הפקולטה להנדסה

המעבדה לעיבוד אותות

שליפת דובר רצוי בעזרת למידה עמוקה

יואב אלינסון

חן סרור

פרויקט שנה ד' לקראת תואר ראשון בהנדסה

מנחה: אביעד אייזנברג

מנחה אקדמי: פרופ' שרון גנות

אוקטובר 2024

1. מבוא

בשנים האחרונות, תחום עיבוד האותות הקוליים ראה התקדמות משמעותית בעקבות פיתוחים בטכנולוגיות למידת מכונה ורשתות עצביות עמוקות. אחת הבעיות המרכזיות בתחום עיבוד הדיבור היא היכולת להפריד בין דובר מסוים לבין דוברים ורעשי רקע בסביבה רועשת, תופעה המכונה "אפקט מסיבת הקוקטייל". בעיה זו מורכבת במיוחד כשמדובר במערכות המשתמשות במיקרופון יחיד, מה שמקשה על המערכת לזהות ולהפריד את הדובר הרצוי.

בפרויקט זה, התמקדנו ביישום אלגוריתם חילוץ דיבור ממוקד (**Target Speech Extraction - TSE**) המבוסס על ארכיטקטורה מתקדמת של **Siamese – Unet** הארכיטקטורה מאפשרת הפרדה של דוברים מתוך תערובת קולות תוך שימוש ברמזים קוליים שהתקבלו מדגימות של הדוברים. המטרה העיקרית של הפרויקט היא לבדוק את ביצועי האלגוריתם הקיים, לבצע שינויים ושיפורים, ולאחר מכן להטמיע את המודל ברובוט המעבדה לעיבוד אותות("ארי"), המיועד לעבוד בבתי חולים.

2. מטרות הפרויקט

א. שיפור ביצועי האלגוריתם במובן מינימום WER, זאת בעזרת שיפור יכולת ההבדלה בין דוברים במרחק cosine בין ייצוג הקולות של כל דובר.

ב. הטמעת המודל ברובוט רפואי אשר יוכל לבצע זיהוי והפרדה של דובר על מנת לשפר את השירות שנותן הרובוט.

3. סקירת ספרות

בסקירת הספרות הכנסנו מאמר אחד מרכזי אשר השפיע רבות על העבודה שלנו בפרויקט זה:
Neural Target Speech Extraction: An Overview (רשימת הכותבים וקישור בפרק הביבליוגרפיה)

3.1 הקשר הפרויקט

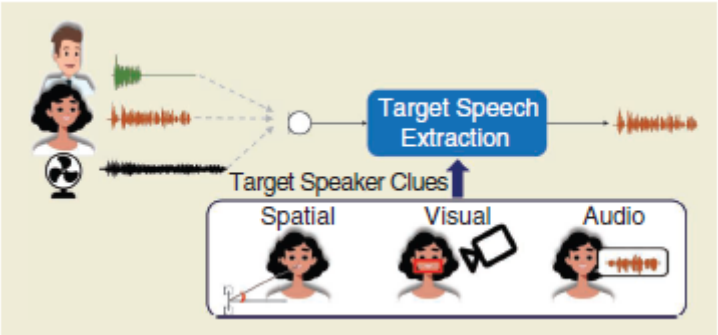
בתהליכי עיבוד אותות קוליים, האתגר המרכזי הוא חילוץ דיבור ממוקד מתוך תערובת של דוברים ורעשי רקע, מצב הקרוי לעיתים "אפקט מסיבת קוקטייל". תופעה זו מתארת את יכולתנו להאזין לדובר ממוקד בסביבה רועשת. היכולת לבודד דובר ספציפי מתוך תערובת קולות חשובה במיוחד ביישומים רפואיים, שם על רובוט רפואי להאזין ולהגיב לדיבור של המטופל בלבד.

3.2 תיאור הבעיה

בעיית חילוץ הדיבור הממוקד (**Target Speech Extraction, TSE**) עוסקת בהפרדת דיבור ממוקד מתוך תערובת של דוברים ורעשי רקע. האתגר כאן נובע מכך שגם האותות הממוקדים וגם אותות הרקע יכולים להיות דומים בתכונותיהם, כמו תדר וקצב דיבור. הפתרונות לבעיה זו מתבססים על רמזים חיצוניים כמו כיוון הדובר, תמונה של פניו, או הקלטה מוקדמת של קולו.

מבחינת רקע היסטרי המאמצים הראשונים לטיפול בבעיית חילוץ דיבור ממוקד החלו בשנות ה-80 עם טכניקות כמו עיבוד קרן (Beamforming) שפותחו להדגשת אות הדיבור של דובר אחד על פני האחרים, תוך שימוש במערכי מיקרופונים. גישה זו מתמקדת בהדגשת הכיוון של דובר ממוקד תוך נטרול רעשי רקע.

בהמשך, בשנות ה-2010, חלה התקדמות משמעותית עם הופעתן של רשתות עצביות עמוקות (DNNs). הטמעת רשתות עצביות בבעיית חילוץ דיבור ממוקד שיפרה את היכולת לבצע זיהוי דובר ולהתמודד עם רעשי רקע, אפילו כשמדובר במיקרופון יחיד. אחת התרומות החשובות של רשתות עצביות היא היכולת ללמוד דפוסים מורכבים בעזרת נתוני קול, דבר שהוביל לשיפור בביצועי המערכות.



3.3 תיאור מתמטי של בעיית ה - TSE

בעיית חילוץ דיבור ממוקד ניתנת לייצוג מתמטי של תערובת אותות כפי שמופיע בנוסחה 1 במאמר:

$$y^m = X_s^m + \sum_{k \neq s} X_k + v^m \triangleq i^m$$

כך ש:

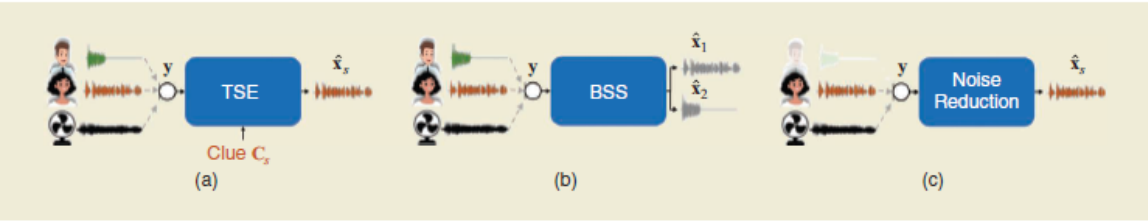
- $y^m = [y^m[0], \dots, y^m[T]] \in R^T$
- $X_s^m, X_k^m, v^m \in R^T$

הם האותות הזמניים של האות המעורבב, אות המטרה, אות ההפרעה, והרעש בהתאמה. המשתנה T מבטא את משך הזמן (מספר הדגימות), והמשתנה m את האינדקס של המיקרופון במערך המיקרופונים. s מבטא את האינדקס של של דובר המטרה וk מבטא את האינדקס של שאר הדוברים.

מטרת ה-TSE היא לשחזר את אות הדיבור של דובר המטרה תוך הפחתת הרעש ודוברי הרקע.

3.4 השוואה בין TSE ל – BSS והפחתת רעשים

בעוד BSS (הפרדת מקורות עיוורת – פירוט מרחיב ברקע תאורטי) מנסה להפריד את כל המקורות מתוך תערובת האותות ללא ידע מוקדם על הדוברים TSE מתמקד אך ורק בדובר אחד תוך ניצול רמזים שמספקים מידע נוסף (כמו כיוון או קול מוקלט). בניגוד לBSS שבו קשה להבטיח שהאות המופרד הוא של הדובר הממוקד TSE משתמש ברמזים חיצוניים כדי להתמקד באות הרצוי.



3.5 רמזים לחילוץ דיבור ממוקד

<u>רמזים קוליים</u>

רמז קולי יכול להיות הקלטה מוקדמת של הדובר הממוקד, המכילה את תכונות הקול שלו. רמזים אלו יכולים לשמש לתהליכי עיבוד נוספים כדי להפריד את הדובר הנכון מתוך התערובת.
במערכות TSE מודרניות, משתמשים בהטמעות של מאפייני קול מבוססי רשתות עצביות (כגון דוקטורים-וקטורים או i-vectors) המאפשרות זיהוי טוב יותר של הדובר גם אם הוא לא הופיע במהלך האימון של המודל.

<u>רמזים חזותיים</u>

שימוש בתמונות או בווידאו של פני הדובר בזמן שהוא מדבר מספק רמזים חזותיים לגבי תנועות השפתיים, מה שמסייע בזיהוי הדובר בתערובת. מערכות המבוססות על רמזים חזותיים מתאימות במיוחד למצבים בהם יש רמקולים עם קולות דומים (כמו בני משפחה), שם הקול בלבד אינו מספיק להפרדה.

<u>רמזים מרחביים</u>

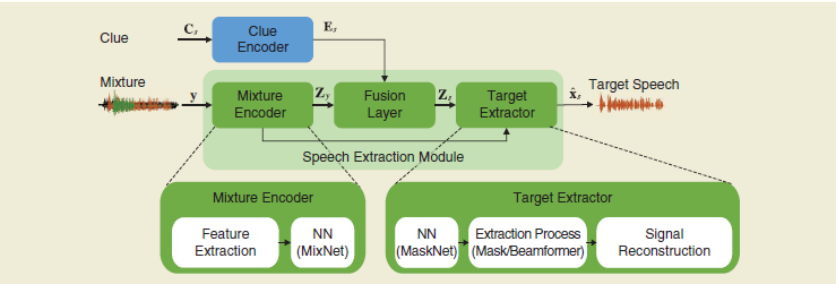
הרמזים המרחביים מספקים מידע על כיוון הדובר הממוקד במרחב, והם מועילים במיוחד במערכות שבהן יש שימוש במערכי מיקרופונים. טכניקות מבוססות מרחב יכולות לסייע בזיהוי הדובר הנכון כאשר יש מספר דוברים במיקומים שונים.

3.6 מסגרת כללית למערכת TSE עצבית

<u>מבנה המערכת</u>

מערכת TSE עצבית טיפוסית מורכבת משני רכיבים מרכזיים:

1. מקודד הרמזים (**Clue Encoder**) רכיב זה מקבל את הרמזים (קוליים, חזותיים או מרחביים) ומפיק מהם ייצוגים המאפשרים לזהות את הדובר הממוקד.
2. מודול חילוץ הדיבור (**Speech Extraction Module**) לאחר קבלת הרמזים, המערכת משתמשת בהם כדי להפריד את אות הדיבור הממוקד מהתערובת.



3.7 הרקע לבעיית חילוץ דובר

חילוץ דובר ממוקד מתערובת של דוברים אחרים ורעשי רקע הוא אתגר מרכזי בעולם עיבוד האותות, במיוחד כשמדובר במיקרופון יחיד. שימוש במיקרופון בודד אינו מאפשר לנצל מידע מרחבי, מה שמוסיף לקושי בהפרדת הדובר הרצוי. שיטות מסורתיות נוטות להתמקד בעיבוד קרן (Beamforming) באמצעות מערכי מיקרופונים, אך כאשר אין מידע מרחבי זמין, נדרשת גישה חכמה יותר המשלבת מידע קולי ישיר או רמזים חיצוניים, כמו דגימות reference של הדובר הרצוי.

3.8 השיטות הקיימות

במשך השנים, פותחו מספר אלגוריתמים לשם הפרדת דוברים, בהם Conv-Tasnet ו-DPRNN. אלגוריתמים אלו פועלים במרחב הזמן, תוך שימוש ברשתות עצביות עמוקות (DNNs) שמבוססות על קידוד אוטומטי של תכונות מתוך הנתונים. הם מציעים פתרון לבעיה באמצעות שימוש בפונקציית איבוד ניגודיות SI -SDR אשר מאפשרת לנצל את המידע הזמני בדיבור. עם זאת, רוב הגישות המודרניות מורכבות ומצריכות חישובים כבדים, מה שמקשה על אימוץ הפתרונות במערכות בעלות משאבים מוגבלים.

