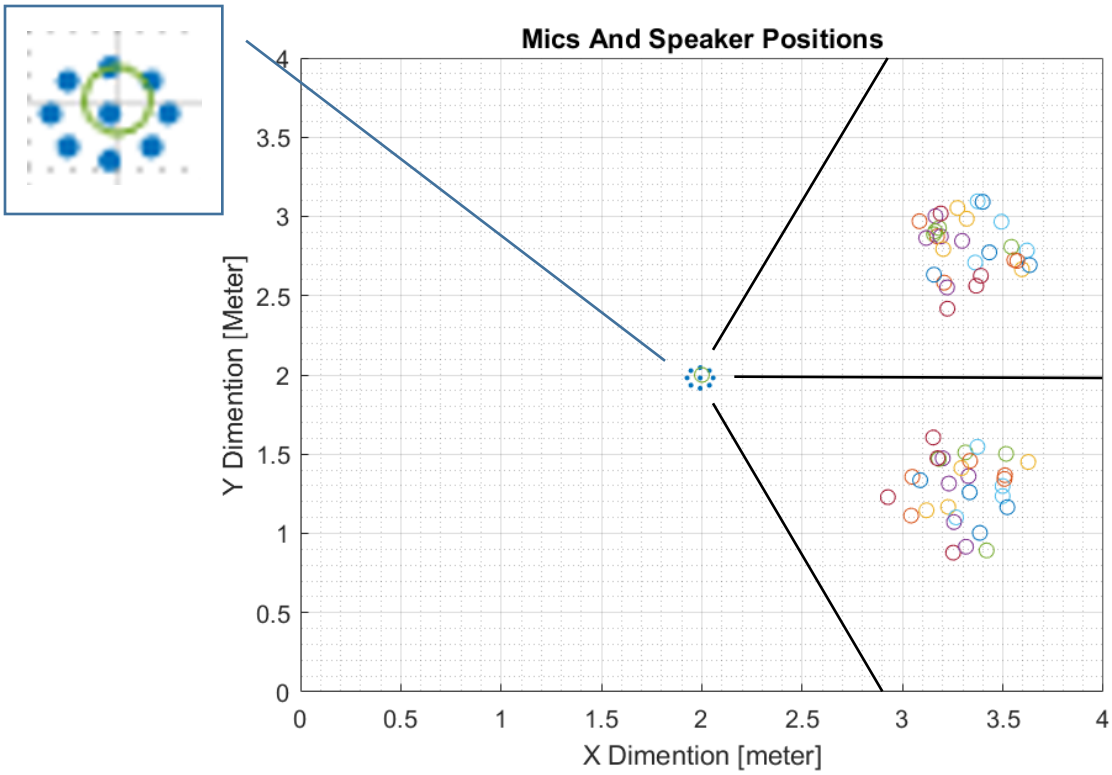
בכל הסיטואציות תחומי הקבלה נקבעים בצורה הבאה: תחום הקבלה הראשון הינו תחום הזוויות $\theta_1 \equiv [0° - 60°]$ ביחס למרכז החדר וציר x, ובאופן זהה תחום הקבלה השני הינו תחום הזוויות $\theta_2 \equiv [-60° - 0°]$.

באיור 6 מוצגות היסטוגרמות של הסיטואציות עבור הפרמטרים. ההיסטוגרמה הראשונה מתארת את התפלגות הזווית של מיקום הדוברים, ניתן לראות בהתפלגות את שני אזורי הקבלה. ההיסטוגרמה השנייה והשלישית מתארים את רדיוס מיקום הדוברים ואת ממד X של החדר. ההיסטוגרמה האחרונה מתארת את זמן ההדהוד.

איור 7 מציג דוגמא לכמה סיטואציות בהם גודל החדר ומיקום המיקרופונים הינו קבוע. נתן לראות באיור את מערך המיקרופונים (העיגולים הכחולים) שמוזז קצת ממרכז החדר (העיגול הירוק). שאר העיגולים הצבעוניים מייצגים את מיקומי הדוברים הממוקמים בתחומי הקבלה שלהם.



איור 5: היסטוגרמות המתארות דגימות של הפרמטרים המגדירים סיטואציות שונות



איור 6: דוגמא לכמה סיטואציות בהם גודל החדר ומיקום המיקרופונים הינו קבוע

כפי שצוין בפרק של מאגר הנתונים, על מנת ליצור מספר גדול של דוגמאות, בחרנו $1/300$ מסך הזוגות האפשריים, וכך נשארנו עם בערך 40 שעות לסט האימון. נציין כי הרעש מתווסף לכל דוגמא לאחר התיוג.

