הפקולטה להנדסה

המעבדה לעיבוד אותות

The Cone of Silence: Speech Separation by Localization

איליה ליפובן

תומר להב

פרויקט שנהד' לקראת תוארראשון בהנדסה

מנחה: מרדכי מוראד י

מנחה אקדמי: פרופ ' שרון גנות

אוקטובר 2024

תוכן עניינים

הצגת הבעיה ומטרת הפרויקט 4
מבוא 4
הגדרת הבעיה: 4
מטרת הפרויקט: 5
השיטה: 5
מימד הזמן מול מימד התדר 6
STFT – short time Fourier transform 6
ה DATA למודל 7
רקע תיאורטי 7
SIR – signal to interference ratio: 7
SNR – signal to noise ratio: 7
RIR – room impulse response: 7
רעש הפרעה 8
מקדם ספיגה 8
יצירת ה DATA 9
כללי 9
הדוברים 9
מערך המיקרופונים 9
יצירת רעש רקע 9
סימלוץ החדרים 9
Data Pipeline: 11
בניית הרשת 12
רקע תיאורטי עבור העקרונות השונים בהם השתמשו 12
RNN – Recurrent neural networks: 12
LSTM – Long short-term memory: 12
שכבות קונבלוציה: 13
Weights Rescaling 14
רשתות ותכונות עליהן מבוססת הרשת שלנו: 15
Wave-U-Net 15
Skip Connections 15
מודל ה- Demucs: 16
Cone of Silence – הרשת המוצעת: 16
ארכיטקטורת הרשת 17
אימון הרשת 18
רקע תיאורטי 18

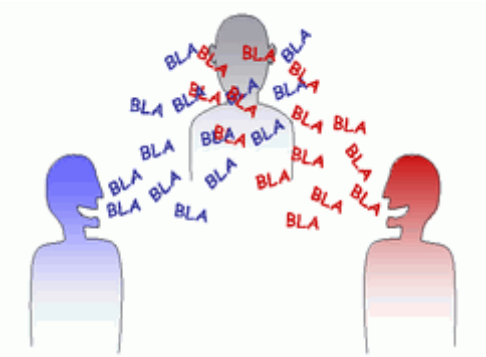
Gradient Descent 18
Adam 18
פונקציית SHIFT 19
האימון של Cone of Silence: 20
תוצאות 22
Train and Val Loss 22
בכחול רואים את האפוקים, באדום את תוצאות הוולידציה שהתבצעה על 1000 דוגמאות בכל אפוק, ובירוק את הTRAIN כנגד מספר הבאצ'ים. 22
אנחנו רואים כי תוצאות הTRAIN והVALIDATION תואמות ואין בעיית OVERFITTING, וגם רואים שקיימת התכנסות. 22
הפרדה מתוך זווית וחלון 23
הפרדת דוברים באמצעות הרשת 24
עקרונות נדרשים 24
SISDR: 24
NMS (Non-Maximum Suppression): 25
אלגוריתם חיפוש בינארי להפרדת דוברים 27
תוצאות מעשיות 28
השוואה למודלים קיימים 29
Ideal ratio mask: 29
Ideal binary mask: 29
Real life + Moving sources: 30
נספחים 31
נספח א' – מעבר הנתונים דרך הבלוקים במודל 31
נספח ב' – ניסויים נוספים של הפרדה דוברים 31
נספח ג' – מיקום דוברים לפי TEST DATA: 35
נספח ד' – תוצאות ROOM GENERATION: 36

הצגת הבעיה ומטרת הפרוייקט

מבוא

התופעה שבה אנשים יכולים להפריד את מיקום הדוברים ואת מה שהם אומרים נקראת בשם " Cocktail party effect". יכולת זו של אנשים מוגבלת, ובאיזורים רועשים שבהם הרבה אנשים מדברים במקביל קשה להפריד ביניהם הדוברים.

בתחום עיבוד האותות הפרדת דוברים היא בעיה מאתגרת, שבה נדרש להפריד דיבור של דוברים שונים מתוך תערובת של קולות. היכולת להפריד בין דוברים חיונית במגוון יישומים מודרניים, כמו עוזרים וירטואליים שצריכים לזהות פקודות קוליות מדוברים שונים, מערכות שמע, וידאו צאט' המסוגל לבודד רעשי רקע, מכשירי שמיעה המסייעים למשתמשים להבחין בין דוברים בסביבה רועשת ועוד.



הגדרת הבעיה:

בהינתן מערך מיקרופונים רב ערוצי, עם מספר דוברים לא ידוע מראש שמדברים בו זמנית. נרצה לדעת למקם את הדוברים ולהפריד בין הקולות של הדוברים השונים, כלומר להפריד לפי מיקום ודובר.

קשה להסיק סטטיסטיקה על הדוברים שהיא אינה ידועה מראש, בפרט על הרעשי רקע שנמצאים בחדר, אילו יכולנו לעשות זאת, על ידי הבנת ההתנהגות הסטטיסטית היינו יכולים להפריד את הדוברים וכן את הרעש.

כמו כן, בעיה זו שבהמשך נראה שנפתרת על ידי רשת עמוקה, נפתרת בדרך כלל על ידי רשת שמשתמשת במישור התדר. עם זאת, לאחרונה מחקרים מראים שרשתות שעובדות במישור הזמן נוטות לעבוד מוצלח גם כן, כגון: Wave-U-Net, Conv-Tasnet, Tasnet. על היתרונות והחסרונות בבחירה לממש רשת במישור הזמן לעומת מישור התדר נפרט בהמשך.

<u>באופן פורמלי</u>:

בהינתן מערך מיקרופונים עם תצורה ידועה מראש המורכב מ- M מיקרופונים כאשר $M > 1$.
בעיית ההפרדה והלוקליזציה של מקורות ב ־ M ערוצים ניתנת לניסוח במונחים של הערכת N מקורות: $s_1, \dots, s_N \in R^{M\times T}$

והמיקומים הזוויתיים שלהם, $\theta_1, \dots, \theta_N$ מתוך גל קול בדיד במערך M ערוצים של התערובת

$x \in R^{M\times T}$ באורך T כאשר: $x = s_i \sum_{i=1}^{N} s_i + bg$

מטרת הפרויקט:

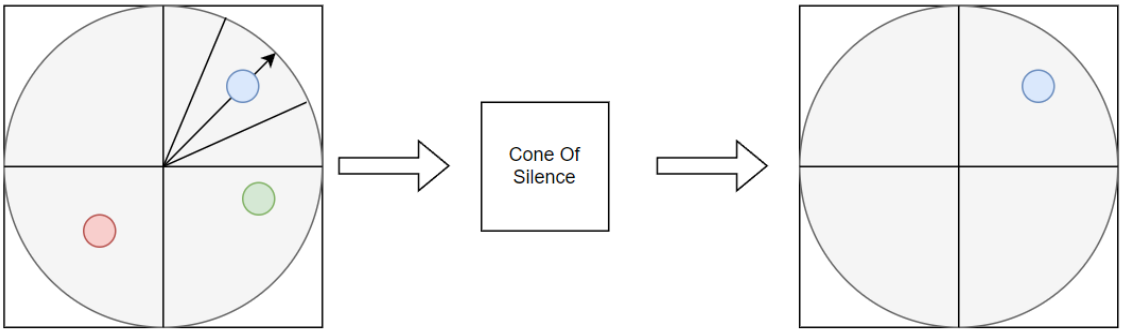
לבנות ולאמן מודל המבוסס על המאמר the cone of silence אשר יכול לפתור את בעיית הפרדת ומיקום הדוברים, מתוך חדר בפרמטרים מוגרלים אקראית ומספר דוברים משתנה במיקומים שונים שמדברים למערך מיקרופונים.

השיטה:

ראשית, נבנה סימולציות של חדרים, המכילים מערך מיקרופונים ודוברים הפזורים ברחבי החדר ומדברים לכיוון המערך. את הדוברים לקחנו מדאטאסט מוכר (VCTK) וסימלצנו אותם במעבר במערך המיקרופונים, על זה הוספנו רעש שכולל 4 דוברים בפינות החדר. שמרנו את כל הסימולציות בנפרד ואת הסופרפוזיציה של כולם : $\sum x_i + s$ יחד עם מטאדאטא שמכיל את המידע על המיקומים בחדר, זאת עבור 15,000 חדרים.

לאחר שיצרנו את כל המידע הנדרש עברנו לשלב האימון בו הגרלנו זווית וחלון באופן יוניפורמי (הסבר על אופן ההגרלה נמצא בחלק האימון), הכנסנו את כלל המיקס יחד עם הזווית, החלון, ותוספת של רעש הפרעה (שונה מרעש הרקע שיצרנו – עליו נסביר בהמשך), והשוונו את המוצא עם הדוברים שבאמת נמצאים בחלון ובזווית המוגדרים.

לאחר שלב האימון קיבלנו מודל שיודע לסנן את הדוברים שנמצאים בחלון ובזווית מסויימים משאר הדוברים ורעשי הרקע הנוספים שיש ברחבי החדר:



לאחר אימון המודל, אנחנו נבצע חיפוש בינארי על גדלי חלון שונים, ובהתאם לאנרגיה שנמצא בכל בדיקה - נדע איפה יש ואיפה אין דוברים בכל אזור ואזור.
לאחר מכן נבצע NMS, אלגוריתם למניעת כפילויות בין הדוברים (עליו נרחיב בהמשך) באמצעות מטריקת SISDR.

מימד הזמן מול מימד התדר

ניתן לנתח אותות בשני מישורים עיקריים: מישור הזמן ומישור התדר. במישור הזמן, האות מוצג לפי ערכי הזמן שלו, בעוד שמישור התדר מציג את ההתפלגות התדרית של האות.

כדי לעבור בין שני המישורים, משתמשים בהתמרת פורייה, שמתארת כל אות כהרכב של תדרים.

אולם כאשר מדובר באותות כמו דיבור, שהתדרים שלהם משתנים לאורך הזמן, נדרש כלי ניתוח נוסף.

STFT – short time Fourier transform

STFT או "התמרת פוריה קצרת זמן" - הוא כלי ניתוח מרכזי בעיבוד אותות משתנים בזמן, כמו דיבור. באמצעות STFT, מחלקים את האות לחלונות זמן קצרים, עליהם מבצעים התמרת פורייה נפרדת, וכך מתקבל ייצוג של התדרים לאורך ציר הזמן. בצורה זו ניתן לעקוב אחרי השינויים התדריים של האות בזמן אמת.

רבות מהשיטות הקיימות להפרדת הדוברים המבוססות על הפרדה במימד התדר באמצעות STFT על ידי הפרדה מבוססת מסכה להערכת המקורות:

שיטות אלו מיישמות מסכה (σ_i) על ספקטרוגרמת התערובת (S) כדי להעריך כל מקור, ולאחר מכן ה-ISTFT משחזרת את אותות המקור. המסכות יכולות להיות בינאריות ({1, 0}) או רציפות (,1]) (SOFT 0]).

קיימות מספר מגבלות בשיטה זו:

ראשית, אין ערובה לכך שהכפלת המסכה ב-S תייצג ספקטרוגרמה אמיתית. התהליך עשוי להוביל להוספת ארטיפקטים בשלב ה-ISTFT, מכיוון שההשלכה הנדרשת אינה מובאת בחשבון בפונקציית ההפסד במהלך האימון.

שנית, בעיית הפאזה נותרת בעינה. שיטות אלו אינן ממדלות את הפאזה, מה שעלול להוביל לאי דיוקים. לדוגמה, אם קולו של זמר מכיל ויברטו, אך הגיטרה מנגנת באותו גובה הצליל, שימוש בפאזה המקורית של התערובת יחיל את הויברטו באופן שגוי על הגיטרה.