3.9 Single Microphone Speaker Extraction using Unified Time-Frequency Siamese-Unet

במאמר זה, עליו מתבסס המימוש בפרויקט שלנו, מציעים אביעד אייזנברג, שרון גנות ושלמה חזן פתרון לבעיית חילוץ דובר ממוקד מתוך תערובת דוברים ורעשים, תוך שימוש במיקרופון יחיד. המאמר מציג ארכיטקטורה של רשת עצבית המבוססת על Siamese – Unet הפועלת במרחב הזמן-תדר (Time – Frequency Domain) מדובר בגישה המשלבת ייצוגי תדר וזמן כדי להפיק את אות הדובר הרצוי מתוך תערובת של דוברים ורעשי רקע, בעזרת דגימת reference (Reference Signal) של הדובר הרצוי.

3.10 הארכיטקטורה המוצעת Siamese – Unet:

במאמר זה, הכותבים מציעים שימוש בארכיטקטורה מסוג Siamese-Unet המשלבת את מרחב הזמן ומרחב התדר לצורך חילוץ הדובר הרצוי. מודל זה מתבסס על שני מקודדים (Encoders) – אחד לדגימת ה- reference ואחד לאות המעורבב. כל מקודד ממפה את הקלט למרחב חבוי (Latent space) ולאחר מכן הפלטים מקודדים מוזנים למפענח (Decoder) משותף, אשר מופק ממנו האות הרצוי.

הארכיטקטורה משתמשת בייצוגים של מרכיבי התדר הממשיים והמדומים (Real and Imaginary Components) מתוך ה- STFT, הן כקלט והן כיעד לאימון המודל. ייצוג זה מאפשר להתגבר על בעיות שנוצרות במרחב התדר, כמו בעיית רעשים בתהליך ההיפוך למרחב הזמן, ומסייע לשפר את ביצועי המודל בתנאי רעש.

3.11 פונקציית איבוד (Loss) והכשרת המודל

המאמר מציין שימוש בפונקציית SI-SDR לאימון המודל, המיועדת לשימור תבניות הזמן והמרחב. הפונקציה נועדה למקסם את היחס בין האות הנקי לאות המעוות שהתקבל בתהליך ההפרדה. השימוש בפונקציה זו כפונקציית LOSS הינו בתצורת פונקציית השיפור המוכרת בשם SI-SDRi, אשר מחושבת על ידי ההפרש בין הSI-SDR של התוצאה אל מול קטע המטרה לבין התוצאה של הSI-SDR אל מול הקטע המעורבב. בנוסף לכך, נעשה שימוש בפונקציית איבוד נוספת, MSE (Mean Square Error), אשר משמשת כרגולריזציה לפונקציה העיקרית כדי לשפר את הייצוגים במרחב התדר.

4. רקע תאורטי

4.1 רשתות עצביות עמוקות – DNN's

4.1.1 מבוא

רשתות עצביות עמוקות (Deep Neural Networks – DNNs) הן כלי חשוב ומרכזי בתחום הבינה המלאכותית ולמידת המכונה. רשתות אלו מהוות הרחבה של המודל הקלאסי של רשתות עצביות מלאכותיות (ANNs) בכך שהן כוללות מספר שכבות נסתרות (Hidden Layers) בין שכבת הקלט (Input Layer) לשכבת הפלט (Output Layer) עומקן של רשתות אלו מאפשר להן ללמוד ייצוגים מורכבים ולהתמודד עם בעיות מגוונות ומורכבות כמו זיהוי תמונות, עיבוד קול וזיהוי דיבור.

4.1.2 עקרונות בסיסיים של רשתות עצביות

רשת עצבית מלאכותית נבנית בהשראת תפקוד המוח האנושי. כל יחידה בסיסית ברשת עצבית, שנקראת נוירון, מקבלת מספר קלטים, מבצעת עליהם חישוב מתמטי (בדרך כלל פונקציה לינארית), ומעבירה את התוצאה דרך פונקציית הפעלה (Activation Function) התהליך הזה מתבצע בכל נוירון ברשת.

4.1.3 מבנה של רשת עצבית עמוקה

רשת עצבית עמוקה בנויה מכמה רכיבים עיקריים:

1. שכבת קלט (**Input Layer**) שכבה זו מקבלת את הנתונים הגולמיים, כמו פיקסלים מתמונה או מאפייני קול.
2. שכבות נסתרות (**Hidden Layers**) אלו השכבות בהן מתבצע עיבוד המידע. ככל שמספר השכבות גדול יותר, הרשת נחשבת "עמוקה" יותר. כל שכבה נסתרת מורכבת ממספר נוירונים המחוברים לשכבות הקודמות והבאות.
3. שכבת פלט (**Output Layer**) השכבה האחרונה ברשת, אשר מספקת את התוצאה הסופית – למשל, החלטה אם תמונה מכילה חתול או כלב, או פלט אודיו בדגם של עיבוד קול.

4.1.4 תהליך הלמידה ברשתות עצביות

רשתות עצביות לומדות על ידי עדכון המשקלות המחברים בין הנוירונים. תהליך זה נקרא למידה מונחית (**Supervised Learning**) והוא מתבצע בדרך כלל על ידי שימוש באלגוריתם שנקרא הקטנת גרדיאנט (**Gradient Descent**) השלבים העיקריים בתהליך הלמידה הם:

1. העברת קדימה (**Forward Propagation**) הנתונים עוברים מהקלט לשכבת הפלט דרך השכבות הנסתרות. בכל שכבה מתבצע חישוב של פונקציה לינארית ואחריה פונקציית הפעלה (למשל RelU)
2. חישוב שגיאה (**Loss Calculation**) הפלט שנוצר מושווה לתוצאה הרצויה בעזרת פונקציית שגיאה (Loss Function) השגיאה מייצגת עד כמה הפלט של הרשת קרוב או רחוק מהתוצאה האמיתית.

3. העברת אחורה (**Backpropagation**) בעזרת השגיאה שחושבה, הרשת מעדכנת את המשקלות של הנוירונים באופן שמקטין את השגיאה, כך שהרשת תלמד להפיק תוצאות טובות יותר ככל שהאימון נמשך.

4.1.5 פונקציית הפעלה / אקטיבציה

פונקציות הפעלה ממלאות תפקיד חשוב בלמידה הלא לינארית של רשתות עצביות. הן מוסיפות אי-לינאריות לרשת, מה שמאפשר לה ללמוד ייצוגים מורכבים יותר של הנתונים. כמה פונקציות הפעלה נפוצות כוללות:

- RelU (Rectified Linear Unit) פונקציית הפעלה פופלרית המגדירה ערכים שליליים כאפס ומשאירה את הערכים החיוביים כפי שהם.
- Sigmoid פונקציה לוגיסטית שמכווצת את הפלט לערכים בין 0 ל-1, ומשמשת בעיקר ברשתות בעלות פלט בינארי.
- Tanh פונקציה דומה ל sigmoid אך מכווצת את הפלט לערכים בין -1 ל 1, מה שמאפשר תוצאות קצת שונות בהתאם לנתוני הקלט.

4.1.6 רשתות עמוקות ויתרונותיהן

היתרון הגדול של רשתות עמוקות הוא יכולתן להתמודד עם בעיות מאוד מורכבות על ידי למידת ייצוגים מופשטים מהנתונים. ככל שיש יותר שכבות נסתרות, הרשת יכולה ללמוד דפוסים מורכבים ועמוקים יותר.

- זיהוי תמונה ועיבוד תמונה: רשתות עצביות עמוקות הראו הצלחה עצומה בזיהוי תמונות, כאשר רשתות מסוג **Convolutional Neural Networks (CNNs)** מתמחות בזיהוי דפוסים ויזואליים כמו קצוות, צורות, ועצמים.

- עיבוד קול: רשתות מסוג **Recurrent Neural Networks (RNNs)** וגרסאותיהן המתקדמות, כמו **Long Short-Term Memory (LSTM),** מצטיינות בעיבוד סדרות נתונים, למשל בעיבוד אותות קוליים.

- רשתות גנרטיביות(**GANs**): רשתות עצביות גנרטיביות מצטיינות ביצירת נתונים חדשים הדומים לנתונים עליהם אומנו. רשתות אלו נפוצות בתחומים כמו יצירת תמונות או עיבוד טקסט.

4.1.7 אתגרים ברשתות עצביות עמוקות

למרות היתרונות הגדולים של רשתות עצביות עמוקות, הן מגיעות עם אתגרים מסוימים:

1. צורך בכמויות גדולות של נתונים: רשתות עמוקות זקוקות לכמות עצומה של נתונים על מנת ללמוד כראוי.

2. הדרישות החישוביות: אימון רשתות עמוקות דורש כוח חישוב גדול מאוד, ולעיתים קרובות משתמשים בחומרה ייעודית כמו יחידות עיבוד גרפיות (GPUs)

3. בעיה של התאמת יתר(**Overfitting**): כאשר הרשת לומדת יותר מדי את דפוסי האימון, היא עלולה לא להצליח להכליל נתונים חדשים.

4.1.8 סיכום

רשתות עצביות עמוקות הן כלי רב עוצמה לפתרון בעיות מורכבות בתחומים מגוונים, אך הן דורשות תכנון נכון, כמות גדולה של נתונים וכוח חישוב משמעותי. בתחום עיבוד הדיבור וחילוץ דיבור ממוקד, רשתות עצביות עמוקות הביאו לשיפורים משמעותיים בזיהוי והפרדת דוברים בתנאים רועשים ומורכבים.

4.2 חילוץ דיבור ממוקד – TSE

4.2.1 מבוא

חילוץ דיבור ממוקד (Target Speech Extraction - TSE) הוא תחום בעיבוד אותות קולים שמטרתו להפריד את הדיבור של דובר ממוקד מתוך תערובת של דוברים ורעשי רקע. בעיית ה TSE עולה בסיטואציות שבהן מקליטים מספר מקורות דיבור או כאשר ישנם רעשי רקע שונים, והמטרה היא לחלץ אך ורק את הדיבור של הדובר הרצוי. תחום זה מתמקד בעיקר בהפרדת דובר מסוים באמצעות שימוש ברמזים חיצוניים, כמו הקלטת קול מוקדמת, תנועות שפתיים, או כיוון דיבור.

4.2.2 אפקט מסיבת הקוקטייל (Cocktail Party Problem)

תופעת אפקט מסיבת הקוקטייל מתארת את היכולת האנושית להקשיב לדובר ספציפי בסביבה רועשת שבה ישנם דוברים נוספים. בני אדם מצליחים להתמקד בדובר אחד בזכות מנגנוני קשב סלקטיביים ומנגנוני שמיעה. האתגר בעיבוד דיבור הוא ליצור מערכות אוטומטיות שמסוגלות לחקות את היכולת הזו ולבודד דובר מסוים מתוך תערובת, גם בתנאים קשים של רעשים או דוברים חופפים.

4.2.3 עקרונות של חילוץ דיבור ממוקד

ב – TSE המטרה היא להפריד דובר ממוקד מתוך תערובת. בניגוד לבעיית הפרדת מקורות עיוורתBSS (Blind Source Separation -), שבה אין מידע מוקדם על האותות ויש צורך להפריד את כל המקורות בתערובת TSE משתמש ברמזים חיצוניים שמסייעים להתמקד בדובר הרצוי.

4.2.4 רמזים המשמשים בחילוץ דיבור ממוקד

ב- TSE נעשה שימוש במגוון סוגי רמזים המסייעים להפריד את הדובר הממוקד. להלן כמה מהסוגים המרכזיים:

1. רמזים קוליים (**Audio Clues**) רמז קולי יכול לכלול הקלטה מוקדמת של הדובר הרצוי (הקלטת זיהוי קול) או חלקים מההקלטה שבהם ידוע מראש שהדובר הממוקד דיבר. שיטות מודרניות משתמשות בהטמעות קול (Speaker Embeddings) כגון דוקטורים-וקטורים (d-vectors) או
i-vectors שמאפשרים לזהות את תכונות הקול הייחודיות של הדובר הממוקד ולחלץ אותו מתוך התערובת.