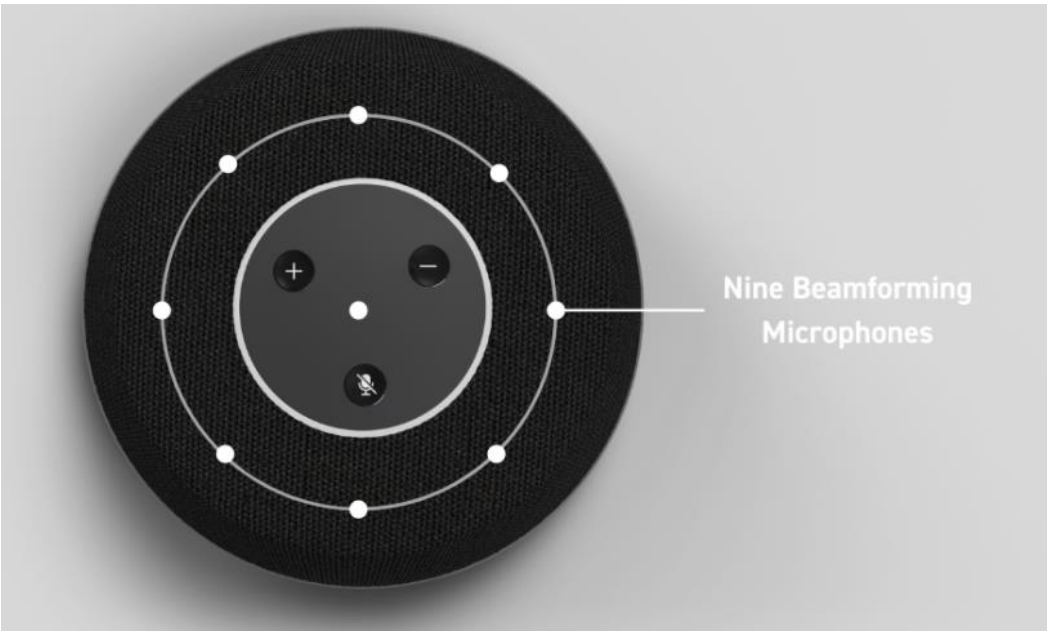
כפי שהוצג בפרק 3.1, בשלב העיבוד המקדים מתבצעת התמרת STFT לאותות שנקלטו במיקרופונים. פרמטרי התמרת STFT שנבחרו הינם:

- אורך חלון – 512 דגימות
- סוג החלון – Hanning
- חפיפה של 50%

אותות הדוברים הינם ממשיים ולכן ניתן לקטום את תמונת ההתמרה. אורך החלון הינו 512 דגימות ולכן עבור כל פריים זמני ניתן לקחת את 257 הדגימות הראשונות של ההתמרה בלי לאבד מידע. אימון רשת הקונבולוציה (פרק 3.2) מתבצע ב- PyTorch עם Adam Optimizer. מספר ה-Epoch נבחר להיות 4, משקלי הרשת נבחרים לפי נקודת המינימום של השגיאה על סט הוולידציה. גודל כל minibatch הינו $100 \times 257 \times 16$ וגודל ה-batch הינו 8. על מנת למצוא את גודל הצעד הטוב ביותר לאימון הרשת, ביצענו חיפוש איטרטיבי עד למציאת גודל הצעד הטוב ביותר.

5.3 הקלטות אמיתיות

לאחר בדיקת ההתכנות של המערכת בפרק הסימולציות, רצינו לבחון את המערכת בעולם האמיתי. לשם כך התשמשנו במערך המיקרופונים STEM TABLE של STEM ECOSYSTEM המכיל 9 מיקרופונים. שמונה מהם בהיקף מעגל ואחד נוסף במרכז המערך. באיור 7 ניתן לראות תמונה של מערך המיקרופונים.