גישה מבוססת גל

במקום להשתמש בספקטרוגרמות (שהן ייצוג תדרי של האות), ניתן ללמד מודלים ישירות על גלי הקול עצמם (waveform). גישה זו יכולה להתגבר על כמה מהמגבלות הקיימות בשיטות מבוססות ספקטרוגרמות. כאשר מודל לומד ישירות מהגל, תהליך האימון הוא מקצה לקצה – כלומר, אין צורך בשלב נוסף של סינתזה לשחזור האות, שלב שיכול להוסיף בעיות או עיוותים לאות הסופי. זה פותר את הבעיה הראשונה שתיארנו קודם.

לגבי הבעיה השנייה (שקשורה לפאזה), המצב עדיין מורכב. לא בטוח אם יש מודל שיכול להתמודד היטב עם מקרים מיוחדים, כמו כאשר יש דוברים שונים שמשמיעים צלילים דומים מאוד מבחינת התדרים שלהם. בתחומים כמו יצירת דיבור או מוזיקה, גישה ישירה זו של יצירת גל כבר החליפה את השיטות הישנות שהתבססו על ספקטרוגרמות.

ה DATA למודל

רקע תיאורטי

SIR – signal to interference ratio:

מדד שמייצג את היחס בין עוצמת האות הרצוי לבין עוצמת אותות מפריעים אחרים, שהם אותות תקשורתיים או אלקטרוניים המגיעים ממקורות חיצוניים אחרים. בניגוד ל SNR, ההפרעה אינה רעש רנדומלי, אלא אותות אחרים שיכולים להפריע לאות הרצוי.

$$SIR = 10 \cdot \log_{10}\left(\frac{P_{signal}}{P_{interference}}\right)$$

SIR גבוה מצגיע על כך שעוצמת האות הרצוי גבוהה בהרבה מההפרעות ולכן האות ברור יותר וקל לקליטה.

SNR – signal to noise ratio:

מודד את היחס בין עוצמת האות הרצוי לבין עוצמת הרעש הרקעי הלא רצוי (רעש סטטיסטי או אקראי). רעש יכול להיות "לבן" או כל צורה אחרת של רעש רנדומלי שנוכח במערכת.

$$SIR = 10 \cdot \log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)$$

SNR גבוה מצגיע על כך שעוצמת האות הרצוי חזקה ביחס לרעש ולכן האות ברור יותר.

RIR – room impulse response:

RIR הוא תיאור של איך צליל מתפשט בתוך חלל (כמו חדר) מנקודת מקור עד נקודת מדידה. הוא מתאר את התגובה של החדר או הסביבה האקוסטית להתרחשות של דחף קול קצר, כמו תנועה של מקורות קוליים. התגובה הזו כוללת את כל התופעות האקוסטיות שמתרחשות בחדר, כמו השתקפויות מהקירות, החזרים, והדהודים.

RIR הוא פונקציה שמתארת את השינוי בעוצמת הצליל ובתדירות במהלך הזמן כאשר דחף קול משוגר בחדר. הוא בעצם מייצג את ה"חותם" של החדר על הצליל, כלומר איך החלל משפיע על הפצת הקול מהמקור לנקודת המדידה.

באמצעות פעולת הקונבולוציה עם קולות הדוברים והRIR אנחנו יכולים לסמלץ את הדיבור של הדוברים בחדר:

$$RIR * X_i = \overline{X}_i$$

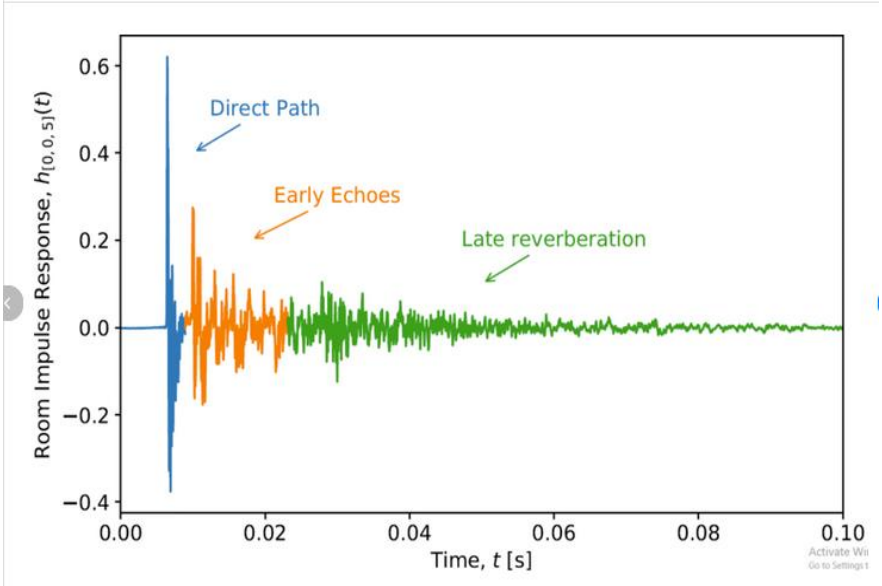
כאשר $\overline{X}_i$ זה אות הדיבור כפי שנקלט במיקרופון בחדר המסומלץ.

RIR מיוצג בתור סיגנל בזמן, והוא מורכב משלושה חלקים עיקריים:

1. **Direct sound:** החלק הישיר החלק הראשון שמגיע לנקודת המדידה, והוא השידור הישיר מהמקור לנקודת המדידה ללא כל השתקפות.
2. **Early Reflections:** השתקפויות מוקדמות קטעי קול שמגיעים אחרי ההתנגשות הראשונה שלהם בקירות או במשטחים אחרים בחדר.
3. **Reverberation:** הדהוד הפולסים הקוליים שמגיעים אחרי הרבה השתקפויות בחדר, והם ממשיכים לדעוך לאט עם הזמן.

מתוצאות ה RIR אנחנו יכולים ללמוד מספר דברים:

- אקוסטיקה של החדר: RIR מכיל מידע חשוב על זמן ההדהוד של החדר, כלומר כמה זמן לוקח לצליל לדעוך לרמה מסוימת לאחר הפסקת השידור.
- מיקום של מקורות ומקלטים: השוואה בין RIR מנקודות שונות בחלל יכולה לתת אינדיקציה על מיקום מקור הצליל או המיקרופון בתוך החדר.
- תדירויות דומיננטיות: ניתן להסיק גם על תדירויות מסוימות שעוברות הגברה או הנחתה בתוך החדר.



רעש הפרעה

רעש הפרעה גאוסי (מכונה לעיתים גם "רעש לבן") הוא סוג של רעש אקראי שמתואר על ידי התפלגות נורמלית (גאוסית). רעש גאוסי נפוץ בתחומים כמו עיבוד אותות ותקשורת, ומשמש לדמות הפרעות טבעיות שמתרחשות במערכות רבות, כמו תנודות אקראיות במתח חשמלי, שגיאות מדידה, ורעשי חיישנים.

אנחנו משמשים ברעש הפרעה על כל האותות שעוברים במודל על מנת שהמודל ידע להתמודד עם רעשים גאוסיים בנוסף לרעש הדיפוזי של הדוברים.

מקדם ספיגה

מקדם ספיגה – (Absorption Coefficient) זהו פרמטר המתאר את מידת האנרגיה של הגל שהחומר "סופג" כשהוא עובר דרכו. ערך גבוה של מקדם ספיגה אומר שהחומר סופג יותר מהאנרגיה של הגל, מה שגורם לדעיכתו ולהפחתת העוצמה שלו. במילים אחרות, חומר עם מקדם ספיגה גבוה "בולע" יותר אנרגיה מהגל המגיע.

יצירת ה DATA

כללי

אנחנו יוצרים חדרים בהם יש מערך מיקופונים, דוברים ורעש רקע כך שבסופו של דבר בהינתן חדר עם אנשים מפוזרים סביבו נדע לבודד קול של דובר מסויים מחלון מסויים ברחבי החדר ולאחר מכן גם להגיד איפה ממוקדים הדוברים בחדר.

בשביל להגיע למצב הזה אנחנו צריכים ליצור כמות גדולה של דוגמאות שונות של חדרים בגדלים שונים עם מס אנשים שונים שאומרים דברים שונים ואת כל הדוגמאות הללו נכניס למודל שנבנה.

M – מיקרופון.
S – אות דיבור של דובר.
Mix – עירבוב כל האותות של כל הדוברים בחדר שנקלט על ידי מיקרופון מסויים.

הדוברים

השתמשנו ב **VCTK DATASET** – מאגר נתונים הכולל נתוני דיבור של 110 דוברים שונים (באנגלית, עם מבטאים שונים), כאשר כל דובר מקריא כ 400 משפטים שונים.
הדוברים מדברים ב MONO – כלומר ערוץ דיבור יחיד (כל הקול מוקלט ונשמע כאילו מגיע ממקור אחד בלבד ללא הבחנה בין ימין לשמאל), ובתדר דגימה ש$[Hz]$ל 44,100 HZ, ככל שתדר הדגימה של הדאטה גבוה יותר זה עוזר לנו באימון במישור הזמן (ההסבר לכך מפורט בהמשך בחלק אימון הרשת), לכן בחרנו את VCTK לעומת DATASETS אחרים כמו LIBRISPEECH למודל שלנו.

מערך המיקרופונים

יצרנו מערך מיקרופנים הנמצא במרכז החדר, מכיל 6 מיקרופונים ברדיוס של 0.0463[m] המייצגים 6 ערוצים – רב ערוצי (Multichannel).

יצירת רעש רקע

הגרלנו 4 דוברים (שהם בהכרח לא אותם דוברים שנבחרו כחלק מהדוברים בחדר הספציפי), ומיקמנו אותם בפינות החדר בשביל לייצר אפקט של רעש דיפוזי.

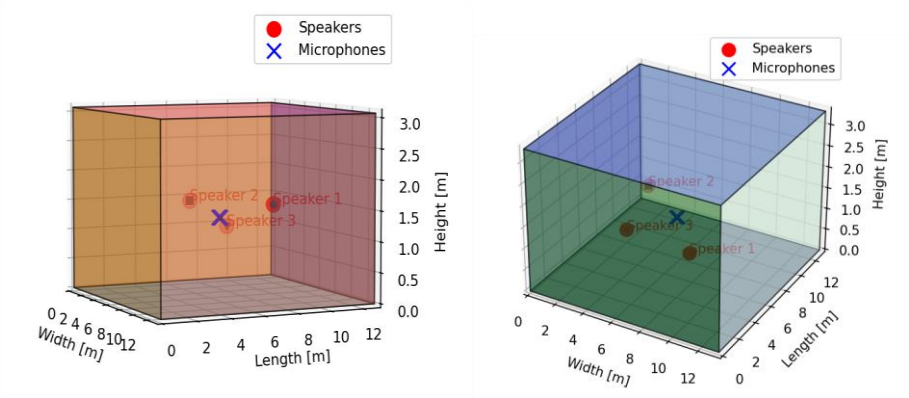
סימלוץ החדרים

הגודל של כל חדר מוגרל בין 12 ל20 מטר אורך ורוחב ובין 3 ל4 מטר גובה.

סימולציות בחדרים גדולים מציעות מספר יתרונות:
רמת הדהוד - חדרים גדולים מספקים סביבות הדהוד גדולות יותר שמתאימות למגוון רחב יותר של תרחישים.
פיזור מקורות הצליל - בחדר גדול ,אפשר לשים את הדוברים במיקומים מגוונים יותר מבלי ליצור קרבה גדולה בין הדוברים או לקירות . מה שיוצר סביבה פחות צפופה ,שבה ניתן להתייחס בצורה מדויקת יותר לתכונות הדיבור של כל דובר ולהתמודד עם רעשי רקע מורכבים יותר.