2. רמזים חזותיים (**Visual Clues**): רמזים חזותיים הם מידע המגיע מתמונות או סרטוני וידאו של הדובר הממוקד. לדוגמה, תנועות השפתיים של הדובר מאפשרות לזהות את הדובר מתוך התערובת הקולית. שיטות אלו שימושיות במיוחד כאשר הדוברים בתערובת נשמעים דומים, כגון במקרים של קרובי משפחה או אנשים עם מאפייני קול דומים.

3. רמזים מרחביים (**Spatial Clues**): רמזים מרחביים מתבססים על מיקום הדובר בחלל (לדוגמה, זווית הגעת הקולDOA) באמצעות שימוש במערכי מיקרופונים, ניתן לחלץ את הדובר הממוקד על פי כיוון הגעת הקול שלו, ובכך להפריד אותו מדוברים אחרים הנמצאים במיקומים שונים.

4.2.5 מודלים ומערכות ב – TSE

מערכות חילוץ דיבור ממוקד כוללות לרוב שני חלקים עיקריים:

1. מקודד רמזים (**Clue Envoder**) תפקידו של מקודד הרמזים הוא להמיר את המידע הקולי, החזותי או המרחבי לייצוגים מתמטיים הניתנים לעיבוד. לדוגמה, מקודד רמזים קולי עשוי להמיר הקלטה מוקדמת של קולו של הדובר לוקטור שמייצג את מאפייני הקול שלו.

2. מודול חילוץ הדיבור (**Speech Extraction Modul**) המודול הזה מקבל את ייצוגי הרמזים יחד עם תערובת הקולות, ומבצע את ההפרדה בין הדובר הממוקד לדוברים האחרים או רעשי הרקע בתערובת.

4.2.6 שימוש בטכניקות למידת מכונה ורשתות עצביות ב TSE

בין הכלים הנפוצים כיום לחילוץ דיבור ממוקד נמצאות רשתות עצביות עמוקות (DNNs), העושות שימוש בכמויות גדולות של נתונים כדי ללמוד כיצד להפריד בין דוברים שונים. שימוש ברשתות עצביות מאפשר למודלים ללמוד תבניות קול מורכבות, מה שמשפר את יכולת ההפרדה בתנאים של רעשי רקע גבוהים או דוברים רבים. בנוסף, שיטות כמו **Siamese Networks** או **Unet** (עליהם מתבסס הפרוייקט שלנו) מספקות ביצועים מצוינים בהפרדת דוברים, אפילו כאשר יש דמיון רב בין הקולות.

4.2.7 יתרונות ה – TSE

1. יכולת מיקוד בדובר ספציפי: היכולת למקד את ההפרדה בדובר אחד מספקת יתרון על שיטות שמנסות להפריד את כל הדוברים בתערובת.
2. שימוש במידע חיצוני: על ידי שימוש ברמזים חיצוניים (כמו הקלטות קודמות או וידאו), ניתן לשפר את איכות ההפרדה ולהפחית את אי הוודאות בתהליך.
3. שימוש במיקרופון יחיד: בניגוד לשיטות אחרות כמו עיבוד קרן (Beamforming) שדורשות מערכי מיקרופונים TSE יכול לפעול גם עם מיקרופון יחיד, מה שהופך אותו לשימושי במגוון רחב יותר של יישומים.

4.2.8 אתגרים ב – TSE

למרות היתרונות של TSE קיימים כמה אתגרים:

1. שונות בין קולות: במצבים בהם הקולות של הדוברים דומים מאוד (כגון קולות של בני משפחה), עלול להיות קשה להבדיל בין הדוברים ולהפריד אותם בצורה מדויקת.
2. תנאי רעש מורכבים: רעשי רקע דינמיים או רעשים בסביבה מאוד רועשת עלולים להקשות על המערכת להפריד את הדובר הממוקד בצורה מיטבית.
3. צורך ברמזים איכותיים: איכות הרמזים החיצוניים (כגון איכות הקלטת הקול או הווידאו) משפיעה באופן ישיר על איכות ההפרדה. רמזים שאינם ברורים עלולים להוביל להפרדה שגויה.

4.2.9 יישומים של - TSE

- מערכות זיהוי דיבור: משמש במערכות לזיהוי דיבור במצבים בהם ישנם דוברים ורעשים נוספים, כגון בעוזרות קוליות, שם המערכת צריכה להאזין למשתמש בודד ולהתעלם מרעשים חיצוניים.
- טלפוניה ובקרה קולית: ביישומי טלפוניה או בקרה קולית על מכשירים חכמים TSE מאפשר למערכת להגיב רק לדובר הנכון גם בסביבה רועשת.
- יישומים רפואיים: ברובוטים רפואיים או מכשירים רפואיים מבוססי קול TSE מאפשר למכשיר להתרכז רק בדיבור של המטופל ולהתעלם מרעשי רקע בסביבה רפואית דינמית כמו חדרי ניתוח או קליניקות.

4.2.10 סיכום

חילוץ דיבור ממוקד (TSE) הוא תחום מרכזי בעיבוד אותות קולים, המספק פתרונות חשובים ביישומים מודרניים בהם יש צורך להאזין לדובר יחיד מתוך סביבה רועשת. שימוש ברמזים חיצוניים יחד עם טכנולוגיות כמו רשתות עצביות עמוקות משפר את היכולת להפריד דוברים בצורה מדויקת, במיוחד במצבים של רעשי רקע מורכבים או תערובת דוברים דומים.

4.3 הפרדת מקורות עיוורת – BSS

4.3.1 מבוא

הפרדת מקורות עיוורת (BSS) היא שיטה בתחום עיבוד אותות שמטרתה להפריד אותות מרובים, המכונים "מקורות", מתוך תערובת של אותות מוקלטים, כאשר אין מידע מוקדם על המאפיינים של המקורות או על אופן הערבוב שלהם. זהו אתגר מרכזי בעיבוד אותות אודיו, במיוחד כאשר ישנם מספר מקורות קולים שמערבבים את עצמם, כמו במצבים של הקלטה עם מספר דוברים. דוגמה קלאסית לשימוש ב-BSS היא מצב שנקרא "אפקט מסיבת הקוקטייל", בו מקליטים מספר דוברים שונים בו זמנית במיקרופון אחד, והמטרה היא להפריד כל דובר מהתערובת.

4.3.2 עקרונות בסיסיים של BSS

ה-BSS מתמקד במצבים שבהם האותות המעורבים נרשמו יחד ללא גישה למידע מוקדם על תהליך ההקלטה, על מקורות האותות או על תבנית הרעש. לכן, מטרתו של BSS היא לבצע הפרדה של האותות בתנאים "עיוורים", כלומר ללא מידע מוקדם על האותות המקוריים עצמם או על הדרך בה הם התערבבו.

4.3.3 ניסוח מתמטי של בעיית BSS

בצורתה הפשוטה, בעיית ה-BSS ניתנת לייצוג מתמטי בתור:

$$As(t) = x(t)$$

כאשר:

- $x(t)$ הוא וקטור התצפיות (האותות שנקלטו)
- A היא מטריצת הערבוב (Mixing Matrix), שמתארת כיצד כל אחד מהמקורות משתקף בתצפיות.
- $s(t)$ הוא וקטור המקורות, כלומר האותות המקוריים אותם מנסים להפריד.

המטרה היא לשחזר את וקטור המקורות $s(t)$ מתוך וקטור התצפיות $X(t)$, מבלי לדעת מראש את A או את $s(t)$

4.3.4 שיטות להפרדת מקורות עיוורת

ישנן מספר גישות פופולריות לביצוע הפרדת מקורות עיוורת, כאשר כל אחת מהן מתבססת על הנחות שונות לגבי תכונות המקורות או תהליך ההקלטה:

1. ניתוח מרכיבים עצמאים: ICA – Independent Component Analysis
   ICA היא אחת השיטות הנפוצות ביותר ל BSS, היא מניחה שהמקורות הם בלתי תלויים סטטיסטית זה בזה ומנסים למצוא את מטריצת ההיפוך למטריצת הערבוב כך שכל אחד מהמקורות יופרד. כלומר, המטרה היא למצוא מטריצה W כך ש:

$$Wx(t) = s(t)$$

גישה זו מבוססת על ההנחה שהמקורות הינם בלתי תלויים, ובכך מאפשרת להפריד ביניהם על סמך תכונות סטטיסטיות בלבד ICA מצטיינת בעיקר במצבים בהם האותות המעורבים אינם גאוסיאניים.

2. ניתוח וקטור עצמאי IVA – Independent Vector Analysis:

IVA היא הרחבה של ICA למצבים בהם יש ערבוב בסביבת תדרים שונים. במצבים כמו הקלטת אודיו או עיבוד דיבור, האותות המעורבים נרשמים בתדרים שונים לאורך זמן IVA מתייחסת לבעיה בהיבט מרחבי ותדרי בו-זמנית, ומאפשרת לבצע הפרדה גם בתנאים שבהם יש ערבוב תדרים חופפים.

3. שיטות מבוססות מערך מיקרופונים Beamforming:
   שיטות אלו מנצלות מידע מרחבי, כגון כיווני ההגעה של האותות DOA – Direction of

Arrival, כדי להפריד מקורות במצבים בהם יש מערך של מיקרופונים. עיבוד קרן (Beamforming) מאפשר לכוון את מערכת הקלט לכיוון ספציפי, ובכך להפריד בין המקורות לפי הכיוון המרחבי ממנו הם מגיעים.

4.3.5 יישומים של BSS

הפרדת מקורות עיוורת משמשת במגוון תחומים, ביניהם:

- עיבוד דיבור: הפרדת דובר אחד מתוך תערובת של דוברים, במיוחד בתנאי הקלטה רועשים.
- עיבוד אותות אודיו: הפרדת מקורות מוזיקה, כאשר המטרה היא להפריד את כלי הנגינה השונים מתוך הקלטת אודיו.
- עיבוד תמונות: שימוש ב – BSS לזיהוי אובייקטים בתמונות תוך נטרול רעשי רקע חזותיים.

4.3.6 יתרונות ואתגרים ב – BSS

היתרון הגדול של BSS הוא היכולת להפריד בין מקורות ללא מידע מוקדם, מה שהופך אותו לכלי גמיש ורב-שימושי במצבים שונים. עם זאת, ישנם מספר אתגרים:

1. חוסר ודאות במבנה הערבוב: אחת הבעיות המרכזיות היא אי הוודאות לגבי מטריצת הערבוב A המקשה על ביצוע הפרדה מדויקת.
2. בעיה של סימניי פרמוטציה (**Permutation Ambiguity**) היכולת לזהות איזה מבין המקורות שהופרדו הוא המקור הרצוי עלולה להיות מוגבלת, שכן לא תמיד ברור איזה מקור שייך לאיזה דובר.
3. ריבוי מקורות וקולות דומים: כאשר יש מספר גדול של מקורות או כאשר האותות המקוריים דומים מאוד זה לזה, קשה יותר לבצע הפרדה יעילה.

4.3.7 סיכום

הפרדת מקורות עיוורת היא שיטה חזקה וגמישה המתמודדת עם בעיות של ערבוב אותות בתנאים עיוורים. היא מתאימה למגוון רחב של יישומים בתחומים כמו עיבוד דיבור, מוזיקה, תמונות ועוד. עם זאת, השימוש ב BSS דורש התמודדות עם אתגרים טכניים הקשורים לאי הוודאות במידע על המבנה המערבב והמקורות עצמם. פרוייקט עצמו לא נשתמש בBSS כלל אך הרעיון המוטיבציה הדומה של הבעיות הכריחה אותנו לקרוא וללמוד גם על נושא זה.

4.4 ארכיטקטורת Siamese-Unet

4.4.1 מבוא

Siamese-Unet היא ארכיטקטורה מתקדמת לשילוב של רשתות עצביות עמוקות (DNNs) לשם עיבוד והשוואת נתונים תוך ביצוע מיפוי מפורט בין מרחבי קלט לפלט. הארכיטקטורה משלבת שני עקרונות מרכזיים: רשת **Siamese** המשמשת להשוואה בין שני מקורות נתונים או תבניות, ו - **Unet** שמטרתה לשמר רזולוציה גבוהה בזמן שהרשת מעבדת מידע רב-שכבתי. ארכיטקטורה זו נפוצה בעיקר בתחומים כמו עיבוד תמונה, עיבוד אותות, והפרדת דוברים.