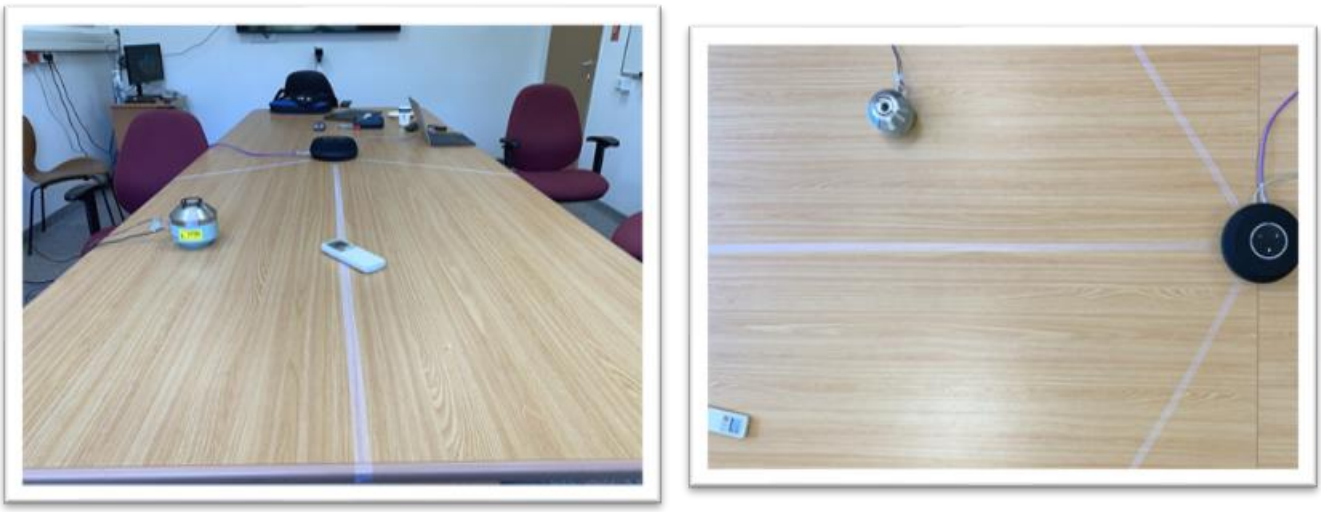


איור 7: תמונה של מערך המיקרופונים *STEM TABLE*

את אותות הדוברים יצרנו בעזרת סימולטור פה Mouth Simulator Types 4227-A של Brüel & Kjær. באיור 8 מוצגת תמונה של סימולטור הפה.

איור 8: תמונה של סימולטור הפה

לצורך אימון המערכת על מערך המיקרופונים STEM TABLE, יצרנו סט אימון בעזרת המערכות שתוארו. את סט האימון הקלטנו בחדר הישיבות של המעבדה SIPL בטכניון. מיקמנו את מערך המיקרופונים במרכז השולחן שבחדר, וסימנו שני תחומי קבלה בדומה לאלו שנבחרו בסימולציה. באיור 9 ניתן לראות תמונה המתארת את ה setup ששימש אותנו לצורך הקלטת סט האימון.



איור 9: תמונה המתארת את ה *setup* ההקלטות ליצירת סט האימון

בכל אחד משני האזורים הקלטנו 20 דקות בשלושה אזורים. בעזרת השימוש ביצירת הזוגות, כפי שתואר במאגר הנתונים, הגדלנו את סט האימון, ולבסוף נשארנו עם סט המכיל בערך 60 שעות.
האותות ממאגר הנתונים הינם בעלי עוצמה דומה, לכן על מנת ליצור גיוון בסט האימון לצורך הכללה, יצרנו הפרשי עוצמות בין הקלטות הדוברים, כך שה $iSIR$, כפי שהוגדר בפרק 4.3.3, ביניהם מתפלג נורמלי: $N(0,16)$. בנוסף, לכל מיקרופון הוספנו רעש לבן עם SNR המתפלג נורמלי: $N(35,25)$.

רצינו לבחון את ביצועי המערכת על הקלטות מחדרים שונים, כלומר כאלו שלא היו בסט האימון. לשם כך, הקלטנו סט בוחן בארבעה חדרים שונים בעלי מאפיינים אקוסטיים מגוונים:

- חדר ישיבות – אותו החדר מסט האימון, אך בקונפיגורציה שונה. המיקרופון וסימולטור הפה מוזזים ומסובבים על מנת ליצור שוני בהדהוד
- משרד 1– חדר שקט ומרווח. החדר מדמה שיחת ועידות בסביבה שקטה וללא הרבה הדהודים
- חדר מחשבים – חדר קטן בעל הרבה רעש והדהוד. חדר זה מדמה שיחה בחדר רועש בעל הפרעות
- משרד 2 – חדר קטן ושקט בעל הרבה מכשולים. חדר זה מדמה שיחה בה חלק מהדוברים חסומים ולא בעלי קו אווירי נקי למערך המיקרופונים

באיור 10 ניתן לראות את סכמת החדרים בהם התבצעו ההקלטות. חדרים שלו ממוקמים במעבדת SIPL בטכניון.
נציין, כי בכל החדרים שמרנו על setup זהה של המערכת (מערך המיקרופונים), כלומר מיקום אזורי הקבלה ביחס למערך המיקרופונים נשמר זהה. נזכיר, כי אזורי הקבלה מוגדרים כתחום זוויות ביחס למרכז מערך המיקרופונים, ולכן אזורי קבלה זהים מגדירים תחום זוויות זהה, אבל בכל תחום, מיקום הדוברים יכול להשתנות.

בכל חדר, בחרנו שני אזורי קבלה ושלושה מיקומים בכל אזור המכילים הקלטה של 5 דקות. כך, יצרנו 15 דקות של הקלטות בכל אזור קבלה. בדומה לסט הבוחן, יצרנו SIR בין הדוברים בהתפלגות דומה לזו שצוינה.
באיור 11 ניתן לראות תמונות של החדרים בהם נרכש סט הבוחן.

איור 10: סכמת החדרים במעבדת *SIPL* בהם התבצעו ההקלטות.

איור 11: תמונות של החדרים בהם נרכש סט הבוחן. בצד ימין למעלה מוצג משרד 1, בצד שמאלה למעלה מוצג חדר המחשבים, בצד ימין למטה מוצג משרד 2, ובצד שמאל למטה מוצג חדר הישיבות.