עבור כל סימולציה בחרנו בצורה אקראית בין 1 ל4 דוברים כשהמיקומים שלהם בחדר גם כן מוגרלים לפי $X \in [12,20], Y \in [12,20], Z \in [1.4,1.5]\ [m]$, המרחק בין דובר לדובר מוגדר להיות לפחות $0.4\ [m]$ אחד מהשני.

מקדם ספיגה הוגרל בין 0.4 ל 0.99. מקדם חזרתיות (מס' הפעמים שהקול חוזר אחרי שהוא פוגע בקיר) – 10 (מקסימלי).
כלומר הקול של כל אחד מהדוברים מוחזר בחדר בעוצמה מוכפלת במקדם הספיגה 10 פעמים כל פעם.



בכל הקלטה שלקחנו, התחשבנו בהפרדה בין הקלטות שמיועדות לשלב ה TRAIN ובין שלב ה TEST.

ביצענו נרמול של SIR על ידי בחירה של דובר אחד ונרמול כך שה SIR בין כל הדוברים יהיה בין 0 ל 5 [dB] ביחס לדובר זה.

ביצענו נרמול של עוצמת הקול של סה"כ הדוברים. למקדם הנרמול קראנו בטא : $\beta = \frac{1}{max(|mix|)}$.

באמצעות הספרייה של PYROOMACOUSTICS חישבנו את הRIR שהוא תגובת התדר של החדר וביצענו קונבולוציה עם אותות הדיבור.

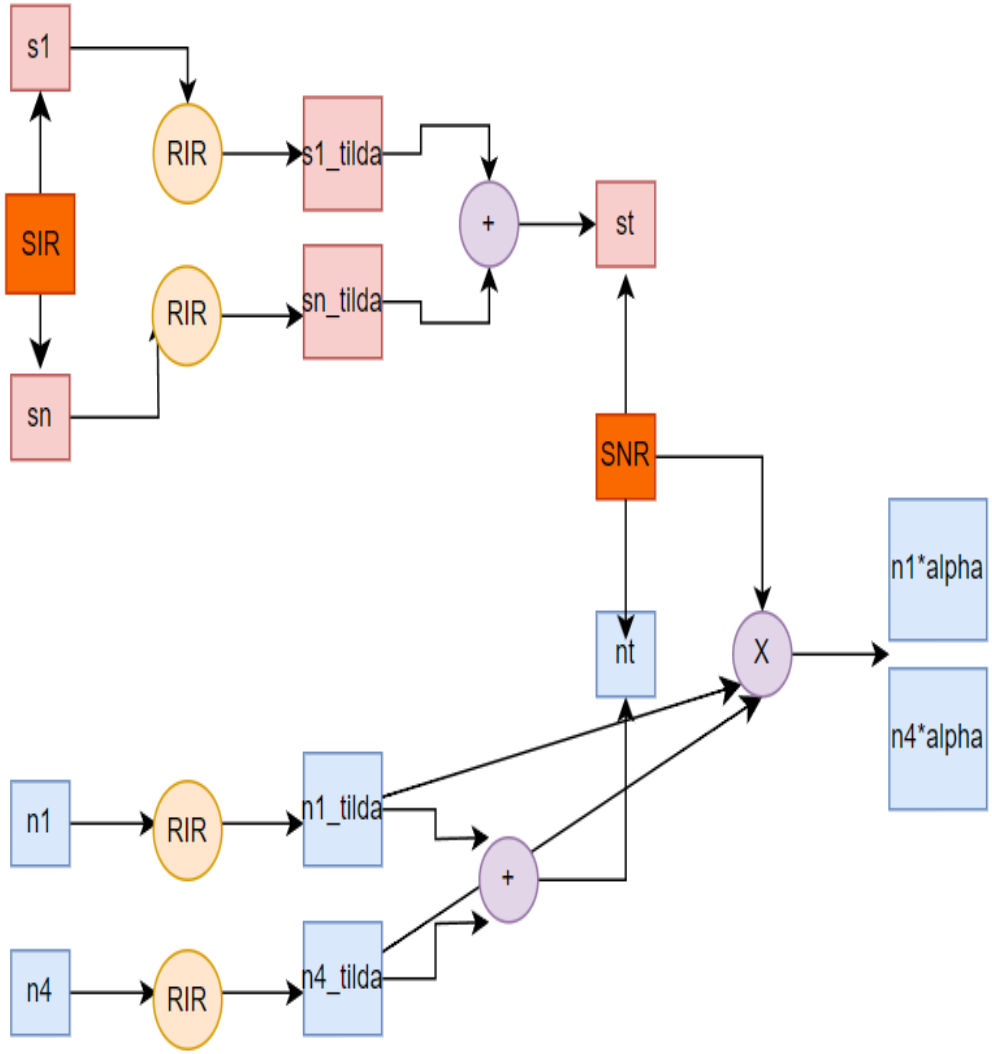
ביצענו קונבלוציה עם ה RIR ועם אותות הדוברים המנורמלים $s_1 \dots s_n$ וקיבלנו את האותות $\widetilde{s_1} \dots \widetilde{s_n}$ חיבור של $s_t = \widetilde{s_1} + \cdots \widetilde{s_n}$ נתן לנו את המיקס של כלל הדוברים המדברים בחדר יחד. זאת עבור כל אחד מ6 הערוצים, כאשר מספר הערוץ נע בין 1 ל6.

סימלצנו את ה Background בנפרד ואת הדוברים ב Foreground. ווידאנו שה SNR בין המיקסים של כל אחד יהיה בין 8 ל10. למקדם הSNR קראנו אלפא:$\alpha = snr(mix\ Foreground, mix\ Background)$.

את ארבעת רעשי הרקע שמרנו בתור $n_1 \dots n_4$ ובדומה לאותות הדיבור העברנו את הרעש ב RIR וקיבלנו את $\widetilde{n_1} \dots \widetilde{n_4}$.

לבסוף שמרנו את $\widetilde{s_1} \dots \widetilde{s_n}$ ואת $\boldsymbol{mix = s_t + \widetilde{n_1} \cdot \alpha + \cdots + \widetilde{n_4} \cdot \alpha}$ יחד עם המטאדאטה המכיל את המיקומים של כולם.

:Data Pipeline



בניית הרשת

רקע תיאורטי עבור העקרונות השונים בהם השתמשו

RNN – Recurrent neural networks:

רשתות נוירונים חוזרות (RNNs) נועדו להתמודד עם נתונים רציפים כמו סדרות זמן, אודיו או טקסט על ידי שמירה על זיכרון מסוים של נתונים קודמים. היתרון המרכזי שלהן הוא היכולת לעבד נתונים בסדרות, שבהן הסדר חשוב, על ידי שימוש במצב מוסתר שמתעדכן לאורך הזמן.

h_t - המצב המוסתר בזמן t

$W_{\{xh\}}$ - מטריצב המשקלים שמכפילה את הקלט x בזמן t

$W_{\{hh\}}$ - מטריצה שמייצגת את הקשרים החוזרים

h_t - המצב המוסתר

b_h - הטייה BIAS

σ - פונקציית אקטיבציה (RELU או TANH)

סה"כ:

$$h_t = \sigma(W_{\{xh\}} * x_t + W_{\{hh\}} * h_{t-1} + b_h)$$

המשוואה המתוארת מייצגת את החישוב של הייצוג הנסתר בזמן t ברשת.
המשוואה מחושבת עבור כל שלב ברצף הזמן ומבוססת על האינפוט הנוכחי והמצב הנסתר הקודם, כך שהרשת יכולה לשמור על מידע מרצף הנתונים.

LSTM – Long short-term memory:

LSTM הוא סוג מתקדם יותר של RNN, שתוכנן במטרה לשמור טוב יותר על זיכרון לטווח ארוך.

ב-LSTM יש מבנה ייחודי עם שלושה שערים מרכזיים:

1. שער כניסה (Input Gate): קובע איזה מידע מהקלט הנוכחי ייכנס למצב הזיכרון.
2. שער שכחה (Forget Gate): קובע איזה חלק מהמידע הישן במצב הזיכרון יש לשכוח.
3. שער יציאה (Output Gate): קובע מה המידע שיוצא מהתא לזמן הנוכחי ולמצב המוסתר הבא.

המבנה הזה מאפשר ל-LSTM לשלוט בצורה מדויקת יותר במידע שהוא שומר, זוכר ושוכח, דבר שמאפשר למודל להתמודד עם תלות ארוכת טווח בצורה טובה יותר. LSTM נועד להתמודד עם בעיית ה- vanishing gradient, שהיא בעיה מרכזית ברשתות - RNNs בסיסיות, אשר מקשה על למידה אפקטיבית של קשרים ארוכי טווח.

שכבות קונבולוציה:

שכבות קונבולוציה **(Convolutional Layers)** הן מרכיב מרכזי ברשתות קונבולוציוניות **(CNNs,** שנפוצות בעיבוד תמונה, קול, ומשימות לזיהוי תבניות. שכבה קונבולוציונית מבצעת קונבולוציה בין הקלט (למשל תמונה) לבין פילטרים **(Kernel)** כדי לזהות תבניות מקומיות, כמו קצוות או טקסטורות, וליצור מפות תכונות **(feature maps)**.

בכל שכבה מוגדרים פילטרים קטנים שמחליקים על הקלט בצעדים **(Strides)** קבועים, ומבצעים מכפלת מטריצה עם חלקי הקלט. התוצאה היא ערכים המייצגים אזורים בהם נמצאה תבנית דומה לפילטר. שכבת אקטיבציה) כמו **(ReLU** יכולה להוסיף אי-לינאריות אחרי הקונבולוציה**.**

היתרונות של שכבות קונבלוציה:

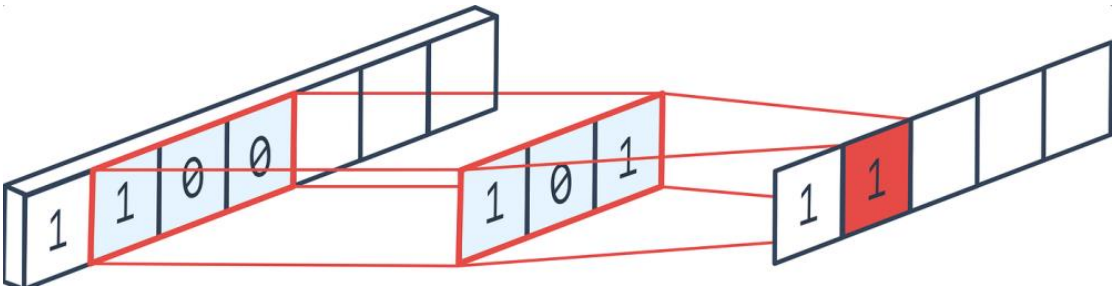
- שמירה על מידע מרחבי.
- הפחתת פרמטרים בזכות שיתוף משקלות.
- זיהוי תבניות בקנה מידה שונה.

החישוב המתמטי של קונבולוציה מבוסס על מכפלת הפילטר במקטעים מקומיים של הקלט, כשהתוצאה היא מפת תכונות חדשה**.**

$$y(t) = (x * h) \cdot (t) = \sum_{n=0}^{N-1} x(t-n) \cdot h(n)$$

כאשר:

- $y(t)$ הוא הפלט של הקונבלוציה בנקודה **t**.
- $x(t-n)$ הוא האות המקורי במיקום **n-t** כאשר **n** הוא ההזזה של המסנן ביחס לאות.
- $h(n)$ הוא המסנן – גודל קבוע שמכפיל את האות **x**.
- N גודל המסנן.