4.4.2 ארכיטקטורת Unet

UNET היא רשת עצבית שהוצגה לראשונה בשנת 2015 לעיבוד תמונה ולמשימות של סגמנטציה, אך היא הותאמה לתחומים נוספים כמו עיבוד קול והפרדת אותות. הרשת בנויה כך שהיא מורכבת מחלק "יורד" (Encoder) וחלק עולה (Decoder) התהליך מתחיל בכך שהרשת מקבלת קלט (למשל תמונה או אות קול), מעבירה אותו דרך שכבות שונות שבהן המידע דחוס ומעובד באופן עמוק יותר (בשלב הEncoder) ולאחר מכן הרשת מחזירה אותו לרזולוציה המקורית (בשלב ה- Decoder) אך בצורה מעובדת ומסווגת.

המאפיין העיקרי של Unet הוא שימוש בנתיבים מקשרים בין חלקי ה – Encoder וה – Decoder כך שבזמן שהרשת מחזירה את המידע לרזולוציה המקורית, היא משמרת את המידע המפורט שנאגר בשלבים המוקדמים של התהליך. קשרים אלו מאפשרים לרשת לבצע עיבוד רב-שכבתי תוך שמירה על מידע חיוני שנאבד ברשתות רגילות.

4.4.3 ארכיטקטורת Siamese

Siamese Neural Networks הן רשתות עצביות שמשתמשות בשני תתי-רשתות זהות, שחולקות את אותן משקלים, לצורך השוואה בין שני מקורות נתונים שונים. הרעיון מאחורי רשתות Siamese הוא לאפשר למערכת ללמוד תבניות הקיימות בשני הקלטים, ולהשוות ביניהם תוך מיצוי מאפיינים רלוונטיים.

ברשתות Siamese, שני הקלטים (שיכולים להיות תמונות, אותות קוליים או כל סוג אחר של נתונים) מעובדים דרך רשתות עצביות זהות, וכל אחת מהן ממפה את הקלט למרחב תכונות. בסיום העיבוד, מבוצעת השוואה בין התכונות של שני הקלטים. רשתות אלו משמשות לרוב במשימות של זיהוי דמיון בין אובייקטים, זיהוי פרצופים, והתאמת דפוסים מורכבים.

4.4.4 שילוב הארכיטקטורות Siamese-Unet:

ארכיטקטורת **Siamese – Unet**משלבת את עקרונות Siamese Neural Networks עם Unet כדי לספק רשת חזקה במיוחד המאפשרת גם ביצוע מיפוי מדויק של הקלטים וגם השוואה עמוקה בין מקורות שונים. הארכיטקטורה מתאימה במיוחד לבעיות שבהן יש צורך להשוות בין שני מקורות קלט תוך כדי שמירה על רזולוציה גבוהה של המידע. בתחומים כמו הפרדת דוברים, עיבוד תמונה והשוואה בין נתונים, שילוב זה מספק יתרונות חשובים.

4.4.5 מבנה ארכיטקטורת Siamese – Unet

בארכיטקטורה זו, שני קלטים עוברים דרך שני מסלולי **Unet** נפרדים אך זהים במשקליהם (כמו ברשת Siamese) התהליך מתבצע באופן הבא:

1. קלטים כפולים: הרשת מקבלת שני קלטים (למשל, שני אותות קוליים משני דוברים שונים, או תמונות ממצלמות שונות).

2. שלב **Encoder**: כל אחד מהקלטים עובר תהליך של דחיסת מידע בשלבי ה – Encoder של ה Unet. בשלב זה, המידע מפורק לתכונות מופשטות יותר, תוך איבוד פרטים מסוימים לטובת ייצוג מופשט של המידע.

3. השוואת תכונות: בשלב מסוים לאחר הקידוד, מתבצעת השוואה בין התכונות שהופקו משני הקלטים. שלב זה קריטי בארכיטקטורה של Siamese ומטרתו לזהות את הדמיון וההבדלים בין הקלטים.

4. שלב **Decoder**: לאחר ההשוואה, המידע משוחזר באמצעות ה – Decoder של ה Unet, במהלכו הרשת מחזירה את הקלטים לרזולוציה המקורית תוך שמירה על הקשרים שנבנו במהלך התהליך.

5. פלט סופי: הפלט הסופי הוא תוצאה שמייצגת את ההבדלים או הדמיון בין הקלטים, או במקרה של הפרדת דוברים – הפרדת הדובר הרצוי מתוך תערובת קולות.

4.4.6 יישומים של Siamese-Unet

1. הפרדת דוברים: בארכיטקטורת Siamese – Unet ניתן להשתמש בזוגות של אותות קוליים שנשמעים מתערובת ולחלץ את הדובר הרצוי תוך התחשבות במאפייני קולו של הדובר. מודל זה מתאים למערכות עיבוד דיבור בהן יש צורך להפריד דוברים על בסיס מאפיינים קוליים.

2. זיהוי והשוואת תמונות: ארכיטקטורה זו שימושית גם בזיהוי והשוואת תמונות, למשל זיהוי שינויים בתמונות רפואיות או זיהוי אובייקטים דומים במסדי נתונים גדולים של תמונות.

3. עיבוד אותות ביו-רפואיים: במערכות רפואיות, כמו עיבוד אותות א.ק.ג. או אותות ביולוגיים אחרים, ניתן להשתמש בארכיטקטורה זו כדי להשוות בין אותות שונים ולזהות אנומליות או שינויים במצב הבריאותי של המטופל.

4.4.7 יתרונות של Siamese – Unet

1. מיפוי מדויק: היכולת של Unet לבצע מיפוי מדויק של המידע המורכב מאפשרת ל-Siamese-Unet לטפל בבעיות הדורשות רזולוציה גבוהה ושמירה על פרטים חשובים.

2. השוואה בין קלטים: ארכיטקטורה זו מאפשרת השוואה מתקדמת בין שני קלטים, מה שהופך אותה לאידיאלית במשימות בהן יש צורך בזיהוי דמיון או הפרדה בין שני מקורות, כגון זיהוי דוברים או תמונות.

3. הפחתת עומס חישובי: מכיוון ששני המסלולים ברשת משתמשים באותם משקלים, יש חיסכון בחישובים ומורכבות הרשת מצטמצמת.

4.4.8 אתגרים

- דרישות חישוביות גבוהות: למרות ההפחתה בעומס החישובי בהשוואה לרשתות רגילות, ארכיטקטורת Siamese-Unet עדיין דורשת משאבי חישוב משמעותיים, במיוחד אם מעבדים קלטים גדולים (כמו תמונות ברזולוציה גבוהה או אותות קול מורכבים).
- הדרכה ואימון: הכשרה של רשת Siamese-Unet מצריכה כמות גדולה של נתונים, ולעיתים גם פרק זמן ארוך יותר לאימון בהשוואה לרשתות עצביות סטנדרטיות.

4.4.9 סיכום

ארכיטקטורת Siamese – Unet משלבת את היתרונות של השוואת תבניות באמצעות רשתות **Siamese** עם יכולות מיפוי המידע המדויקת של Unet. הארכיטקטורה מתאימה למגוון רחב של יישומים בעיבוד תמונה, עיבוד אותות והפרדת דוברים, והיא מספקת יתרונות חשובים בתהליך של שמירה על רזולוציה גבוהה והשוואת תכונות.

4.5 מודל WavLM

4.5.1 מבוא

WavLM הוא מודל למידת ייצוגים מונחית עצמית (Self-Supervised Learning) שפותח על ידי מיקרוסופט, במטרה להתמודד עם אתגרים בעיבוד דיבור במצבים אמיתיים הכוללים רעשי רקע ודוברים חופפים. המודל מאפשר זיהוי דיבור מדויק גם בסביבות רועשות ובתנאי אודיו מורכבים, תוך שימוש בכמות מינימלית של נתונים מתויגים. השימוש בלמידת ייצוגים ללא צורך בתמלולים רבים מאפשר הפחתה בעלויות ובזמן הנדרש לתמלול נתוני אודיו, במיוחד עבור שפות מגוונות.

4.5.2 ארכיטקטורה

הארכיטקטורה של WavLM מבוססת על גישת למידת ייצוגים של דיבור עם התאמות ייחודיות שמטרתן לספק יציבות וביצועים גבוהים יותר בתנאים של דיבור רב-ערוצי, רעשי רקע ודוברים חופפים. המודל כולל את השלבים הבאים:

1. קידוד אודיו גולמי WavLM : מקבל קלט של אודיו גולמי, ומקודד אותו באמצעות רשת קונבולוציונית (CNN) מרובת שכבות לייצוגים חבויים.
2. הסתרת ייצוגים חבויים: המודל מסווה חלק מהייצוגים החבויים בדומה למודל BERT, וכך לומד לזהות ולהשלים חלקים מוסתרים על בסיס הייצוגים הגלויים, מה שמסייע לו לפתח הבנה עמוקה יותר של ההקשר הדיבורי.
3. שימוש ב-**Transformer**: המודל מעביר את הייצוגים דרך רשת Transformer, שמאפשרת למודל ללמוד תלות ארוכת טווח בין חלקי האודיו השונים, ובכך לבנות ייצוג קונטקסטואלי עשיר.
4. איבוד ניגודיות (**Contrastive Loss**): לאחר הסתרת חלק מהייצוגים החבויים, המודל מבצע איבוד ניגודי כדי לחזק את יכולתו לזהות את הייצוג הנכון מתוך סט של ייצוגים מוסתרים, מה שמסייע לו לפתח מודעות למבנה הדיבור הכללי.

4.5.3 יישומים

- זיהוי דיבור: משפר את דיוק זיהוי הדיבור גם בסביבות רועשות או במצבי דיבור חופפים.
- זיהוי דוברים: מאפשר זיהוי דוברים במצבים של דיבור מרובה דוברים.
- סינתזת דיבור: מסייע ביצירת ייצוגים איכותיים לסינתזת דיבור טבעית יותר.
- תרגום דיבור: מקל על העברת הדיבור לשפה אחרת תוך שמירה על הקשר ותוכן.

4.5.4 יתרונות וחסרונות

<u>יתרונות</u>

- יכולת למידה מונחית עצמית: שימוש בנתוני אודיו לא מתויגים מביא להפחתה בעלויות התמלול.
- ביצועים גבוהים בסביבות רועשות: מתאים במיוחד לזיהוי דיבור בסביבות מורכבות, ומציג ביצועים טובים גם בדיבור רב-דוברי.
- דרישות נמוכות לנתונים מתויגים: יכול להניב תוצאות מצוינות עם כמות מזערית של נתונים מתויגים.

חסרונות

- תלות בחומרה חזקה: האימון והכיוונון של המודל דורשים משאבים חזקים מבחינת חומרה, כמו GPUs.
- מורכבות ארכיטקטונית: הארכיטקטורה מבוססת Transformer יחד עם איבוד ניגודיות עלולים להוסיף למורכבות הפיתוח והכיוונון של המודל.

4.5.5 אתגרים

- התאמה לשפות שונות: למרות יכולתו ללמידה מונחית עצמית, התאמת המודל לשפות מגוונות עשויה לדרוש אופטימיזציה נוספת או אימון מחדש.
- ביצועים בדיבור בלתי סדיר: זיהוי דיבור עם דוברים המשתמשים בשפת גוף או במבנים לא סדירים יכול להיות מאתגר.
- דרישות חומרה גבוהות: העבודה עם קלטי אודיו גדולים ומורכבים מחייבת שימוש בחומרה מתקדמת.