עיבוד האותות זהה לזה הנעשה בסימולציה. מספר ה-Epoch נבחר להיות 30, משקלי הרשת נבחרים לפי נקודת המינימום של השגיאה על סט הוולידציה. גודל כל minibatch הינו $16 \times 257 \times 100$ וגודל ה-batch הינו 8.

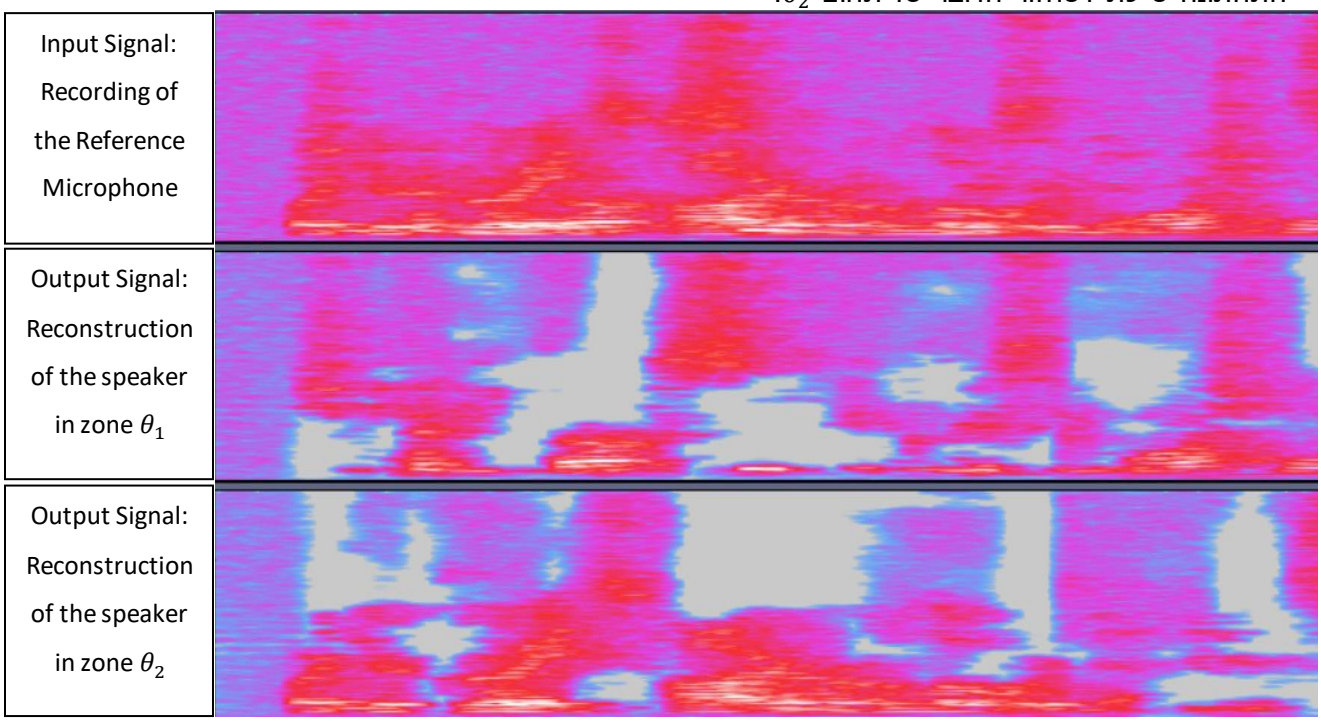
5.4 תוצאות

5.4.1 תוצאות הסימולציה

הניסוי הראשון מתאים לסימולציה שתוארה בפרק 5.2. את תגובות החדר המגדירות את הסיטואציות יצרנו באמצעות MATLAB. האותות נלקחים ממאגר הנתונים, עוברים קונבולוציה עם תגובות החדר שנוצרות בהתאם למיקום שלהם ולמיקום המיקרופונים. בשלב העיבוד המקדים מחושבות ההתמרות כפי שהוגדרו במשוואות (3),(2) מחושבים התיוגים כפי שהוצג בפרק 3.2, ומתווסף הרעש לאותות. לאחר מכן אימנו את הרשת ב- Python וקיבלנו את המסכות כפי שמוצגות במשוואה (4). את שלב הסינתזה וחישוב מדדי הביצועים ביצענו ב- Python לקבלת האותות המשוחזרים והמדדים הכמותיים.

המדד הראשון שהשתמשנו בו על מנת להעריך את ביצועי השיטה הינו מבחן האוזן הסובייקטיבי. ע"י האזנה לאותות המשוחזרים ואותות השמע המקוריים, ניתן להבין בבירור את האות הרצוי בכל אזור קבלה ולהבחין כי האות הלא רצוי מונחת בצורה ניכרת.

באיור 12 מוצגות ספקטוגרמות אותות השחזור ואות המקור. הספקטוגרמה העליונה שייכת לסכום שני הדוברים כפי שנקלטו במיקרופון הרפרנס, הספקטוגרמה האמצעית שייכת לשחזור הדובר של תחום θ_1 והספקטוגרמה התחתונה שייכת לשחזור הדובר של תחום θ_2.