Weights Rescaling

במודל שלנו השכבות השונות מקבלות קנה מידה שונה בסטיית התקן במשקלים לפי עומק השכבות במהלך האימון. כדי למנוע אימון עם קצב למידה אחיד לעומק שכבות שונה או להתכנסות איטית עם קצב למידה אחיד אנחנו עושים מניפולציה למשקלים לפי סטיית התקן של כל שכבה.

שינוי קנה מידה של משקלים בשכבות קונבולוציוניות במודל הוא קריטי בגלל המורכבות של הנתונים (אודיו) והעובדה ששכבות שונות עוסקות בתכונות שונות – למשל, השכבות הראשונות עוסקות במידע גלובלי יותר (טווח רחב של תדרים), והשכבות האחרונות מתמקדות במאפיינים ספציפיים. שינוי קנה מידה של המשקלים מאפשר למודל ללמוד בצורה מאוזנת יותר ולהימנע ממצבים שבהם השכבות האחרונות לא מתעדכנות בצורה מספקת או מתעדכנות יתר על המידה.

השימוש בטכניקה הזו במודל מאפשר גמישות רבה יותר במהלך האימון, במיוחד כשמדובר במודלים שדורשים איזון בין שכבות רבות עם תכונות שונות.

שיטות אופטימיזציה מודרניות כמו Adam מתאימות את שיעור הלמידה עבור כל פרמטר בנפרד, כך שבממוצע כל משקל יקבל עדכונים באותו סדר גודל. אך אם נרצה להשתמש בשיעור למידה גדול כדי לכוון את המשקלים בשכבה הראשונה, זה יהיה גדול מדי לשכבה האחרונה.

פתרון: שינוי קנה מידה של המשקלים: המחברים מציעים "טריק" שמאפשר להשתמש בשיעור למידה מותאם לכל שכבה, בלי לשנות את המודל באופן מהותי:

- נסמן את המשקלים כ- W, ונחשב את

$$\frac{std(w)}{a} = \alpha$$

כאשר a הוא קנה מידה ייחוס (במקרה הזה, $a = 0.1$)

- נחליף את המשקל W ב-

$$\frac{w}{\sqrt{\alpha}} = w'$$

ואת תוצאת הקונבולוציה ב-

$$x * w' \cdot \sqrt{\alpha}$$

כך שהתוצאה תישאר זהה, אך המשקלים יהיו מותאמים לפי כל שכבה.

טכניקה זו תסייע לזרז את ירידת פונקציית האבידה במהלך האימון ולהגיע לאופטימום טוב יותר.

רשתות ותכונות עליהן מבוססת הרשת שלנו:

Wave-U-Net

מודל המשמש בעיקר להפרדת מקורות (source separation) באותות אודיו, כמו הפרדת כלי נגינה מתוך הקלטה או הפרדת דיבור מרעשי רקע. המודל מבוסס על הארכיטקטורה של U-Net, שנמצא בשימוש רחב בתחום הראייה הממוחשבת, אך הותאם לעיבוד ישיר של אותות אודיו גולמיים, כלומר במישור הזמן ולא במישור התדר כמו שציינו קודם

Wave-U-Net בנוי ממספר שכבות של Downsampling (הקטנת ממדים) ו-Upsampling (הגדלת ממדים) באופן סימטרי. כל שכבה ב -downsampling מקטינה את הרזולוציה של האות תוך חישוב מאפיינים (features) מתקדמים יותר, בעוד שב -upsampling הרשת מחזירה את האות לרזולוציה המקורית, תוך שימוש במאפיינים שנלמדו בשלבים הקודמים. במהלך התהליך הזה, המודל לומד לייצג את האות בצורות מופשטות ומפורטות יותר, מה שמאפשר לו להפריד את המקורות בצורה יעילה.

היתרון המרכזי של Wave-U-Net הוא בכך שהוא עובד ישירות על האות הגולמי (waveform), בניגוד למודלים כמו STFT או spectrogram, ולכן הוא לא דורש שום הנחה מוקדמת על המבנה התדרי של האות. בנוסף, השימוש באות במישור הזמן מקנה יתרון בכך שהמודל יכול לשמור על פרטים מרחביים (spatial details) יותר בקלות.

Skip Connections

Skip Connections היא טכניקה ברשתות נוירונים המאפשרת העברת מידע ישירות משכבה אחת לשכבה עמוקה יותר, תוך דילוג על שכבות ביניים. כלומר, במקום שהמידע יעבור רק דרך כל השכבות ברצף, הוא יכול "לדלק" שכבות מסוימות ולהעביר מידע בצורה ישירה לשכבות עמוקות יותר ברשת.

הטכניקה הזו פופולרית במיוחד בארכיטקטורות כמו U-Net ו-ResNet חיבורים אלו מאפשרים:

1. שמירה על פרטים חשובים: במודלים כמו U-Net, החיבור המדלג מאפשר לשמור על פרטים שהתקבלו בשכבות מוקדמות (ב-downsampling) ולהעביר אותם ישירות לשכבות של upsampling. כך נשמר מידע חשוב שלא אובד בתהליך של הקטנת ממדים.
2. התגברות על בעיות של היעלמות גרדיאנט: ברשתות עמוקות, skip connections מסייעים להקטין את בעיית היעלמות הגרדיאנט, שבה עדכוני המשקל מתעממים ככל שהרשת מעמיקה. באמצעות Skip Connections, הגרדיאנט מועבר ישירות לשכבות עמוקות יותר, מה שמקל על תהליך האימון.

ב-Wave-U-Net, skip connections מאפשרים לשלב בין המידע שנלמד בשכבות הנמוכות (downsampling) לבין המידע מהשכבות הגבוהות (upsampling), כך שהמודל משלב ידע מרמות שונות של עיבוד.

מודל ה- Demucs:

Demucs (deep music separation) היא ארכיטקטורת למידת עמוקה שנועדה לבצע הפרדת מקורות באותות שמע, במיוחד בתחום המוזיקה. המטרה המרכזית של Demucs היא לבודד רכיבי מוזיקה בודדים (כגון שירה, תופים, בס וכלים נוספים) מתוך אות שמע מעורב. יכולת זו חשובה במיוחד בהפקת מוזיקה, רמיקסים, וניתוח מוזיקלי.

שכבות המודל

המבנה של Demucs כולל מספר שכבות חשובות שמסייעות בהפרדת המקורות:

1. שכבות קונבולוציה:
   - השכבות הקונבולוציוניות אחראיות על חציבת תכנים מהספקטרוגרם של האודיו, ומסייעות לזהות דפוסים ומבנים באות השמע. השכבות הללו מזהות תבניות כמו כלי נגינה ושירה.
2. שכבות חזרתיות:
   - כדי ללמוד את התלות הזמנית באות השמע, Demucs משלבת שכבות חזרתיות, כמו LSTMs או GRUs השכבות הללו מאפשרות לרשת להבין כיצד רכיבים שונים מתפתחים עם הזמן, מה שמסייע לשפר את איכות ההפרדה.
3. שכבות פלט:
   - הפלט של המודל כולל מספר ערוצים, כל אחד מהם מייצג מקור שמע שונה. האותות המופרדים יכולים להיות משוחזרים בחזרה לדומיין הזמן באמצעות טכניקות כמו ההמרה ההפוכה של טרנספורמת פורייה הקצרה ISTFT.

Cone of Silence – הרשת שלנו:

כפי שנאמר הרשת לוקחת קלט Multichannel, במקרה שלנו 6, במימד: $x_{mix}' \in R^{6X44100*3}$

כך שהקלט מועבר כבר בפונקצייה של SHIFT שהיא פונקצייה שנותנת התאמה חח"ע לזווית המטרה שלנו - w

ווקטור שמייצג את גודל החלון בתצורת One-Hot: h^{1X5}

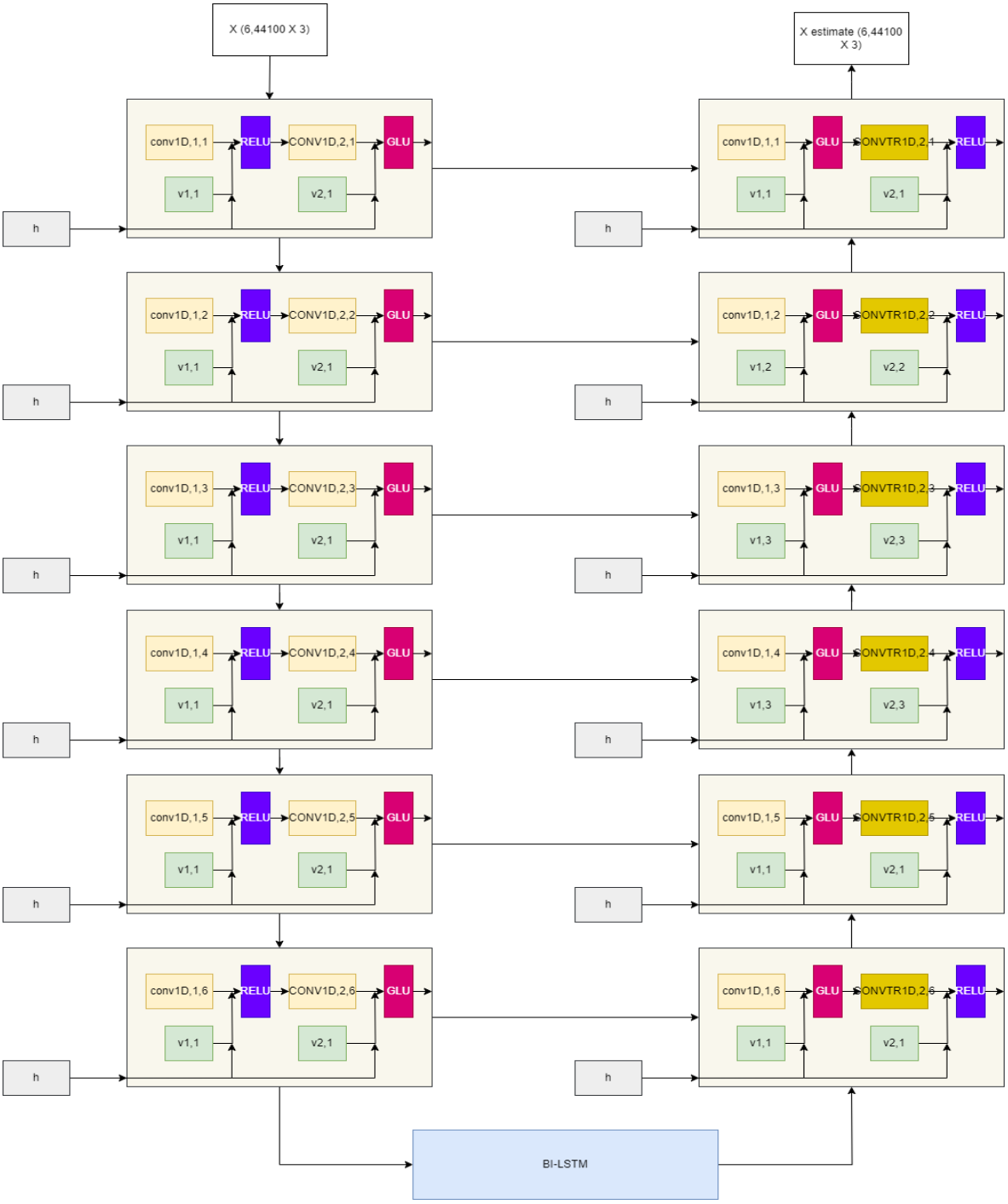
הרשת מורכבת מארכיטקטורת DECODER וENCODER עם SKIP CONNECTION במוטיבציה מרשת DEMUCS, היכולת של הרשת להפריד במימד הזמן והיכולת להשתמש בטעינה של חלק מהשכבות ממודל מוכן (PRETRAIN) מהווה מוטיבציה להשתמש במודל שדומה לה.