4.5.6 סיכום

WavLM הוא מודל עוצמתי ללמידת ייצוגים של דיבור, המספק מענה מצוין לאתגרים של עיבוד דיבור בסביבות רועשות ורב-דובריות. הארכיטקטורה של המודל מבוססת על למידה מונחית עצמית, מסווה ייצוגים חבויים ומסתמכת על איבוד ניגודיות כדי לפתח הבנה סמנטית עמוקה של הדיבור. המודל מהווה כלי רב עוצמה בזיהוי דיבור, זיהוי דוברים, ויישומים נוספים הדורשים ייצוגים קונטקסטואליי. לכן בחרנו במודל זה על מנת לנסות ולשפר את הפתרון המוצא לשליפת דובר.

4.6 RIR Generator

4.6.1 מבוא

RIR Generator – Room Impulse Response Generator הוא כלי שפותח על ידי מעבדות האודיו בארלנגן (AudioLabs) ומשמש ליצירת תגובת ההלם של חדרים (Room Impulse Responses) באמצעות שיטת התמונות (Image Method) שהוצעה על ידי אלן וברקלי בשנת 1979. כלי זה, המיושם

ב-MATLAB כפונקציית func_mex, מאפשר למשתמשים לשלוט בפרמטרים כמו סדר ההשתקפויות, ממדי החדר, כיווניות המיקרופון ועוד, כדי ליצור סימולציות אקוסטיות מדויקות. כמוכן ישנו מימוש בפייתון שבו השתמנו.

4.6.2 ארכיטקטורה

1. הגדרת פרמטרים: המשתמש מגדיר את מהירות הקול, תדר הדגימה, מיקום המקור והקולט, ממדי החדר, זמן ההדהוד, סוג המיקרופון, סדר ההשתקפויות, כיוון המיקרופון והפעלת מסנן מעבר גבוה.
2. חישוב תגובת ההלם: הפונקציה מחשבת את תגובת ההלם של החדר בהתבסס על הפרמטרים שהוגדרו, תוך שימוש בשיטת ההשתקפויות.

4.6.3 יישומים

- פיתוח ובדיקת אלגוריתמים לעיבוד אותות דיבור: הכלי מאפשר יצירת נתוני אימון ובדיקה בסביבות אקוסטיות שונות.
- סימולציה של סביבות אקוסטיות: ניתן לדמות חדרים עם תכונות אקוסטיות מגוונות לצורך מחקר ופיתוח.
- הערכת ביצועים של מערכות שמע: הכלי מסייע בבחינת תגובת מערכות שמע בתנאים אקוסטיים שונים.
- ייצור מאגרי מידע: בעזרת תגובות להלם שונות ניתן לייצר מאגרי מידע סינטטיים המכילים תנאים אקוסטיים רבים מאוד.

4.6.4 סיכום

RIR Generator הוא כלי חיוני למחקר ולפיתוח בתחום עיבוד הדיבור והאקוסטיקה, המאפשר הדמיה של תנאי חדר שונים. באמצעותו, ניתן ליצור נתונים עשירים ואותנטיים המשפרים את יכולת המודלים לעיבוד דיבור להתמודד עם סביבות רועשות ובלתי צפויות. בעזרת כלי זה ניתן להגיע לתנאים אקוסטיים קיצוניים על מנת ליצור מודלים אשר יכולים לעבוד בתנאים קיצוניים שכאלה.

4.7 זיהוי דיבור אוטומטי - ASR

4.7.1 מבוא

מערכות זיהוי דיבור אוטומטיות (ASR - Automatic Speech Recognition) משמשות להמרת דיבור אנושי לטקסט באופן אוטומטי, ומהוות כלי חשוב בתחומים רבים כמו עוזרים קוליים, שירותי תמלול, מערכות ניווט, ועוד.בפרוייקט שלנו השתמשנו בAPI שלDeepGram אשר משתמש באכיטקטורת Whisper **Whisper,** היא מערכת ASR מתקדמת שפותחה על ידי OpenAI. המערכת מתבלטת בדיוק הגבוה שלה וביכולת לזהות דיבור במגוון רחב של שפות, מבטאים וסביבות אקוסטיות, כולל סביבות רועשות. Whisper מתבסס על למידת עומק ומביאה יכולות חדשות ומתקדמות בתחום זיהוי הדיבור.

4.7.2 ארכיטקטורה

הארכיטקטורה של Whisper מבוסס על מודלים עצביים מסוג Transformer שעברו אימון על כמויות עצומות של נתוני דיבור. הנה כמה מהמאפיינים המרכזיים בארכיטקטורה של Whisper:

1. קידוד דיבור Whisper : מקבלת קלט בצורת אודיו גולמי (גל קול) ומקודדת אותו לייצוגים חבויים באמצעות רשתות עצביות עמוקות.
2. שימוש ב-**Transformer**: המודל מבוסס על רשת Transformer שידועה ביכולתה ללמוד תלות ארוכת טווח בין חלקי האודיו, מה שמאפשר למערכת לזהות הקשר ומבנה של שפה בצורה מדויקת יותר.
3. למידה רב-לשונית ומגוונתWhisper : אומנה על מאגר נתונים רב-לשוני עצום, המכיל הקלטות משפות ומבטאים מגוונים, מה שמאפשר לה להבין ולהמיר דיבור במגוון שפות בצורה מדויקת, גם בתנאים של רעשי רקע.
4. אימון מונחה עצמי (Whisper: **Self-Supervised Learning)** מבצעת למידת ייצוגים בדומה ל-Wav2Vec 2.0, תוך הסתרת חלק מהנתונים במהלך האימון, מה שמאפשר לה ללמוד תבניות דיבור ללא צורך במלוא הנתונים המתויגים.

4.7.3 יישומים

עוזרים קוליים Whisper : יכולה לתרום לשיפור הבנת הדיבור של עוזרים קוליים כמו Siri ו- Google Assistant.

שירותי תמלול: יכולות המודל לספק תמלולים מדויקים גם במגוון רחב של מבטאים וסביבות רעשיות הופכות אותו לכלי תמלול יעיל ואמין.

תרגום בזמן אמת: היכולת של Whisper לזהות דיבור בשפות שונות מאפשרת יישומים של תרגום דיבור בזמן אמת.

מערכות נגישות: המערכת יכולה לשפר את חוויית השימוש של אנשים עם לקויות שמיעה על ידי המרת דיבור לטקסט בצורה מדויקת ומהירה.

4.7.4 יתרונות וחסרונות

יתרונות

- דיוק גבוהWhisper : מצליחה להגיע לדיוק גבוה במגוון רחב של מבטאים ושפות.
- עמידות לרעשי רקע: המערכת מצטיינת בזיהוי דיבור גם בסביבות רועשות, מה שמאפשר שימוש רחב יותר בתנאי אמת.
- רב-לשוניות: תמיכה בשפות ומבטאים רבים מאפשרת ל-Whisper לשרת קהל רחב ומגוון.

חסרונות

- דרישות חישוביות גבוהותWhisper : מבוססת על מודל Transformer גדול, ולכן דורשת משאבים חישוביים רבים, מה שעשוי להגביל את השימוש במערכות עם חומרה חלשה.
- זמן עיבוד: עבור מערכות בזמן אמת, זמן העיבוד עלול להיות ארוך יותר עקב גודל המודל.

4.7.5 סיכום

Whisper היא מערכת זיהוי דיבור עוצמתית ומדויקת, אשר נבנתה כדי להתמודד עם אתגרים של ריבוי שפות, מבטאים ורעשי רקע. המערכת מתבססת על ארכיטקטורת Transformer מתקדמת ושיטת אימון עצמית, שמאפשרת לה להגיע לדיוק גבוה גם בסביבות מורכבות. Whisper מהווה קפיצת מדרגה בזיהוי הדיבור ומשתלבת בקלות ביישומים מגוונים כמו תמלול, עוזרים קוליים, תרגום ועוד.

4.8 VAD - זיהוי פעילות דיבור

4.8.1 מבוא

זיהוי פעילות דיבור (VAD - Voice Activity Detection) הוא תהליך שמטרתו לזהות קטעי דיבור מתוך זרם אודיו, ולהבדילם מרעש רקע או שתיקה. זיהוי פעילות דיבור הוא רכיב חשוב בתהליכי עיבוד דיבור, ומשמש במגוון יישומים כמו מערכות זיהוי דיבור אוטומטי, הקלטות קול, ושירותי שיחות קוליות. **WebRTC VAD** הוא מנגנון זיהוי פעילות דיבור שמפותח כחלק מפרויקט WebRTC של גוגל, ומיועד ליישומים בזמן אמת הדורשים יכולות זיהוי דיבור יעילות ומדויקות בסביבות רועשות. בשל תכונות העמידה בזמן אמת של מערכת זו בחרנו להתשמש בה ביישום המערכת על הפלטפורמה הרובוטית.

4.8.2 ארכיטקטורה

הארכיטקטורה של WebRTC VAD מתבססת על עקרונות פשוטים אך אפקטיביים, מה שמאפשר לו לספק תוצאות בזמן אמת עם דרישות עיבוד נמוכות:

1. חלוקת האודיו למקטעים קצרים: המערכת מחלקת את זרם האודיו למקטעים קצרים (Frames), בדרך כלל באורך של 10-30 מילישניות. כל מקטע מוערך בנפרד לזיהוי נוכחות דיבור.
2. חישוב תכונות בסיסיותWebRTC VAD : מנתח תכונות אודיו פשוטות כמו עוצמה, תדרים, ומידע ספקטרלי כדי לקבוע אם ישנה פעילות דיבור במקטע.

3. אלגוריתם מבוסס סף: המערכת משתמשת באלגוריתמים מבוססי סף (Threshold-based algorithms) כדי להחליט אם קטע מסוים מכיל דיבור. התאמת הספים מאפשרת לשלוט ברגישות המערכת לדיבור בסביבות עם רעשי רקע משתנים.
4. משאבים נמוכיWebRTC VAD : מתוכנן להיות יעיל מבחינת צריכת משאבים ולכן הוא מתאים לשימוש במכשירים ניידים וביישומי רשת, בהם חשוב להימנע משימוש מופרז במשאבים חישוביים.

4.8.3 יישומים

- שיחות קוליות בזמן אמת: זיהוי דיבור ב-WebRTC VAD מאפשר לשפר את איכות השיחה על ידי הפחתת רעשי רקע ושליחת נתונים רק כאשר יש דיבור.
- מערכות זיהוי דיבור: מערכות ASR (זיהוי דיבור אוטומטי) משתמשות ב-WebRTC VAD כדי למנוע עיבוד של שתיקה או רעש ולהתמקד רק בקטעי הדיבור.
- הקלטות קוליות: הקלטות קוליות יכולות להתבצע בצורה מדויקת יותר כאשר המערכת יודעת להקליט רק קטעי דיבור רלוונטיים.

4.8.4 יתרונות וחסרונות

<u>יתרונות</u>

- יעילות גבוהה בזמן אמתWebRTC VAD : מספק תוצאות מהירות ויכול לפעול במכשירים עם משאבים מוגבלים.
- התאמה לרעשי רקע: המערכת מתמודדת היטב עם סביבות רועשות, ומשפרת את זיהוי הדיבור גם כאשר יש רעשי רקע.
- פשטות ושילוב קל: היותו חלק מ-WebRTC הופך את השילוב במערכות זמן אמת קל ופשוט, ומספק יכולת זיהוי פעילות דיבור במגוון יישומי אינטרנט וטלפוניה.

<u>חסרונות</u>

- רגישות לדיבור חלש: המערכת עלולה להתקשות בזיהוי דיבור בעוצמה נמוכה, מה שעשוי להוביל לטעויות בזיהוי.
- דיוק מוגבל בסביבות דינמיות מאוד: בתנאים של רעשי רקע משתנים מאוד, WebRTC VAD עשוי להיתקל בקשיים בזיהוי דיבור בצורה יציבה.

4.8.5 אתגרים

- שמירה על דיוק בסביבות רועשות: התאמת האלגוריתם לסביבות עם רעש משתנה ודיבור חלש דורשת כיוונון עדין של הספים.
- עמידה בדרישות זמן אמת: האתגרים בעיבוד אודיו בזמן אמת מתבטאים בעיקר בכך שהמערכת צריכה להיות מהירה ויעילה, תוך כדי שמירה על רמת דיוק גבוהה בזיהוי פעילות דיבור.