איור 12: ספקטוגרמות של אות הכניסה והאותות המופרדים בסימולציה
הספקטוגרמה העליונה שייכת לסכום שני הדוברים כפי שנקלטו במיקרופון הרפרנס, הספקטוגרמה האמצעית שייכת לשחזור הדובר של תחום θ_1 והספקטוגרמה התחתונה שייכת לשחזור הדובר של תחום θ_2.

המדד השני שישמש אותנו על מנת להעריך את ביצועי האלגוריתם יהיה מדדי הביצועים שהוגדרו בפרק 4. לצורך המדד השני שישמש אותנו על מנת להעריך את ביצועי האלגוריתם יהיה מדדי הביצועים שהוגדרו בפרק 4.
בטבלה 1 מוצגים מדדי הביצועים שהופעלו על סט הבוחן עבור אלגוריתם DDESS. המדדים הינם ממוצע הערכים על המשפטים מסט הבוחן וסטיית התקן שלהם. בעמודה האמצעית דובר 1 הינו הדובר הרצוי ודובר 2 הינו הדובר הלא רצוי, ובעמודה הימנית דובר 2 הינו הדובר הרצוי ודובר 1 הינו הדובר הלא רצוי. ניתן לראות כי הן מבחינת מדד העיוות והן מבחינת מדדי ההפרדה, התוצאות טובות. נדון על כך במסקנות.

טבלה 1: מדדי הביצועים (תוחלת וסטיית תקן) על אותות השחזור כפי שהתקבלנו בסימולציה עם האלגוריתמים DDESS

Evaluation Criterion	Speaker 1	Speaker 2
fwSNR [dB]	13 ± 7	14 ± 7
Output SIR [dB]	12 ± 4	12 ± 5
SIR gain [dB]	12 ± 5	12 ± 5

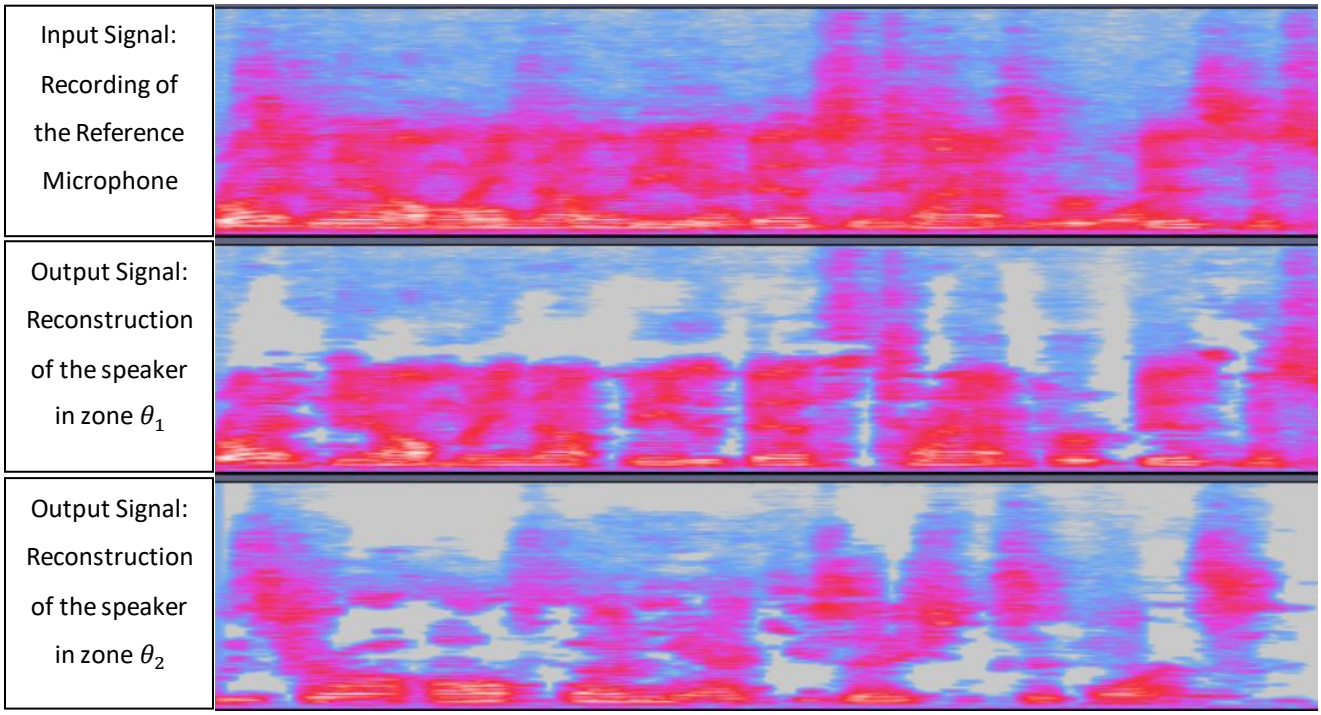
5.4.2 תוצאות ההקלטות האמיתיות

הניסוי השני מתאים להקלטות האמיתיות כפי שתוארו בפרק 5.3. אימון הרשת מתבצע על סט האימון המכיל רעש ויחס SIR כפי שתואר. בעזרת הרשת המאומנת מתבצעת הפרדה של הקלטות הדוברים מסט הבוחן. את מדדי הביצועים על סט הבוחן, הפעלנו על כל חדר בנפרד, על מנת לראות את השפעות אקוסטיקת החדר על יכולת ההפרדה.