הרעיון הוא להמיר את הקלט x_{mix}' באמצעות שכבות ENCODER שמורכב משכבות קונבולוציה לווקטור CONTEXT שמכיל מידע על הקלט במימד נמוך יותר, להעביר את ווקטור הCONTEXT דרך LSTM דו כיווני כדי שהמודל יבין את ההקשר בנקודות הזמן השונות של הקלט , ומכיוון שהLSTM הוא דו כיווני אז נקודות בעתיד "יודעות" את ההקשר לעבר ונקודות בעבר "יודעות" את ההקשר בעתיד.

לאחר מכן להעביר בחזרה דרך DECODER שמורכב מקונבולוציות טרנספוז. במשך כל המודל אנחנו נותנים משתנה גלובלי h שהוא קידוד בתצורת ONE HOT של גודל החלון:
(90,45,22.5,12.25,6.125), עבור זווית של 45 מעלות יומר ל:(0,1,0,0,0), הווקטור מועבר באופן גלובלי לכל השכבות מקודד מפענח.

הסבר על מעבר הקלט בשכבות המודל וגודלו מצורף בנספח.

ארכיטקטורת הרשת



*המקומות בהם החצים נפגשים הם חיבור ווקטורי.

הבלוקים בצד שמאל הם encoders ובצד ימין הם decoders כפי שהסברנו מעל.

אימון הרשת

רקע תיאורטי

Gradient Descent

Gradient Descent הוא אחד מהאלגוריתמים הנפוצים ביותר באימון מודלים של למידת מכונה. מטרתו היא למצוא את הערכים המיטביים של פרמטרי המודל כדי למזער את פונקציית העלות- loss function האלגוריתם פועל על ידי עדכון המשקלים בצורה מחזורית, כך שהערכים שלהם זזים בכיוון השיפוע השלילי של פונקציית העלות ביחס לכל פרמטר – כלומר, בכיוון שבו הירידה בפונקציית העלות היא החזקה ביותר.

Adam

Adam הוא אחד מאלגוריתמי האופטימיזציה הפופולריים ביותר שמשתמשים בהם לאימון רשתות נוירונים. הוא שדרוג של Gradient Descent קלאסי, המשלב בין היתרונות של שני אלגוריתמים Momentum ו-RMSProp.

המטרה של Adam היא לא רק לזוז בכיוון השיפוע, אלא גם לעשות זאת בצורה חכמה יותר, כך שהוא מתאים את גודל הצעד (learning rate) לכל פרמטר בנפרד, על סמך השיפוע וההיסטוריה שלו.

חישוב גרדיאנט: תחילה, בדיוק כמו בגרדיאנט דיסנט רגיל, מחושב הגרדיאנט של פונקציית האיבוד ביחס לכל פרמטר במודל. נניח שהגרדיאנט בזמן t מסומן ב- g_t

ממוצע ראשון (Momentum - תנע): האלגוריתם מחשב את ממוצע הגרדיאנטים מהעבר כדי לקבל כיוון יציב יותר לעדכון. ממוצע ראשון (תנע) מסומן כך:

$$tg \cdot (\beta_1 - 1) + m_{t-1} \cdot \beta_1 = m_t$$

כאשר β_1 הוא היפרפרמטר שמייצג את המשקל שניתן לגרדיאנטים מהעבר (ערך בו השתמשנו: 0.9).

ממוצע שני :(RMSprop) אדם משתמש במידע על שינויים בגרדיאנטים כדי להתאים את קצב הלמידה בהתאם. מחושב ממוצע משוקלל של ריבועי הגרדיאנטים:

$$v_t = g_t^2 \cdot (\beta_2 - 1) + \beta_2 \cdot v_{t-1}\beta_2$$

הוא היפרפרמטר שמייצג את ההשפעה של גרדיאנטים קודמים על הערך הנוכחי (ערך בו השתמשנו: 0.999).

תיקון ההטיות (Bias Correction): מכיוון שבשלבים המוקדמים הגרדיאנטים יכולים להיות מוטים כלפי ערכים קטנים מדי, האלגוריתם מתקן את ההטיות של m_t ושל v_t

$$m_{t^\wedge} = \frac{m_t}{\beta\frac{t}{1} - 1}, v_t = \frac{v_{t^\wedge}}{\beta\frac{t}{2} - 1}$$

עדכון המשקולות: לבסוף, האלגוריתם מעדכן את הפרמטרים לפי הממוצעים שתוקנו:

$$\theta_t = \frac{m_{t^\wedge}}{\epsilon + \sqrt{v_{t^\wedge}}} \cdot \alpha - \theta_{t-1}$$

כאשר α הוא קצב הלמידה, ו- ϵ הוא ערך קטן שמונע חלוקה באפס (בחרנו ב 10^{-8}).

פונקציית SHIFT

פונקציית ה-shift משמשת לשלב הראשוני בהפרדת הדוברים. הפונקציה נועדה ליישר (align) את האותות מכל מיקרופון כך שאותות שמגיעים מזווית מסוימת יהיו מיושרים בזמן בכל ערוצי הקלט, מה שמאפשר שמאפשר לנו להטמיע את זווית המטרה בתוך ערוצי הדיבור במקום להכניס לרשת את הזווית כקלט נוסף.

רעיון

מטרת ה-shift

כאשר אנחנו עוסקים במערך מיקרופונים, האות המגיע לכל מיקרופון משתנה בהתאם לזמן הגעת האות, שהוא פונקציה של מיקום המקור וזווית ההגעה.
הפונקציה shift נותנת לנו את האפשרות להשתמש בזווית המטרה שעליה אנחנו מסתכלים - בתוך הדאטא עצמו במקום להעביר אותה כפרמטר למודל.
נעשה זאת על ידי חישוב הפרשי זמני ההגעה TDOA - Time Difference of Arrival בין המיקרופונים למקור מרוחק שמייצג את זווית המטרה, והזזת הערוצים ביחס להפרשי הזמן כמפורט מטה. בצורה כזו, הזווית מיוצגת כהזזות בערוצים ואנחנו טומנים אותה בהם.

חישוב הפרשי זמני ההגעה:

נסמן:

- c הוא מהירות הקול באוויר (כ 343 מטרים לשנייה).
- sr הוא קצב הדגימה (sampling rate).
- p_θ מקור מרוחק שמייצג את מיקום זווית המטרה – קבענו רדיוס 3 מטר.
- $d(p_\theta, mic_i)$ הוא המרחק בין p_θ למיקרופון ה- i.

בהינתן מקור שנמצא במיקום p_θ והמרחק של מקור האות מהמיקרופון mic_i, הפרש זמני ההגעה T_{dealy} עבור כל מיקרופון מחושב לפי המשוואה:

$$T_{delay}(p_\theta, mic_i) = \lfloor \frac{d(p_\theta, mic_i)}{c} \cdot sr \rfloor$$

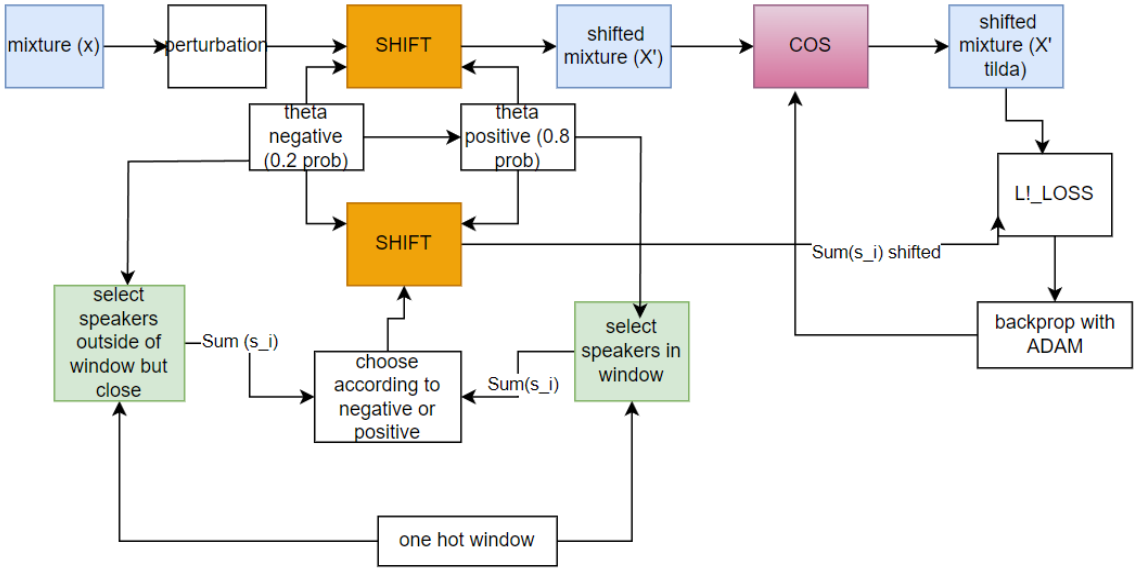
לאחר חישוב ה-TDOA, הפונקציה מיישמת הזזה (shift) בזמן על האות מכל מיקרופון כך שכל ערוצי הקלט מיושרים בהתאם לזווית המטרה θ.
לדוגמה, אם מיקרופון 0 נבחר כעמדת ההתייחסות, כל שאר ערוצי הקלט x_i יוזזו כך שיהיו מיושרים ל mic_0:

$$x_i' = shift\left(x_i, T_{delay}(p_\theta, mic_0) - T_{delay}(p_\theta, mic_i)\right) \; i \in \{1..M-1\}$$

ניתן לראות שככל שקצב הדגימה sr יותר גבוה כך T_{delay} יהיה גדול יותר, לכן ה shift בכל ערוץ יקבל יותר משמעות בכך שהוא גדל יותר ביחד ל sr.

הפעולה של הפונקציה היא ריפוד האותות באפסים.

האימון של Cone of Silence:



אנחנו מגרילים :

חלון - w: ווקטור One Hot Encoded של אחת מהזוויות 90,45,22.5,12.25,6.125.

זווית – θ: בהתאם לגודל החלון שהגרלנו ניקח אחת מהזוויות במרחב באופן יוניפורמי.

שלילי או חיובי: דוגמא חיונית שכוללת דוברים בחלון בהסתברות 0.8 אחרת 0.2 (דוגמא שלילית: דוגמא שבה אין דוברים בתוך תחום החלון והזווית. דוגמא חיובית: דוגמא בה יש דובר אחד לפחות בתוך תחום החלון והזווית).

לאחר ההגרלות, לוקחים את הדוברים שנמצאים בתוך החלון - w והזווית - θ , סוכמים אותם, מטמיעים את הזווית בתוך סכום ערוצי הדוברים בחלון עם פונקציית השיפט והתוצאה תהיה ה-Ground Truth.

מצד שני לוקחים את המיקס ששמרנו קודם בבניית הדאטא ומבצעים עליו שיפט כדי להטמיע את הזווית – θ, בערוצי המיקס בחדר, ומעבירים את המיקס שעבר שיפט במודל במטרה שהמודל יסנן את כל מה שמחוץ לחלון - w.

משווים בין התוצאות בL1 LOSS ומבצעים ADAM בהתאם לתוצאה.