4.8.6 סיכום

WebRTC VAD הוא כלי חשוב לזיהוי פעילות דיבור בסביבות זמן אמת, המתאפיין ביעילות חישובית וביכולת לעבוד בתנאים רועשים. המערכת מתאימה לשימוש ביישומי אינטרנט ושיחות קוליות, ומשפרת את חוויית המשתמש על ידי שליחת מידע רק כאשר יש דיבור רלוונטי. Whisper מתוכנן להצליח להתמודד עם תנאי רעש מגוונים והוא מהווה פתרון מהיר ואמין עבור יישומי קול ו-ASR בזמן אמת.

4.9 ROS (Robot Operating System)

4.9.1 מבוא

ROS (Robot Operating System)היא מערכת תוכנה המיועדת לשימוש ברובוטיקה. על אף שמה ROS אינה מערכת הפעלה במובן המסורתי, אלא מסגרת תוכנה (Software Framework) המיועדת לפיתוח רובוטים. ROS מאפשרת לחוקרים ומהנדסים לפתח, לבדוק, ולהפעיל מערכות רובוטיות מורכבות בצורה יעילה, תוך שימוש במודולים ויכולות מוכנות לשימוש. אחד היתרונות המרכזיים של ROS הוא יכולתה להפעיל רכיבים שונים ברובוט תוך שמירה על מודולריות, פעולה במקביל (Parallelism) ואינטראקציה בין תתי-מערכות שונות.

4.9.2 עקרונות מערכת ROS

1. **<u>מבנה מודולרי</u>**:

   - ROS בנוי ממודולים עצמאיים הנקראים **Nodes** (צמתים), שכל אחד מהם מטפל בתת-מערכת אחרת של הרובוט. צמתים אלו מתקשרים זה עם זה בעזרת פרוטוקול תקשורת מבוסס הודעות (Messages) בכך, ROS מאפשר לפתח ולנהל מערכות רובוטיות מורכבות תוך חלוקת התפקידים בין תתי-המערכות.

2. **<u>פרוטוקול תקשורת</u>**:

   - התקשורת ב-ROS מתבצעת דרך פרוטוקול מבוסס הודעות. צמתים יכולים לתקשר זה עם זה על ידי שליחה וקבלה של הודעות, בפרוטוקולים מוגדרים מראש. הצמתים יכולים לפרסם (publish) הודעות לנושאים (topics) מסוימים, וצמתים אחרים יכולים להירשם (subscribe) לנושאים אלה כדי לקבל את ההודעות.

3. **<u>שירותים ופעולות (Services & Actions)</u>**

   - בנוסף להודעות שמתבצעות על גבי ROS, ה – Topics מאפשרים גם תקשורת סינכרונית בין צמתים באמצעות **Services** (סרוויסים). סרוויסים משמשים לבקשות קצרות בין שני צמתים, שבהן צומת אחד שולח בקשה ומקבל תשובה.

   - **Actions** דומות לסרוויסים אך משמשות למשימות מתמשכות שמצריכות עדכון תדיר, למשל מעקב אחרי תנועה של רובוט.

4. **<u>מודולריות והתאמה אישית</u>**:

   - ROS בנוי בצורה כזו שכל חלק במערכת פועל כיחידה עצמאית. דבר זה מאפשר להחליף, לשפר, או להוסיף רכיבים למערכת מבלי לשנות את שאר המערכת. לדוגמה, ניתן להוסיף צומת חדש שמבצע עיבוד אותות קוליים, מבלי לשנות את הצומת האחראי על ניווט הרובוט.

5. **<u>תמיכה במערכות מרובות</u>**:

   - ROS תומך במערכות מבוססות רובוטים מרובים (Multi Robot Systems) המשמעות היא שמספר רובוטים יכולים לפעול יחדיו במערכת אחת, ולחלוק מידע בצורה סינכרונית דרך תקשורת מבוזרת.

4.9.3 כלים ויכולות עיקריות של ROS

1. **Rviz**:

- **Rviz** הוא כלי ויזואליזציה המיועד להציג מידע בזמן אמת על תהליכים שונים במערכת הרובוטית. ניתן להשתמש בו כדי לראות את מיקומו של הרובוט בסביבה, את חיישני הרובוט (כגון לייזרים ומצלמות), ולבחון את הפעולות שהוא מבצע.

2. **Gazebo**:

- **Gazebo** הוא סימולטור פיזיקלי תלת-ממדי שמשמש לבדיקת תכנון של רובוטים. באמצעות Gazebo ניתן לבדוק את התנהגות הרובוטים בסביבה וירטואלית לפני שמיישמים אותם בעולם האמיתי. הסימולטור משולב עם ROS ומאפשר להריץ ניסויים ולבדוק אלגוריתמים מבלי לסכן את הרובוטים בעולם הפיזי.

3. **ROS Bag**:

- **ROS Bag** הוא כלי המשמש לאיסוף וניתוח נתונים שהתקבלו מצמתים במערכת. ניתן להקליט נתונים מרשת ה – ROS לצורך ניתוח או שימוש מאוחר יותר, למשל לאימון רשתות עצביות או לניתוח ביצועי רובוט.

4. **<u>מנועי תכנון</u>** (**Planning Engines**)

- ROS מספק ממשקים למנועי תכנון המסייעים לרובוט בתכנון תנועה (Motion planning) כלים כמו **Movelt!** מאפשרים לרובוט לתכנן מסלולים ולהימנע ממכשולים תוך כדי תנועה, ולשלב זאת בתקשורת עם צמתים אחרים במערכת.

4.9.4 יישומים של ROS

1. **<u>רובוטיקה תעשייתית</u>**:

- ROS נמצא בשימוש במערכות רובוטיות בתעשיות כמו ייצור ותחבורה, במיוחד לאוטומציה של תהליכים מורכבים שדורשים אינטגרציה בין רכיבים מרובים כמו זרועות רובוטיות ומערכות חישה.

2. **<u>רובוטיקה רפואית</u>**:

- בתחום הרובוטיקה הרפואית ROS משמש לאינטגרציה בין מערכות רובוטיות לניתוחים ולביצוע משימות עדינות שדורשות דיוק רב. לדוגמה, ברובוטים רפואיים המשמשים בניתוחים בהם יש ביצוע של הליך זעיר ופולשני ROS מנהל את החיישנים והזרועות הרובוטיות בו-זמנית.

3. **<u>רובוטים אוטונומיים</u>**:

- מערכות אוטונומיות רבות, כמו רובוטי משלוח אוטונומיים או רכבים אוטונומיים, מתבססות על ROS לצורך ניווט ותיאום תנועות בזמן אמת. המודולריות של ROS מאפשרת לרובוטים לפעול בסביבה דינמית תוך קבלת החלטות על סמך מידע מחיישנים וניווט בזמן אמת.

4.9.5 יתרונות השימוש ב ROS

1. **<u>קוד פתוח ושיתוף ידע</u>:**

   - ROS הוא פרויקט בקוד פתוח, דבר שמאפשר למפתחים מכל העולם לתרום, לשתף, ולפתח מערכות רובוטיות מורכבות. הקהילה הרחבה של ROS מציעה ספריות ותוספים רבים לשימוש חופשי, מה שמאיץ את תהליך הפיתוח.

2. **<u>חיסכון בזמן פיתוח</u>:**

   - המודולריות של ROS מאפשרת למפתחים לחבר יחדיו רכיבים קיימים וליצור מערכות חדשות מבלי להתחיל מאפס. הספריות השונות של ROS מספקות פתרונות למגוון רחב של בעיות רובוטיקה, מה שמאפשר לחוקרים ולמהנדסים להתמקד בבעיות העיקריות ולא לבנות את כל התשתית מאפס.

3. **<u>תמיכה בסביבות סימולציה ובדיקה</u>:**

   - התמיכה המובנית של ROS בסימולטורים כמו Gazebo מאפשרת לפתח ולבדוק מערכות רובוטיות באופן וירטואלי, מבלי להשקיע משאבים בניסויים פיזיים יקרים ומורכבים. סימולציות אלו מדמות את סביבת העבודה האמיתית של הרובוט ומאפשרות בדיקה ואימות של האלגוריתמים.

5. יישום ומתודולוגיה

5.1 שלבי הפרוייקט:

1. הכנת הדאטה לטובת בדיקות הביצועים.
2. בדיקת ביצועים ראשונית, שימוש באלגוריתם ללא שינויים, הסקת מסקנות וסטטיסטיקות על התוצאות.
3. ביצוע שינוי באלגוריתם ואימון המודל מחדש.
4. בדיקת ביצועי האלגוריתם לאחר השינויים.
5. הטמעת המודל ברובוט ARI - אינו מותנה בשינוי האלגוריתם.

5.1.1 הכנת הדאטה לטובת בדיקות הביצועים

לטובת הפרוייקט השתמשנו ב Dataset שנקרא LRS3 – Lip Reading Sentence, מאגר זה הוא אוסף נתונים חשוב המשמש למחקר בתחום של קריאת שפתיים ועיבוד דיבור חזותי, הוא פותח על ידי אוניברסיטת אוקספורד ונמצא בשימוש נרחב במערכות של קריאת שפתיים אוטומטית, זיהוי דיבור מבוסס וידאו, ובפרויקטים הדורשים שילוב של נתונים חזותיים וקוליים יחד. ב – LRS ישנם לא רק נתונים חזותיים אלא גם קטעי אודיו איכותיים שמתעדים את הדוברים אומרים משפטים ברורים ומלאים. ה – Dataset מספק הזדמנות לעבוד על עיבוד דיבור בתנאים של משפטים קצרים ומובנים, המשמשים למגוון יישומים כמו זיהוי דיבור, הפרדת דוברים, וניתוח אותות קוליים.

5.1.1.1 <u>מאפיינים עיקריים של LRS3:</u>

1. **<u>מקור נתונים:</u>** ה – LRS3 מבוסס על וידאו מהרצאות TED (Ted Talks) וכולל קטעי וידאו והקלטות קוליות שבהם אנשים מדברים באנגלית במגוון נושאים שונים.

2. **<u>גודל ומגוון</u>:**
   - ה-Dataset כולל מעל ל-400 שעות של וידאו.
   - מכיל מעל ל-150,000 קטעי דיבור, עם משך זמן כולל ארוך למדי.
   - הדוברים המופיעים בקטעי הווידאו הם מגוונים, והם כוללים מבוגרים ממגוון גילאים, רקעים שונים, גברים ונשים.

3. **<u>תוכן</u>:**
   - כל קטע וידאו כולל דובר שמדבר מול קהל, כאשר ניתן לראות בבירור את הפנים והשפתיים של הדובר.
   - קטעי הדיבור מתועתקים (Transcriptions) וכוללים טקסט תואם לדיבור המדויק בווידאו. תיעוד זה מאפשר שימוש במודלים של עיבוד שפה טבעית יחד עם עיבוד חזותי.

4. **<u>פורמט הנתונים</u>:**

- קטעי האודיו שמורים בפורמט המתאים לעיבוד דיבור מודרני, וכוללים תמלול מדויק של מה שנאמר בקטעים. זה מאפשר לאמת את תוצאות הפרדת הדוברים ולהשוות אותן לתמלול המקורי (לצורך מדידת דיוק בעיבוד האותות).

5.1.1.2 יתרונות ה – LRS3:

- גודל ועומק: הLRS3 נחשב לאחד מה Datasets הגדולים ביותר לקריאת שפתיים ועיבוד דיבור חזותי. הוא מספק מגוון רחב של דוברים, תכנים שונים וסוגי דיבור מגוונים, מה שהופך אותו למתאים לאימון מודלים מתקדמים בעיבוד שפה חזותית.
- איכות גבוהה: הווידאו באיכות גבוהה, דבר שמאפשר ניתוח מדויק של תנועות שפתיים ופרטים נוספים על פני הדובר.
- יישומים נרחבים: השימושים שלו מגוונים, ומותאמים למשימות כמו קריאת שפתיים, זיהוי דיבור מבוסס חזות, הפרדת דוברים והשלמת דיבור.