בדומה לסימולציה, מוצגות באיור 13, ספקטוגרמות אותות השחזור ואות המקור. הספקטוגרמה העליונה שייכת לסכום שני הדוברים כפי שנקלטו במיקרופון הרפרנס, הספקטוגרמה האמצעית שייכת לשחזור הדובר של תחום θ_1 והספקטוגרמה התחתונה שייכת לשחזור הדובר של תחום θ_2.

בטבלה 2, מופיעות תוצאות מדדי הביצועים. המדדים הינם ממוצע הערכים על המשפטים מסט הבוחן וסטיית התקן שלהם. ניתן לראות כי תוצאות ההפרדה שונות בין החדרים השונים, נרחיב על כך במסקנות.

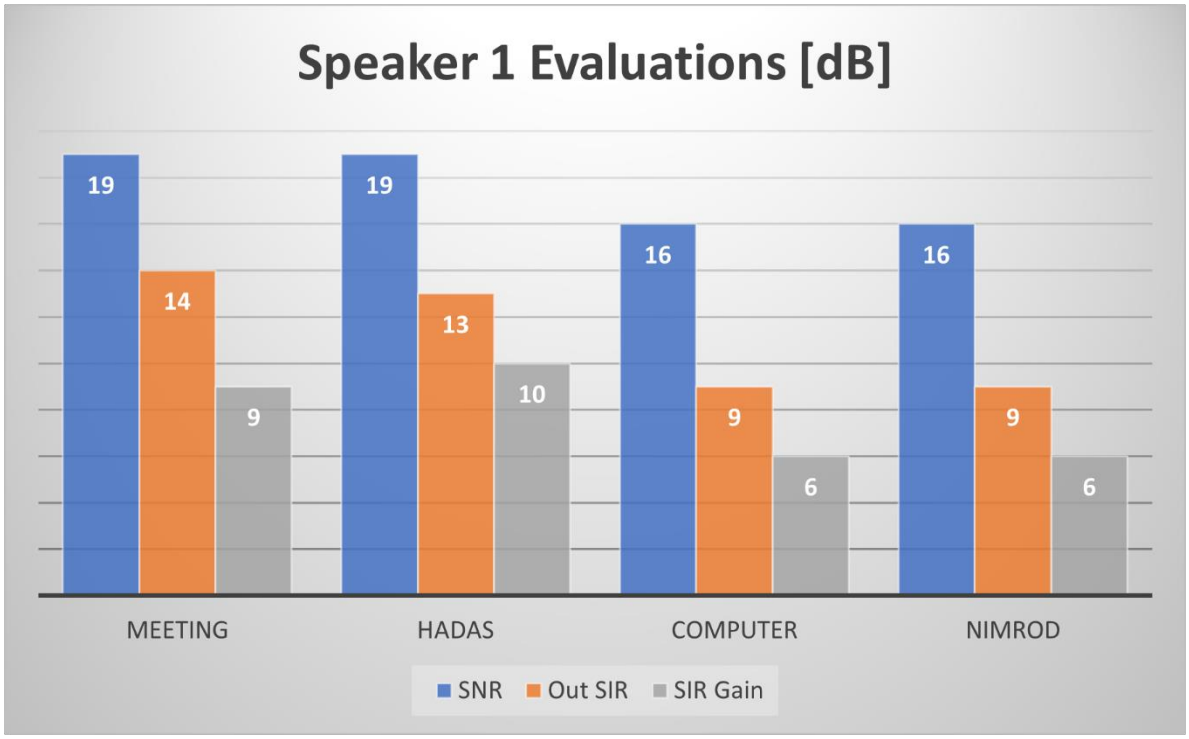
ניתן לבחון את הנתונים מטבלה 2 באיור 14. באיור זה התוצאות מוצגות בעזרת גרף עמודות וניתן להבדיל כך בין התוצאות של החדרים השונים.