מטרת האימון היא למזער את פונקציית ה**LOSS** הבאה:

$$L(x; s_1, \dots, s_N, \theta_t, w) = \|\tilde{x} - \sum_{i=1}^{N} s_i' \cdot I\left(\theta_t - \frac{w}{2} \le \theta_i < \theta_t + \frac{w}{2}\right)\|_1$$

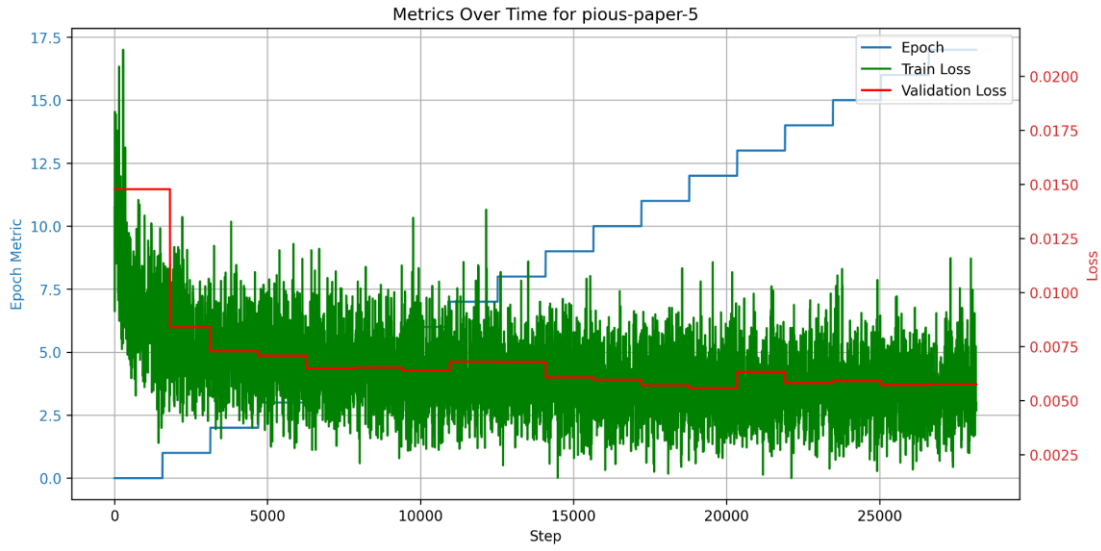
המשמעות היא שאנחנו עושים L1 LOSS וצריכים לחסר את התוצאה של המודל אחרי השיפט עם סה"כ הדוברים בתוך תחום החלון.

ביצענו את האימון על המודל כ 40 אפוקים (מעברים על כל המידע) בגדלי באץ' של 8, קצב אימון (Learning rate) של 0.0003.

תוצאות

Train and Val Loss

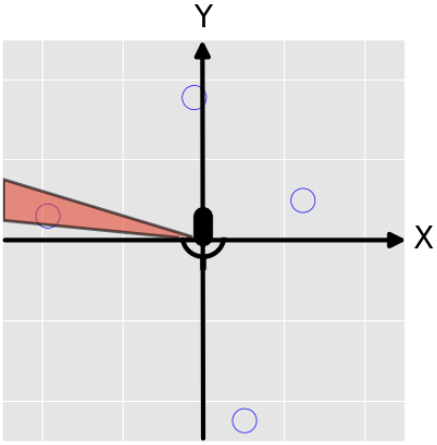
לאחר שאימנו את הרשת קיבלנו את גרף הלוס הבא:

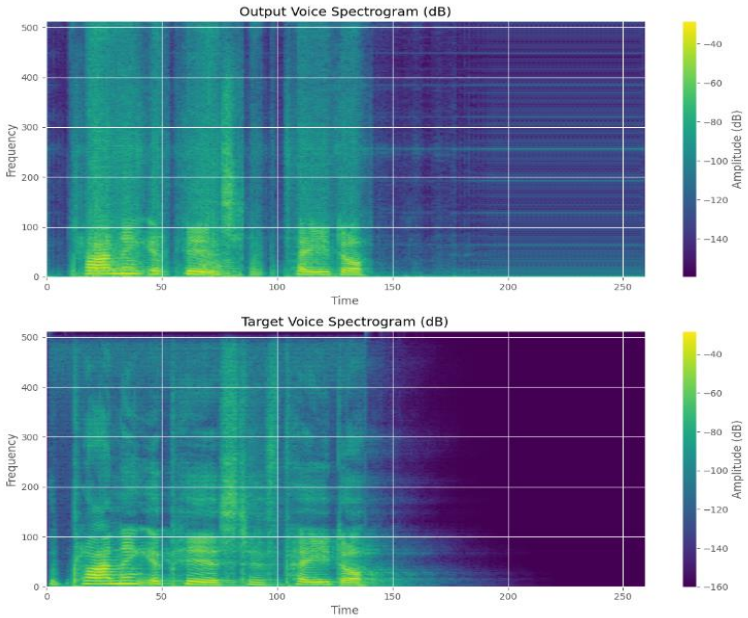


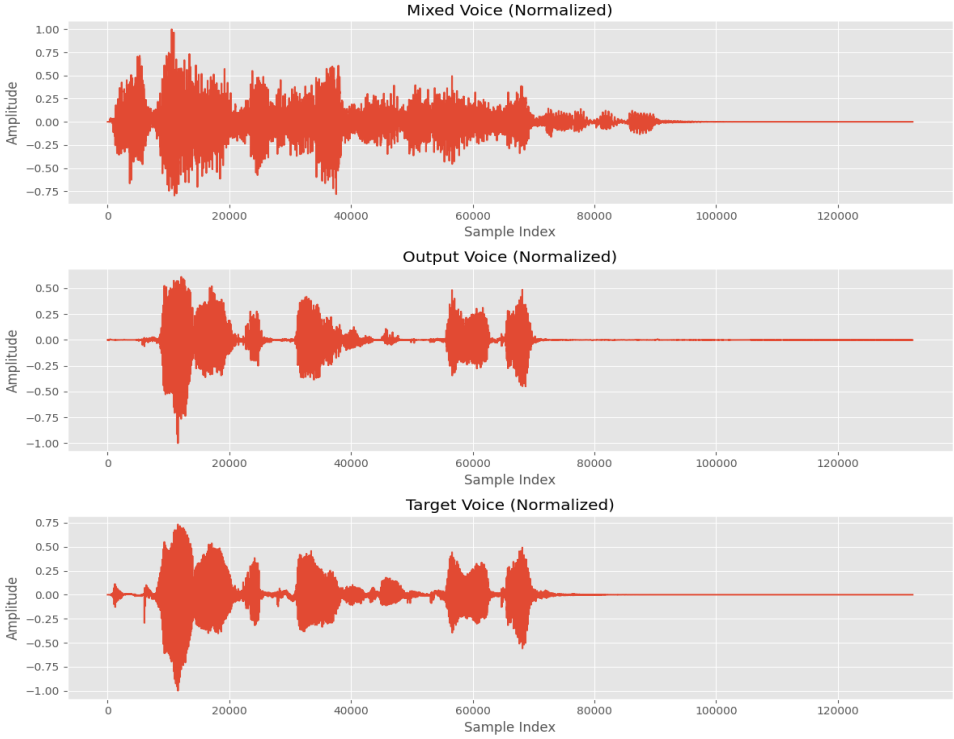
בכחול רואים את האפוקים, באדום את תוצאות הוולידציה שהתבצעה על 1000 דוגמאות בכל אפוק, ובירוק את הTRAIN כנגד מספר הבאצ'ים.

אנחנו רואים כי תוצאות הTRAIN והVALIDATION תואמות ואין בעיית OVERFITTING, וגם רואים שקיימת התכנסות.

הפרדה מתוך זווית וחלון







בדוגמא הנ"ל ניתן לראות שביצענו את הניסוי על 4 דוברים שממוקמים בחדר.
ניתן לראות שעבור מימד הזמן, האמפליטודה של התוצאה דומה בגודל עם מעט שוני בין המטרה לבין התוצאה מהמודל.

עבוד מימד התדר קיבלנו תוצאה דומה מאוד עבור הזמן אשר בו הדובר בזווית המטרה דיבר. לאחר שהדובר סיים לדבר המודל לא הנחית לגמרי את הקול אבל עדיין ביצע הנחתה טובה.

הפרדת דוברים באמצעות הרשת

עקרונות נדרשים

SISDR:

SISDR (Scale-Invariant Signal-to-Distortion Ratio) הוא מדד להערכת איכות אותות לאחר עיבוד, במיוחד בהקשרים של הפרדת דוברים ועיבוד שמע. המונח מתייחס ליחס בין האות המקורי לבין העיוותים שנוספו במהלך תהליך העיבוד, תוך כדי התחשבות בשינויים בקנה המידה של האות.

חשיבות **SISDR**:

SISDR משמש ככלי להערכה של מערכות עיבוד שמע, כמו מערכות הפרדת דוברים, והוא מספק מדד איכות אובייקטיבי שמאפשר להשוות בין שיטות שונות של עיבוד אותות.

זהו מדד יעיל במיוחד כאשר רוצים לקבוע את ההשפעה של אלגוריתמים על איכות האות, תוך התמקדות בהפרדה נכונה של דוברים.

באופן כללי, SISDR הוא כלי חשוב בהערכת ביצועי מערכות עיבוד שמע ועוזר להבטיח שהתוצאות הן באיכות גבוהה.

חישוב SISDR

1. הגדרת האותות: נניח שיש אות מקורי $s(t)$ ואות מעובד $\hat{s}(t)$.
2. חישוב רכיבי השגיאה: שגיאה או עיוות נחשב כהפרש בין האות המקורי לאות המעובד:
$$e(t) = s(t) - \hat{s}(t)$$
3. עוצמת האות:

חישוב עוצמת האות המקורי:

$$P_s = \frac{1}{T}\int_0^T |s(t)|^2 dt$$

חישוב עוצמת העיוות:

$$P_e = \frac{1}{T}\int_0^T |e(t)|^2 dt$$

4. חישוב $SISDR$:

$$SISDR = 10log_{10}\left(\frac{P_s}{P_e}\right)$$

P_s היא עוצמת האות המקורי והמשתנה P_e היא עוצמת העיוות.

משמעויות

אם ה-SISDR גבוה, זה מצביע על כך שהאות המעובד קרוב יותר לאות המקורי עם פחות עיוותים.

SISDR מתמקד בשימור התכנים הרצויים של האות, ולכן הוא כלי שימושי בהערכת ביצועי אלגוריתמים להפרדת דוברים או עיבוד שמע.

NMS (Non-Maximum Suppression):

NMS הוא אלגוריתם המיועד למנוע כפילויות או זיהוי שגוי של אובייקטים (כמו מקורות קול או אובייקטים בתמונה) בתהליכי עיבוד אותות ובזיהוי תמונה. הוא משמש בדרך כלל בסביבת למידת מכונה, במיוחד בעבודות של זיהוי אובייקטים, כמו גם בהפרדת דוברים.

דרך הפעולה של האלגוריתם:

פלטים ראשוניים: בתהליכי זיהוי ומיקום דוברים, המודל מספק מספר זיהויים אפשריים של דוברים בזוויות שונות, כאשר לכל זיהוי יש:

- זווית (angle): המיקום הזוויתי של הדובר.
- אנרגיה (energy): מדד לעוצמת הקול של הדובר.
- נתונים קוליים (data): האות הקולי המופרד של הדובר.סינון ראשוני:

האלגוריתם מתחיל בסינון זיהויים עם דרגת ביטחון נמוכה. נשאר עם רשימה של זיהויים שמעל לסף ביטחון שנקבע.

סינון ראשוני: האלגוריתם מתחיל בסינון זיהויים עם אנרגיה נמוכה מ-**ENERGY CUTOFF**. נשארים עם רשימה של זיהויים שמעל לסף האנרגיה שנקבע, מה שמבטיח שרק דוברים משמעותיים ייכללו בהמשך.

סידור לפי רמת ביטחון: ממיינים את הזיהויים שנשארו לפי האנרגיה מהגבוהה לנמוכה. האנרגיה משמשת כמדד לביטחון בזיהוי הדובר.