תחילה ייצרנו טסט סט (test set) אשר שימש לבדיקת האלגוריתם שלנו ללא שינויים. לשם כך, השתמשנו בקטעי אודיו ממאגר הנתונים 3LRS אשר כפי שתואר קודם, מכיל קטעי דיבור מגוונים של דוברים שונים. בחרנו דגימות של דוברים שונים מתוך מאגר הנתונים, במטרה לבדוק את יכולת האלגוריתם להפריד דוברים במקרים של ערבוב קולות.

5.1.1.3 יצירת הטסט סט:

בשלב הראשון, חילקנו את קטעי האודיו שנבחרו לדגימות קצרות המייצגות דוברים שונים, כך שכל דגימה תכיל משפט קצר שנאמר על ידי דובר אחד מתוך ה-3LRS. לאחר מכן, לצורך יצירת סימולציה של תופעת הדהוד (**Reverberation**) והתגובות האקוסטיות הנלוות, הגרלנו תגובות אקוסטיות (Impulse Responses) על פי זמן הדהוד (60RT) ומיקום הדוברים בחלל. תגובות אלו נועדו לדמות את האפקט של הדהוד בסביבות שונות.

5.1.1.4 ערבוב דגימות אודיו

בשלב הבא, ערבבנו את הדגימות שהכנו על ידי יצירת תערובות אודיו רנדומליות. בכל ערבוב, יצרנו מיקס (Mix) של שתי דגימות שונות, כך שכל תערובת כוללת שני דוברים המדברים יחד, כאשר הדוברים בתערובת נבחרו באופן אקראי מתוך הדגימות המקוריות. בנוסף לכך, שמרנו דגימת רפרנס אחת של אחד הדוברים בכל מיקס. דגימת רפרנס זו משמשת במהלך השלב של חילוץ הדובר הרצוי מהתערובת.

5.1.1.5 בחירת המאפיינים האקוסטיים של החדר

את התגובות להלן יצרנו בעזרת הסימולטור – RIR Generator כאשר גודל החדר היה 4x6x6, וזמן ההדהוד הוגרל מהתפלגות אחידה $R_{T60} \sim U(0.2,0.8)$. כמוכן גם מיקומי הדוברים הוגרלו בהתפלגות אחידה על צירי הx,y של החדר. עבור כל דוגמא בסט הגרלנו חדר לדוברים וחדר לרפרנסים, כאשר המיקומים היו שונים עבור כל מקרה.

5.1.2 מדידת ביצועי האלגוריתם

ביצוע הניסויים:

לאחר יצירת הסט המעורבב, הרצנו את כל קטעי האודיו בתערובות דרך האלגוריתם. במהלך ההרצה, מדדנו מספר פרמטרים כדי להעריך את ביצועי המודל:

1. מדד **SISDR (Scale-Invariant Signal-to-Distortion Ratio)** – מדד זה משמש להערכת איכות ההפרדה בין הדוברים, ומודד את היחס בין עוצמת האות המופרד לעיוותים שנגרמו במהלך תהליך ההפרדה. את המדד הזה נרצה למקסם על מנת לקבל אות מופרד טוב ונקי מעיוותים.

2. מדד **WER (Word Error Rate)** – מדד זה בוחן את מידת הדיוק בזיהוי המילים של הדוברים בתוצאה הסופית על ידי מערכת זיהוי דיבור אוטומטית (ASR). מערכת הASR שהשתמשנו בה בניסויים הינה deepgram. ככל שהערך נמוך יותר, כך האלגוריתם מצליח לזהות את הדובר בצורה מדויקת יותר לאחר ההפרדה.

3. המרחק בין הייצוגים של שני הדוברים – שמרנו את הייצוג של הרפרנס של הדובר הנוצר על ידי האלגוריתם ומדדנו את מרחק הקוסינוס בין הייצוגים של הדוברים המעורבבים בכל סצנה.

$$(x, y) = 1 - \frac{x \cdot y^T + 1}{2}$$

ניתוח התוצאות, הסקת מסקנות וסטטיסטיקות:

<u>תוצאות WER על האודיו המופרד מול האודיו המעורבב:</u>

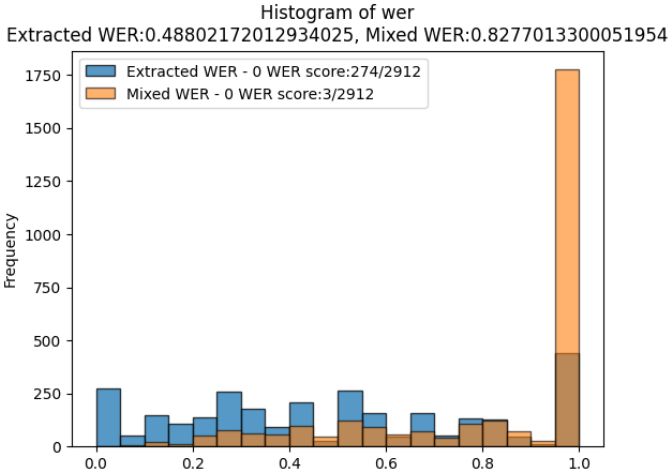


Figure 5.1.3.1

ניתן לראות כי ישנו שיפור משמעותי בWER עבור האודיו המופרד, אך התוצאות אינן מושלמות משום שבממוצע על כל הסט אנו מגיעים למעל 48% שגיאה.

על מנת לבסס את הכיוון למחקר שלנו רצינו להראות כי ישנו קשר בין מרחק הקוסינוס בין ייצוג הדוברים לבין טיב ההפרדה, הן מבחינת sisdr והן מבחינת WER.יצרנו גרף המציג בדיוק זאת:

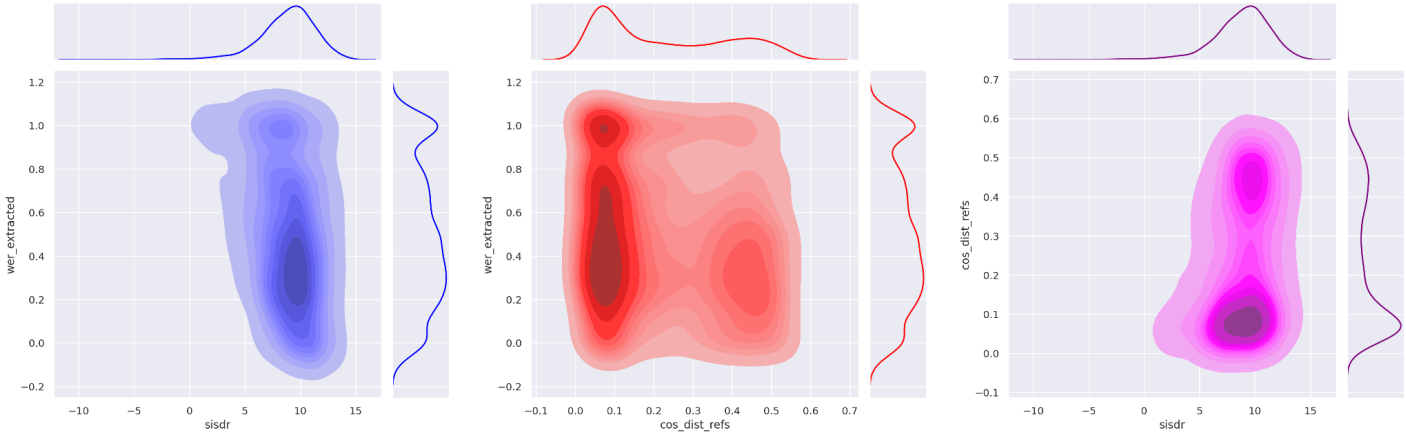


Figure 5.1.3.2

ניתן לראות בגרף זה כי רוב הדוגמאות אינן רחוקות מבחינת מרחק קוסינוס אך למרות זאת תוצאות הsisdr מרוכזות סביב הערך $10dB$ – אשר מצביע על הפרדה טובה. בהתחשב בתוצאות שגיאת הASR שהצגנו, ניתן להסיק כי אם נחזק את יכול ההבדלה של האלגוריתם בין דוברים שונים אולי גם נוכל לשפר את תוצאות הASR.

5.1.3 ביצוע שינוי באלגוריתם ואימון המודל מחדש

עלו מספר אופציות לשיפור האלגוריתם בביצועי WER:

- הוספת פונקציית loss אשר כוללת בדיקת ASR של האודיו המופרד תוך כדי אימון (ctc loss) ובכך להכריח את המודל ליצור אודיו אשר משיג WER טוב יותר.
- הוספת מודל ייצוג נוסף אשר אומן להבדיל בין דוברים (למשל ecapa).
- אימון המודל התצורה הקיימת תוך הוספת מגבלה על מרחק הקוסינוס בין הדוברים.

אופציית הctc-loss לא נבחנה משום שהיא מגבילה את האלגוריתם לפתרון ASR יחיד. כמוכן, מגבלה על מרחק הקוסינוס יכולה להיות מסובכת ולכן בחרנו לנסות לאמן את המודל על מודל ייצוג דוברים נוסף.

בחרנו להשתמש במודל wavlm, ספציפית בחרנו במודל שאומן למשימת הבדלה בין דוברים – wavlm for x-vector.

כמו שהזכרנו ברקע לפרוייקט, הרפרנס עובר ברשת הסיאמית ומתקבל הייצוג של הדובר שבעזרתו נשלף הדיבור הרצוי. על מנת לשלב את המודל החדש באלגוריתם בחרנו להוסיף שכבה (fc) אשר תקבל את הייצוג של האלגוריתם הנוכחי ואת הייצוג של המודל החדש ותלמד את השילוב הטוב ביותר בין השניים. משלב זה ואילך האלגוריתם זהה לאלגוריתם המקורי.

מפאת חוסר הזמן הנובע משירות מילואים ארוך במהלך הסמסטר, תהליך האימון לא הושלם.
בכל מקרה האימון רץ מספר epochs והגיע לתוצאות sisdr סבירות:

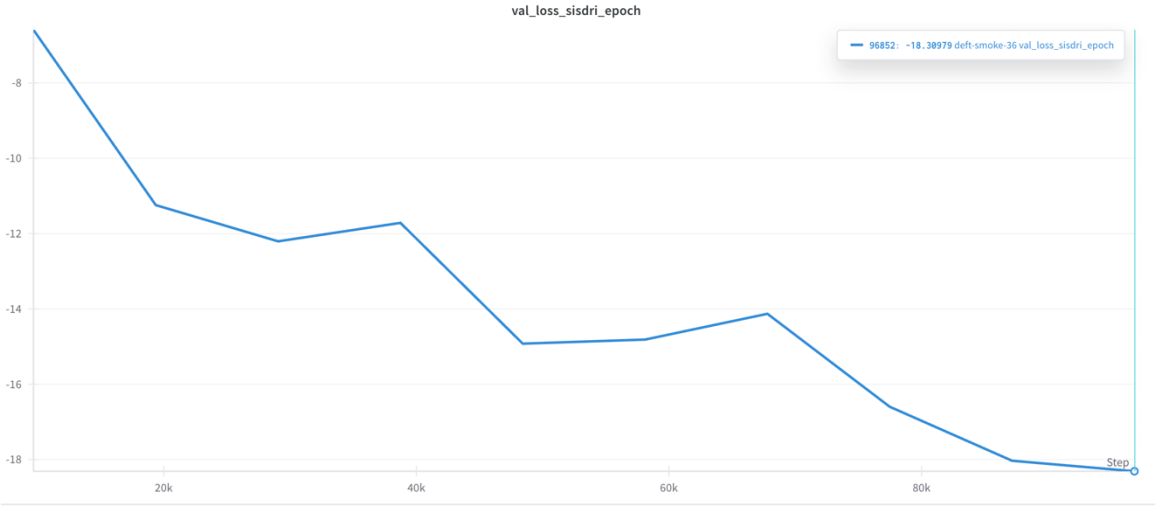


Figure 5.1.3.1

לצערנו אימון זה קרס לאחר 10 אפוקים (הפסקת חשמל כנראה, קרה לאחר ההקפצה למילואים). בדקנו את המודל על דוגמאות מסט הבדיקה ונראה כי המודל לא למד מספיק (המודל באלגוריתם המקורי עבר כ370 אפוקים).