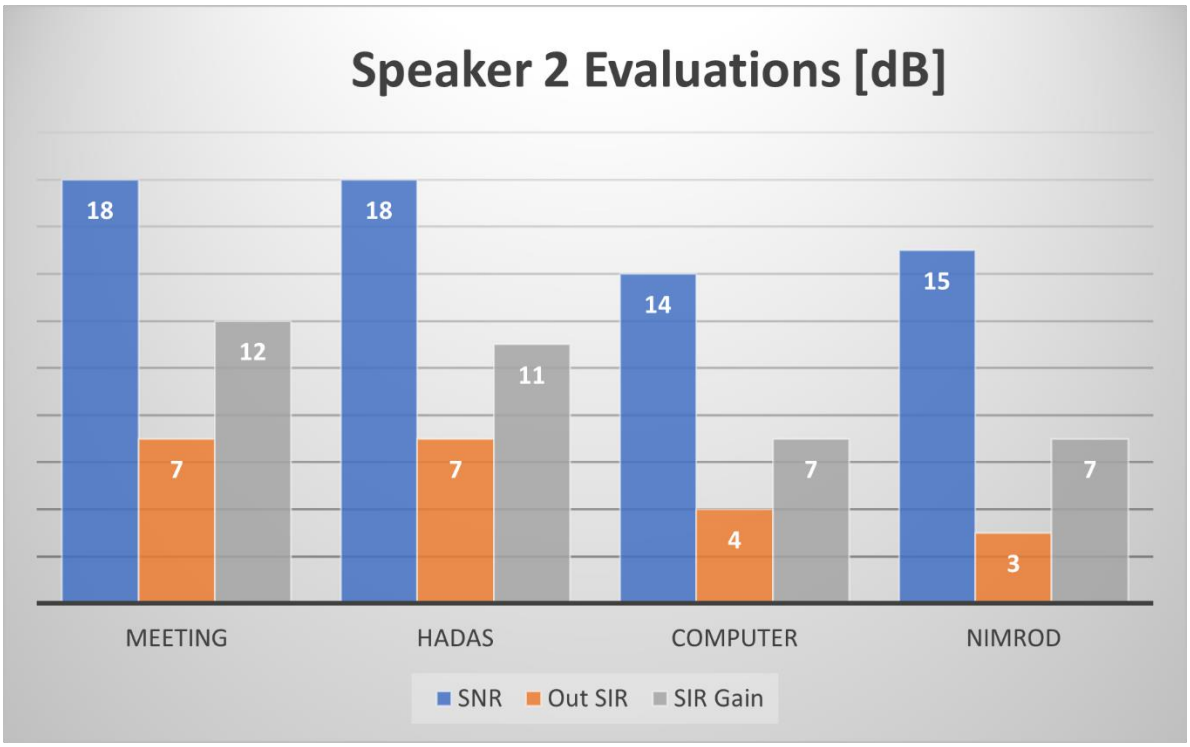


איור 13: ספקטוגרמות של אות הכניסה והאותות המופרדים בהקלטות אמיתיות.
הספקטוגרמה העליונה שייכת לסכום שני הדוברים כפי שנקלטו במיקרופון הרפרנס, הספקטוגרמה האמצעית שייכת לשחזור הדובר של תחום θ_1 והספקטוגרמה התחתונה שייכת לשחזור הדובר של תחום θ_2.

טבלה 2: מדדי הביצועים על אותות השחזור כפי שהתקבלנו על הקלטות אמיתיות עם האלגוריתמים DDESS

Speaker	Evaluation criterion	Meetings room	Hadas	Computers' room	Nimrod
1	fwSNR [dB]	19 ± 4	19 ± 4	16 ± 3	16 ± 3
	Output SIR [dB]	14 ± 3	13 ± 4	9 ± 3	9 ± 4
	SIR gain [dB]	9 ± 2	10 ± 3	6 ± 2	6 ± 3
2	fwSNR [dB]	18 ± 4	18 ± 4	14 ± 4	15 ± 4
	Output SIR [dB]	7 ± 2	7 ± 3	4 ± 3	3 ± 3
	SIR gain [dB]	12 ± 3	11 ± 3	7 ± 2	7 ± 2





איור 14: גרף עמודות המציג את תוצאות מדדי הביצועים מטבלה 3 בחדרים השונים. בגרף העליון, דובר 1 הינו הדומיננטי ודובר 2 הינו ההפרעה. ובגרף התחתון דובר 2 הינו הדומיננטי ודובר 1 הינו ההפרעה.

6 סיכום ומסקנות

6.1 דיון בתוצאות – ביצועי הסימולציה

בפרק 5.4.1 הצגנו את תוצאות ניסוי הסימולציה והצגנו את מדדי הביצועים. בחנו את תוצאות השחזור של אלגוריתם DDESS בניסוי ע"י התבוננות בספקטוגרמות (איור 7) והקשבה לאותות המשוחזרים. ניתן להסיק מהאותות המשוחזרים כי ההפרדה טובה וניתן לשמוע את האות הרצוי בצורה ברורה ומובנת. בנוסף, כמעט ולא ניתן לשמוע את אות הדובר הלא רצוי, כלומר ההנחתה מתבצעת בצורה מיטבית. מעבר לכך, ניתן לראות בספקטוגרמות כי המסיכה שנוצרה מנחיתה את אנרגיית הדובר הלא רצוי בצורה טובה.