בחירת זיהוי מקסימלי: מתחילים עם המועמד הראשון ברשימה (זה עם האנרגיה הגבוהה ביותר) ומשאירים אותו כתוצאה סופית ראשונית.

השוואת זיהויים: משווים את הזיהוי הנבחר עם שאר הזיהויים ברשימה כדי לזהות כפילויות.

בדיקת חפיפה במיקום: מחשבים את המרחק הזוויתי בין המועמד הנוכחי למועמד הנבחר. אם המרחק הזוויתי קטן מרדיוס **CUTOFF** מוגדר, זה מצביע על כך שיש חפיפה משמעותית במיקום, והמועמד נחשב כחופף.

בדיקת חפיפה בתוכן: מחשבים את ה-**SDR** (יחס אות לעיוות) בין הנתונים הקוליים של המועמד והנתונים של המועמד הנבחר באמצעות **SI-SDR**. אם ה**SDR** גבוה מ**CUTOFF** מוגדר אז המועמד חופף בתוכן.

חזרה: חוזרים על תהליך הבחירה וההשוואה עבור הזיהויים הנותרים. בכל פעם, המועמד עם האנרגיה הגבוהה ביותר מהרשימה המעודכנת נבחר ומשווים אותו למועמדים הנותרים, עד שלא נשארים עוד זיהויים.

פלט סופי: התוצאה הסופית היא רשימה של זיהויים ללא כפילויות.

במקרה של הפרדת דוברים, אם המודל מזהה שני דוברים קרובים מאוד זה לזה עם תוכן דומה (למשל, כשיש חפיפה גבוהה בין התחומים), **NMS** מסייע לשמור על אחד מהם בלבד (לרוב זה עם האנרגיה הגבוהה יותר) כדי למנוע כפילויות בשלב הסופי.

יתרונות של NMS**:**

- פשטות: NMS הוא פשוט יחסית ליישום ודורש חישובים מינימליים.

- יעילות: האלגוריתם מספק סינון יעיל של תוצאות, במיוחד כאשר יש ריבוי זיהויים קרובים.

חסרונות של NMS**:**

- איבוד מידע: לפעמים NMS עלול להפסיד זיהויים חשובים כאשר יש חפיפות גבוהה מאוד בין התחומים.

- רגישות לסף: התוצאה תלויה במידה רבה בסף החפיפה שנבחר, שיכול להשפיע על מספר הזיהויים שיישמרו.

NMS הוא כלי יעיל ועוצמתי בתהליכי עיבוד אותות ובזיהוי אובייקטים, והוא עוזר להבטיח שהתוצאות הסופיות יהיו מדויקות ורלוונטיות.

אלגוריתם חיפוש בינארי להפרדת דוברים

האלגוריתם מבצע חיפוש בינארי במרחב הזוויתי כדי למצוא את המיקום של מקורות השמע (הכיוון שממנו מגיע כל מקור), ובמקביל מפריד את האותות מכל מקור. הוא מתחיל מחלון זוויתי גדול ומצמצם אותו בהדרגה עד שהוא מוצא את הזוויות המדויקות של המקורות:

1. חלון זוויתי התחלתי: מתחילים עם חלון זוויתי גדול ω של $90°$ וזוויות יעד ראשוניות של $\theta_0 = \{-135°, -45°, 45°, 135°\}$, המייצגות ארבעה כיוונים במרחב הזוויתי.
כמו כן, יש לנו את תערובת האותות שהוקלטו מ-M מיקרופונים המתוארת כ $x^{MxT} \in R$

2. מעבר ראשון: במעבר הראשון, מפעילים את הרשת (המודל) COS על הזוויות הראשוניות θ_0 כדי להפריד את התערובת לפי רבעונים במרחב הזוויתי. רבעונים ללא מקורות קול ייצרו תוצאות ריקות, ולכן ניתן לדחות אותם ולהתמקד בזוויות שבהן יש פעילות אקוסטית משמעותית.

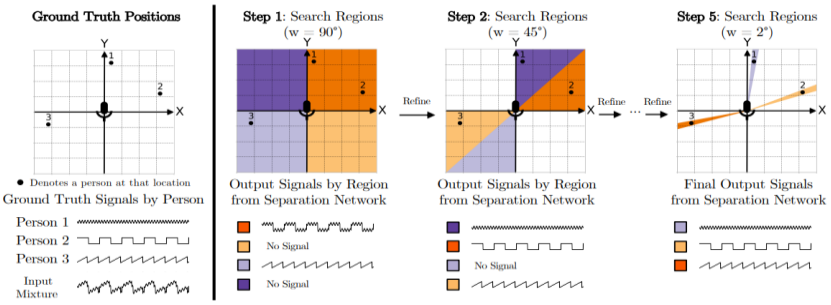
3. צמצום החלון: לאחר המעבר הראשון, מצמצמים את גודל החלון הזוויתי ω בחצי (למשל, $\omega_1 = 45°$ כזכור התחלנו ב $90°$) ומעדכנים את הזוויות לפי האזורים שבהם הייתה פעילות גבוהה במעבר הראשון θ_{0i}.
הזוויות החדשות יהיו $\theta_1 = \{\frac{\omega_0}{2} \pm \theta_{0i}\}$.

4. המשך החיפוש: ממשיכים לצמצם את החלון הזוויתי ω במעברים הבאים על פי אותו עיקרון, עד שמגיעים לרזולוציה הרצויה של זוויות המדויקות.

5. הפרדת מקורות סופית: בשלב האחרון, האלגוריתם מבצע Non-Maximum Suppression כדי למנוע כפילויות בין מקורות קרובים. אם יש שני מקורות הממוקמים קרוב אחד לשני ויש להם תוכן דומה, האלגוריתם מסיר את אחד מהם (את המקור עם האנרגיה הנמוכה יותר).

עקרונות מרכזיים:

1. חיפוש בינארי במרחב הזוויתי: האלגוריתם מחפש את מקורות הקול בזוויות באמצעות צמצום הדרגתי של החלון הזוויתי, מה שמאפשר להגיע **לתוצאה בזמן לוגריתמי**.
2. הפרדת מקורות תוך כדי החיפוש: בכל שלב של החיפוש, הרשת מפרידה את המקורות השונים ומסננת את האזורים שבהם לא נמצאים מקורות, כדי לייעל את החיפוש.
3. Non-Maximum Suppression: כדי למנוע כפילויות, האלגוריתם מסיר מקורות קרובים שהם דומים זה לזה, תוך התחשבות במרחק הזוויתי שלהם ובתוכן האקוסטי.



תוצאות מעשיות

Precision - מייצג את היחס בין מספר המקורות שזוהו נכון על ידי המודל ביחס לכל המקורות שזוהו

Precision = TP/(TP+FP)

recall - מייצג לנו את היחס בין מספר המקורות שהמודל זיהה נכון ביחס לכל המיקומים הנכונים

Recall = TP/(TP+FN)

False Positive (FP) - מיקום דובר שהמודל זיהה אבל הדובר לא נמצא שם.

False Negative (FN) - מיקום דובר שהמודל היה צריך לזהות אבל לא זיהה.

True Positive (TP) - מיקום דובר שהדובר היה שם באמת והמודל זיהה נכון.

sdri -נותן לנו מדד על ההפרדה, מייצג דומה הצליל של ה **GT** ל**OUTPUT**

angular error - מייצג לנו כמה רחוקים בממוצע הזיהויים של המיקומים בזווית מה**GT**

בצענו הערכה של SISDR על סט של 200 חדרים בהם 2 דוברים ורעש רקע.
קבלנו את התוצאות הבאות:

Overall Precision: 0.96

Recall: 0.82

Median Angular Error: 2.01

Median SDRi: 12.21

השוואה למודלים קיימים

בחנו את יכולת המודל להפריד את הדוברים ביחס לאלגוריתמים אחרים קיימים שפועלים במימד התדר: **Ideal binary mask, Ideal ratio mask**:

Ideal ratio mask:

Ideal Ratio Mask (מסכת יחס אידאלית) מיועדת להפרדת מקורות קול על ידי קביעת כמה מהאות של כל מקור קיים בכל תיבה של זמן-תדר באות האודיו. לדוגמה, אם ישנם שני מקורות קול (כמו שני דוברים), ה-**IRM** יחשב את היחס בין כמות האנרגיה של כל מקור בתיבה מסוימת ויישם את היחס הזה כמסכה. הנה כיצד זה עובד:

- חישוב יחס: עבור כל תיבה בזמן-תדר, ה-**IRM** מחשב את היחס בין האנרגיה של הצליל המבוקש לבין האנרגיה המשולבת של כל המקורות.
- יישום מסכה: היחס הזה פועל כגורם קנה מידה (בין 0 ל-1) שמצביע על כמה מהצליל בתיבה הזו יש לשייך למקור המבוקש.

על ידי יישום **IRM**, משמרים את החלקים של האות שבהם הדובר הרצוי דומיננטי, אבל בצורה חלקה יותר מאשר החלטות בינאריות, מה שמוביל להפרדות רכות יותר.

Ideal binary mask:

Ideal Binary Mask (מסכת בינארית אידאלית) היא גישה פשוטה יותר שמבצעת הפרדה על ידי קבלת החלטות בינאריות (0 או 1) בכל תיבה בזמן-תדר. היא נועדה לשמור על מקור היעד בתיבות שבהן האנרגיה שלו חזקה מזו של מקורות אחרים, ולדכא את היתר. הנה כיצד היא פועלת:

- החלטה בינארית: עבור כל תיבה בזמן-תדר, **IBM** נותן ערך 1 אם האנרגיה של מקור היעד בתיבה עולה על סף מסוים, ו-0 אחרת.
- יישום מסכה: על ידי יישום המסכה הבינארית הזו, **IBM** שומר רק את התיבות שבהן מקור הצליל הרצוי חזק ביותר, ומסלק את השאר לחלוטין.

ביצענו הערכה של **SISDR** על סט של 200 חדרים של 2 דוברים יחד עם רעש רקע,

$N = 2$, $x = \sum_{i=1}^{N} s_i + bg$, קיבלנו את התוצאות הבאות:

Ideal Binary Mask:

Median SI-SDRi: 11.23

Ideal Ratio Mask:

Median SI-SDRi: 6.13

Cone Of Silence:

Median SDRi: 12.21

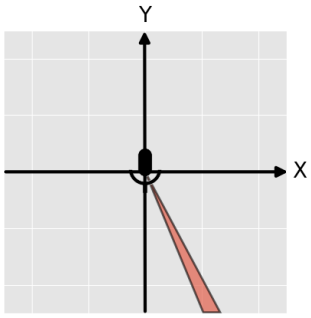
Real life + Moving sources:

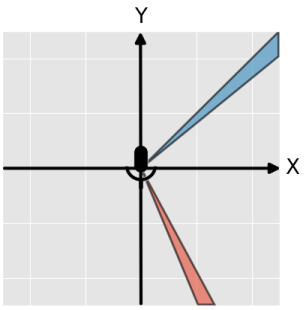
הרצנו את המודל על מידע שקיבלנו ממערך מיקרופונים של **RESPEAKER CORE V2** עם סאונד בפורמט 32 **BIT** בתדר דגימה של 44100 לשנייה , חילקנו למקטעים של 3 שניות וכל אחד מהם הכנסנו למודל בנפרד.

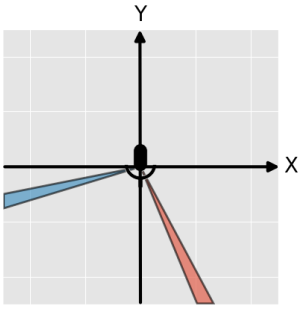
ניסוי:

2 דוברים כאשר בהתחלה הדיבור היה ביחד ואז אחד מהדוברים לא דיבר ושינה מיקום בכ - 180 מעלות.