5.1.4 הטמעת המודל ברובוט ARI

על מנת להטמיע את האלגוריתם לros ולרובוט ARI בפרט נדרשנו לבצע מספר דברים:

5.1.4.1 יצירת מאורע של המודל אשר מחשב את ייצוג הדובר.

5.1.4.2 יצירת מאורע של המודל אשר בהנתן ייצוג דובר רצוי מחשב את שליפת הדובר.

5.1.4.3 מימוש חלון נע – עבור כל מסגרת חדשה שמגיעה נבצע inference למאורע המוזכר בסעיף 2.

5.1.4.4 מימוש תהליך הרשמה. במידה ואין דובר רצוי ברגע נתון, ואכן מזוהה דיבור, קול הדובר יועבר דרך מאורע המודל מסעיף 1 והייצוג ישמר. לאחר מכן הקול של אדם זה ישלף מרגע זה עד לרגע בוא יוחלט על דובר רצוי אחר או על הפסקת השליפה.

5.1.4.1 **<u>יצירת מאורע של המודל אשר מחשב את ייצוג הדובר:</u>** קלאס פשוט שכל מטרתו היא להחזיר את הייצוג המחושב על ידי המודל. כמוכן למודל זה יש מודל vad אשר מודא שאכן בדגימה יש דיבור.

```python
class SpeakerEmbedder:
    '''
    1. enroll speaker service
    2. publish embedding to speaker extraxtion node
    3. work on improving extraction with length and normalization of embedding.
    '''
    def __init__(self,debug=False) -> None:
        self.embedder_model = Embedder(device=DEVICE)
        self.vad_eval= webrtcvad.Vad(3)
        self.emb = None
        self.debug = debug

    def close_sub(self):
        self.sub.unregister()
        self.emb = None

    def enroll_speaker(self):
        '''
        1. start recording audio from an input topic of choice, default : /audio/raw_audio
        2. when vad is done or time is above 5 sec stop recording.
        3. embed voice
        4. set publish embeddings to speaker_extractor
        '''

        self.emb =None
        self.buffer = np.zeros((MSG_COUNT,512),np.int16)
        self.msg_count = 0
        self.vad_done = False
        self.zero_vad = 0
        self.sub = rospy.Subscriber(INPUT_TOPIC,RawAudioData,self._enrollment_callback)
```

Figure 5.1.4.1.1

חישוב הייצוג מתבצע על ידי הפונקציה:

```python
def embed_ref_audio(self,ref_audio,fs=16000):
    if not isinstance(ref_audio, torch.Tensor):
        ref_audio = torch.tensor(ref_audio, device=DEVICE)
    if not torch.is_floating_point(ref_audio):
        ref_audio = ref_audio.float()/2**15
    if type(ref_audio) != torch.FloatTensor:
        ref_audio = ref_audio.float()
    if fs != 8000:
        ref_audio = F.resample(ref_audio,orig_freq=fs,new_freq=8000)
    ref_audio = ref_audio*0.9/max(abs(ref_audio))
    emb = self.embedder_model.extract_embedding(ref_audio)
    return emb
```

2Figure 5.1.4.1.

5.1.4.2 **<u>יצירת מאורע של המודל אשר מחשב את שליפת הדובר:</u>** גם כאן מדובר בפתרון יחסית פשוט, כאשר לקלאס זה יש מאורע של הקלאס שהוגדר בסעיף 1. כמוכן המצב ההתחלתי הוא ללא דובר רצוי. נראה בהמשך גם שכאשר אין דובר רצוי לשליפה האלגוריתם לא מבצע כלום – כלומר מעביר את האודיו כמו שהגיע.

```python
class SpeakerExtractor():
    def __init__(self) -> None:
        rospy.init_node('speaker_extractoion')
        self.device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
        self.extractor_model =  Extractor(device=self.device)#self.hp,extractor_model,self.device)
        self.extractor_model.has_reference=False
        self.frames = 12
        self.input_buff = InputBuffer(size=512*self.frames,dtype=torch.int16,device=self.device)
        self.gcc_buff =InputBuffer(size=512*self.frames,dtype=torch.int16)
        self.p = AudioPublisher()
        self.enrollment = SpeakerEmbedder(debug=True)
        self.set_ref_service = rospy.Service('/set_voice_ref',Trigger,self.set_ref_srv)
        self.remove_ref_service = rospy.Service('/remove_voice_ref',Trigger,self.remove_ref_srv)
        # self.speaker_loader = Loader()
        self.latch = False
        rospy.loginfo('Loading model without a known refrence voice')
        rospy.Subscriber(INPUT_TOPIC,RawAudioData,self.callback,queue_size=30)
```

2.1Figure 5.1.4.

5.1.4.3 **<u>מימוש חלון נע:</u>** בrosa כל פעולה מתבצעת על ידי פונקציית callback אשר נקראת בכל הגעה של הודעה חדשה. אנשחנו הגדרנו שהודעה תהווה מסגרת לטובת האלגוריתם, ולכן בכל הגעת מסגרת השליפה תחושב. ישנן תוספות הנוגעות לאדפטציה לרובוט – כמו השתקט הדיבור של הרובוט בעצמו (משום שלא מומש בו eco- cancelation מוצלח).

```
    def callback(self,msg):
        if INPUT_TOPIC == '/audio/raw_audio':
            s = np.array(msg.data).reshape(512,6)[:,5]
            frame = np.array(msg.data).reshape(512,6)
            sep_frame = torch.from_numpy(frame[:,1]).to(device=self.device)
            gcc_frame = torch.from_numpy(frame[:,2])
        else:
            s = 0
            sep_frame = torch.tensor(msg.data)
            gcc_frame = torch.zeros((512,))
        if len(np.unique(s)) > 1:
            self.ari_speech = 'ARI'
        else:
            self.ari_speech = None

        # self.gcc_buff.push(gcc_frame)
        self.input_buff.push(sep_frame)
        if self.input_buff.is_full():
            data = self.input_buff.try_read_buffer()
            # gcc = self.gcc_buff.try_read_buffer()
            gcc = None
            if data is not None:
                audio = self.extractor_model(data,fs=16000)
                self.data_tuple =
(audio,data,gcc,self.ari_speech,'voice_mock')#self.speaker_loader.target_voice)
                publisher_thread = Thread(target = self.send_frames_to_publisher)
                publisher_thread.start()
            else:
                raise EmptyBufferError(self)
```

Figure 5.1.4.3.1

5.1.4.4 **<u>מימוש תהליך הרשמה:</u>** בros ניתן להגדיר service אשר יכול להקרא מכל מחשב אשר מחושב לרובוט. לכן בחרנו באופציה זו למימוש תהליך ההרשמה, כאשר היא נקראת – קול הדובר נשמר עד לקבלת מידע מספיק לשליפה (הגדרנו זאת ל35ממסגרות אך זה ניתן לשינוי).

```
    def set_ref(self,ref):
        rospy.loginfo('Extractor has a target speaker')
        self.extractor_model.set_ref(ref)

    def set_ref_srv(self,request):
        self.enrollment.enroll_speaker()
        t = 0
        while self.enrollment.emb == None: #wait for a valid
embedding   rospy.sleep(0.1)
            t += 0.1
            if t > 15: #to avoid waiting for ever
                break

        emb = self.enrollment.emb
        if emb != None:
            self.set_ref(emb)
            return TriggerResponse(success=True,
                message='New Voice embedding added')
        else:
            self.enrollment.close_sub()
            return TriggerResponse(success=False,
                message='No Voice embedding found')

    def remove_ref_srv(self,request):
        self.extractor_model.remove_ref()
        return TriggerResponse(success=True,
                message='Refrance voice removed. Back to bypass-mode')
```

Figure 5.1.4.4.1

לאחר שיצרנו את אותן מחלקות ופונקציות, עטפנו הכל בחבילת ros והרצנו על הרובוט לבדיקה.

כל הקוד למימוש נמצא כאן:
https://gitlab.inria.fr/spring/wp3_av_perception/1ch_speaker_extraction

6. תוצאות ודיון

ניתן לראות (מהמצגת) שהמימוש על הרובוט אכן עובד ואכן מצליח לשלוף את הדובר הרצוי, זאת למרות הירידה בביצועים כאשר השליפה נעשית בזמן אמת. בנוסף לכך ראינו שעבור שפות שונות (אנגלית צרפתית ועברית) המודל מצליח לשלוף בצורה טובה – עובדה מפתיעה שלא תמיד ניתנת להשגה במודלים עמוקים כאלו, לרוב אנו מצפים לקריסה בשינויים דרסטיים כאלו.

7. מסקנות ועבודה להמשך

מהפרויקט עלו מסקנות רבות על עבודת האלגוריתם וכמוכן עלו הצעות רבות לשיפור (חלקן מומשו וחלקו ימומשו בעבודה עתידית).

ראשית, שיפור השליפה על ידי מודל שאומן להבדלה בין דוברים נראה אפשרי וצריך אימון נוסף על מנת להראות אפקטיביות. כמוכן יש לשקול מודלים נוספים חוץ מהמודלים שהוצעו. מעבר לכך, ניתן גם לאמן את המודל לא השלב השני על מנת למנוע עיוותים בעת הורדת ההדהוד.

שנית, מימוש המודל על הפלטפורמה הרובוטית היה מאתגר. הן בהיבט התאמת האלגוריתם הקיים לעבודה בזמן אמת על ידי חלון נע. והן בהיבט התאמת האלגוריתם לעבודה בתצורה סיאמית כאשר לא תמיד מתקבלות שתי כניסות (לעיתים ישנו רפרנס קיים וישנה רק כניסה של אודיו, ולפעמים ישנו רק אודיו שהופך לרפרנס). אך אנו מרוצים מהמימוש ומהעובדה שניתן להפעיל את האלגוריתם על הרובוט.

8. מקורות

Zmolikova, K., Delcroix, M., Ochiai, T., Kinoshita, K., Černocký, J., & Yu, D. (2023). Neural Target Speech Extraction: An Overview. IEEE Signal Processing Magazine, May 2023.

Eisenberg, A., Gannot, S., & Chazan, S. E. (2023). Single Microphone Speaker Extraction Using Unified Time-Frequency Siamese-Unet. Proceedings of IEEE, March 2023.

Ronneberger, O., Fischer, P., & Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation." *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. https://arxiv.org/abs/1505.04597

Koch, G., Zemel, R., & Salakhutdinov, R. (2015). "Siamese Neural Networks for One-shot Image Recognition." *ICML Deep Learning Workshop*.

Afouras, T., Chung, J. S., & Zisserman, A. (2018). "LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition." *ArXiv preprint arXiv:1809.00496*. https://arxiv.org/abs/1809.00496

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." *NeurIPS 2020*. https://arxiv.org/abs/2006.11477

Quigley, M., Gerkey, B., & Smart, W. D. (2015). "Programming Robots with ROS: A Practical Introduction to the Robot Operating System." *O'Reilly Media*.

Koenig, N., & Howard, A. (2004). "Design and Use Paradigms for Gazebo, An Open-Source Multi-Robot Simulator." *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... & Wei, F. (2022). "Wavlm: Large-scale self-supervised pre-training for full stack speech processing". *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505-1518.

Wikipedia – ROS, DNN's, TSE, BSS and more

9. מילון מונחים

ראשי התיבות	משמעות	תרגום	הסבר נוסף
TSE	**T**arget **S**peech **E**xtraction	חילוץ דיבור ממוקד	
DNNs	Deep Neural Networks	רשתות עצביות עמוקות	
ANNs		רשתות עצביות מלאכותית	
CNNs	**Convolutional Neural Networks**		
RNNs	**Recurrent Neural Networks**		
LSTM	**Long Short-Term Memory**		
BSS	Blind Source Separation		
ICA	Independent Component Analysis		
IVA	Independent Vector Analysis		
NMF	Non-negative Matrix Factorization		