מהאזנה של האותות המשוחזרים, ניתן לשמוע כי אחד מהאותות המשוחזרים מכיל את רוב הרעש מהכניסה, כלומר הרעש שהתווסף למיקרופון. הסבר אפשרי לתופעה זו הינו שאזורי הרעש בספקטוגרמה (אזורים ללא דיבור) מתויגים בתהליך האימון לדובר בעל האנרגיה היותר חזקה. הסיבה לכך היא שבאזורים אלו הדובר החזק יותר דומיננטי, אפילו שבמקצת. לכן, הרשת מסווגת את אזורי הרקע (הרעש) לדובר הדומיננטי וכתוצאה מכך הרעש מסווג אליו. הסבר נוסף הינו שמבחינה סטטיסטית הרשת תעדיף לסווג תא זמן-תדר לדובר הדומיננטי במידה ואין הכרעה משמעותית לגבי שייכותו. כתוצאה מכך תאים קשים לסיווג, לדוגמא תאים המכילים רק רעש, יסווגו לדובר הדומיננטי.

6.2 דיון בתוצאות – ביצועי ההפרדה בהקלטות אמיתיות

בפרק 5.4.2 הצגנו את תוצאות ההפרדה של המערכת על הקלטות אמיתיות. המערכת אומנה על סט אימון המכיל הקלטות מחדר יחיד, ואת מדדי הביצועים בחנו את סט הבוחן המכיל הקלטות מארבעה חדרים שונים. תוצאות ההפרדה של ההקלטות מחדרים שונים מחדר האימון, מצביעות על יכולת ההכללה הטובה של המערכת. כך, אימון המערכת יכול להתבצע ללא תלות בחדרים בהם היא תופעל בהמשך.

נדון בתוצאות ההפרדה בחדרים השונים: בחדר הישיבות ובמשרד 1, ראינו תוצאות הפרדה טובות ודומות לתוצאות ההפרדה בסימולציה. תוצאות אלו ראויות לציון, שהרי הקלטות הבוחן ממשרד 1 לא הופיעו בסט האימון. מכנה משותף לחדרים אלו הינו העובדה כי חדרים אלו מרווחים ושקטים, ותכונות אלו תרמו להפרדה. בחדר המחשבים ובמשרד 2 התקבלו תוצאות הפרדה פחות טובות, דבר שניתן לשמוע גם במבחן האוזן. בחדר המחשבים קיים הדהוד משמעותי ורעש חזק ברקע, תופעות אלו עשויות להסביר את התוצאות. במשרד 2, היו מכשולים בין מערך המיקרופונים וסימולטור הפה. עובדה זו גם כן גרעה מהפרדת הדוברים בחדר.

ניתן להבחין כי תוצאות ההפרדה עבור דובר 2 הינן טובות יותר מדובר 1 בכל החדרים שנבחנו. תוצאה זו צפויה, מכיוון שעקב האילוץ על SIR בין הדוברים כפי שתואר בפרק 5.3, דובר 1 בממוצע חזק יותר (במובן RMS) מדובר 2 ב- 4 [dB] .

6.3 כיוונים להמשך

לאור התוצאות המוצלחות שהתקבלו בניסוי 5.4.2, אפשר לבחון את המערכת על הפרדה של יותר אזורי קבלה ויותר דוברים. בהתאם לתוצאות שהוצגו בפרק 5.4.1 ולהסבר הניתן על סיווג הרעש, ניתן לבחון את הקונספט של הוספת יכולת סינון רעשים למערת באמצעות יצירת אזור רעש וירטואלי. פרויקט זה הוביל לפרסום מאמר אקדמי.

רשימת מקורות

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A consolidated perspective on multimicrophone speech enhancement and source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[2] E. Vincent, T. Virtanen, and S. Gannot, Eds., "Audio Source Separation and Speech Enhancement," Wiley, Sep. 2018.

[3] S. Rickard and O. Yilmaz, “On the approximate w-disjoint orthogonality of speech,” in *Proc. ICASSP*, 2002.

[4] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger and S. Gannot, "Multi-Microphone Speaker Separation based on Deep DOA Estimation," in *Proc. EUSIPCO*, A Coruna, Spain, pp. 1-5, 2019.

[5] E. Sejdi´c, I. Djurovi´c, and J. Jiang, “Timefrequency feature representation using energy concentration: An overview of recent advances,” *Digital Signal Processing*, vol. 19, no. 1, pp. 153—183, Jan. 2009.

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256-1266, Aug. 2019.

[7] S. Rickard and O. Yilmaz, “On the approximate w-disjoint orthogonality of speech,” in *Proc. ICASSP*, 2002

[8] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015

[9] I. Goodfellow, Y. Bengio and A. Courville, "Deep Learning," Cambridge, MA, USA:MIT Press, 2016, http://www.deeplearningbook.org.

[10] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *preprint arXiv:1412.6980*, 2014.

[11] Philipos C Loizou. Speech enhancement: theory and practice. 2nd ed. Boca Ratton: CRC press, 2013.

[12] P. J. Price, W. Fisher, J. Bernstein, and D. S. Pallet, "The DARPA 1000-word Resource Management database for continuous speech recognition," in *Proc. ICASSP*, vol. 1, pp. 651-654, New York, 1988.