הקלטה רציפה.







המודל הורץ על חדר במימד קטן יותר ממה שהמודל אומן עליו ובהדהוד גבוה, ועדיין נתן הפרדת דוברים ומיקום באופן נכון.

נספחים

נספח א' – מעבר הנתונים דרך הבלוקים במודל

בכניסה יש שכבת קונבולוציה עם פילטר בגודל 8 וסטרייד בגודל 4 ומספר ערוצים במוצא של 64,
(33074=batchSize,) -> (BatchSize,64,1 + $\frac{132300-8}{4}$6,132300) מה שמגדיל את מספר הערוצים פי 2
ומקטין את הווקטור הכניסה.

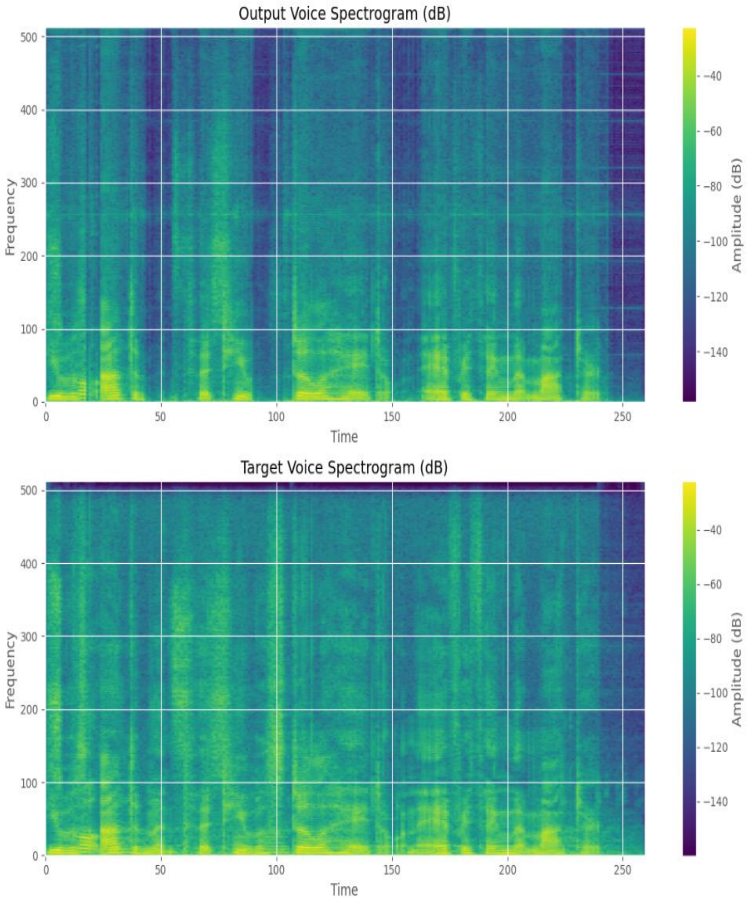
לאחר מכן מוסיפים את החלון שמועבר בשכבה ליניארית כך שהמימד שלו יהיה (- BatchSize,5)
(64,1,>BatchSize). מחברים חיבור ווקטורי בין 2 התוצאות ומפעילים פונקציית RELU על התוצאה.

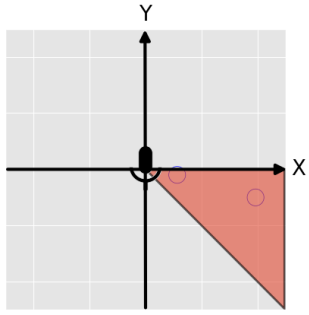
נקח את התוצאה הזו ונפעיל עוד קונבולוציה שתשנה (batchSize,) -> (BatchSize,64 ∗ 2,3307364,33074)
ונוסיף שוב את הזוויות לתוצאה זו ונעביר בפונקציית אקטיבציה של GLU שתחזיר אותנו למימד של
(BatchSize, 64 , 33074) שיכנס בתור הקלט לכבה הבאה וגם בתור קלט דרך הSKIP CONNECTION למפענח
המקביל.

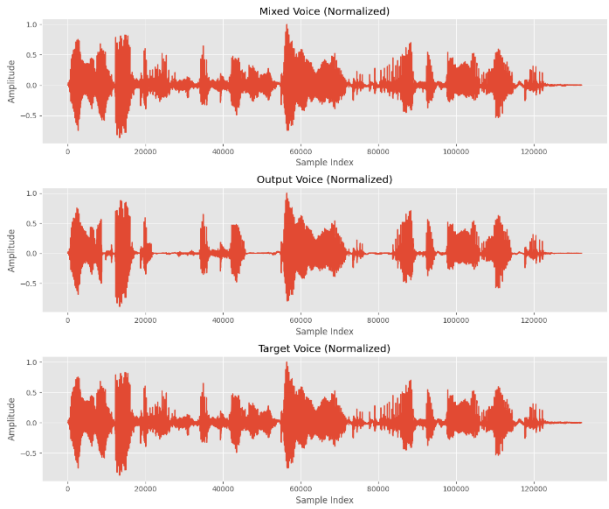
בכל מקודד נמשיך להגדיל פי 2 את מספר הערוצים תוך כדי הקטנה של ווקטורי הכניסה, כדי להגיע לווקטור
CONTEXT שיעבר בLSTM , המוצא עובר חזור בתהליך דומה רק הפוך.

נספח ב' – ניסויים נוספים של הפרדה דוברים

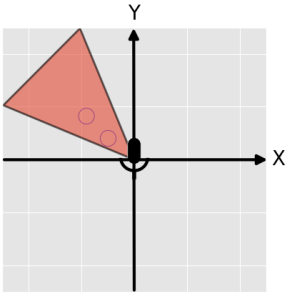
ניסוי 1: (2 דוברים)

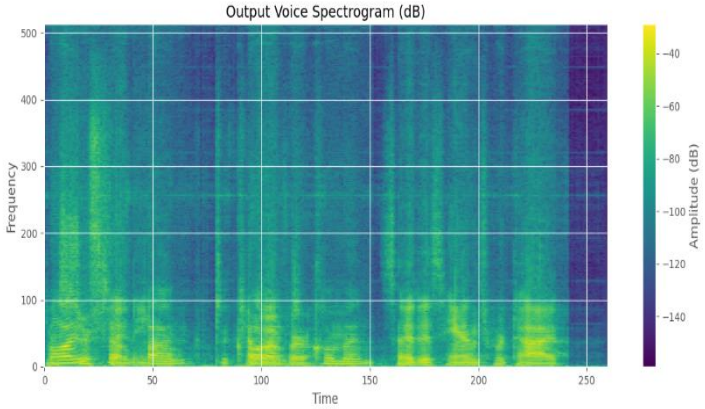


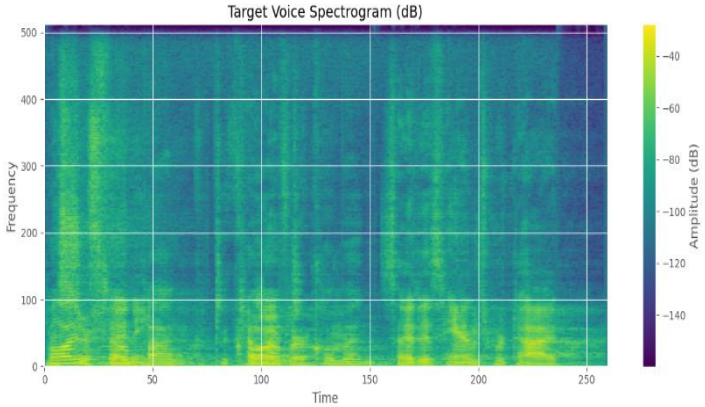


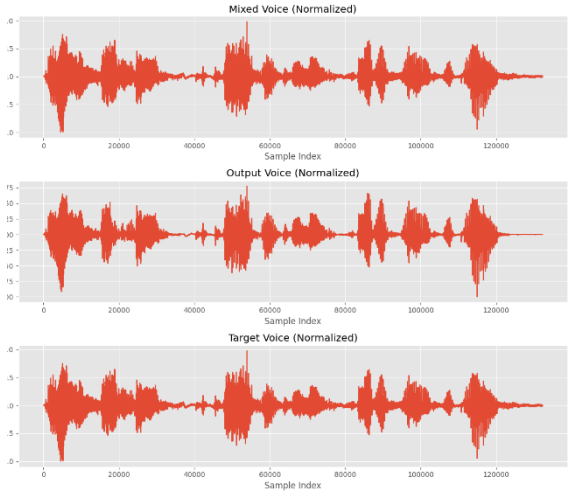


ניסוי 2: (2 דוברים)

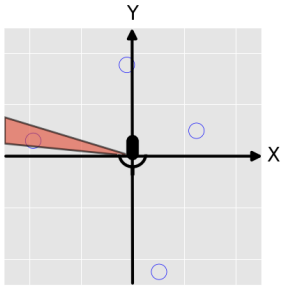


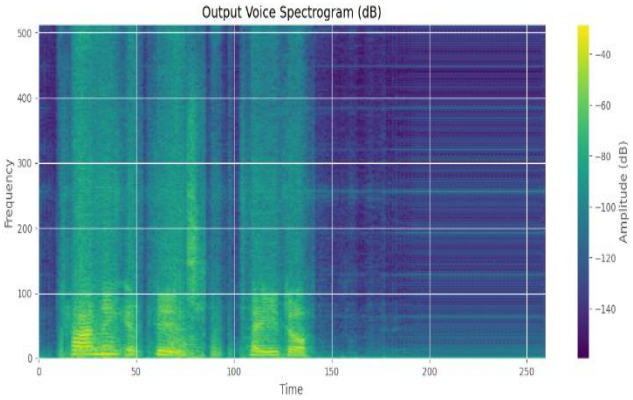


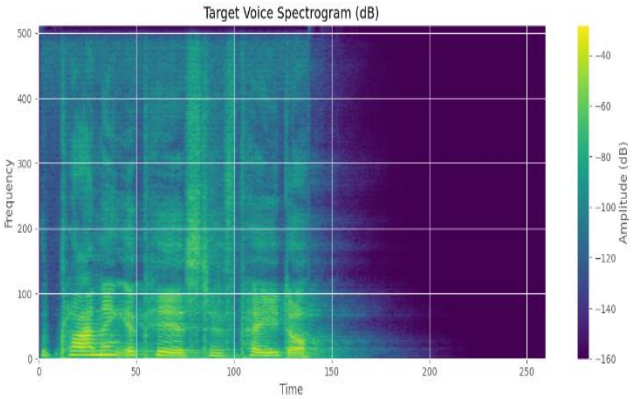


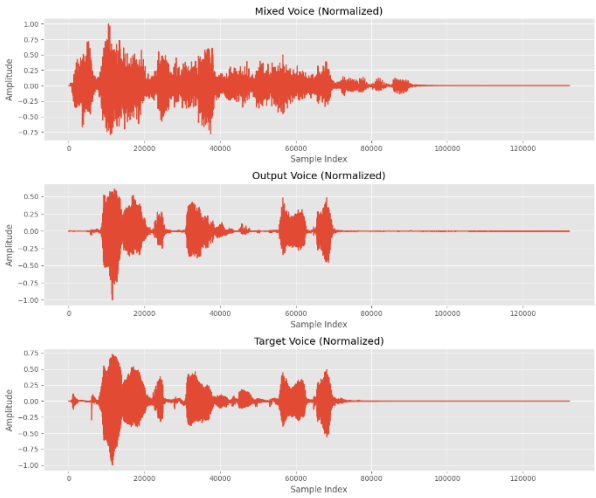


ניסוי 3: (4 דוברים)

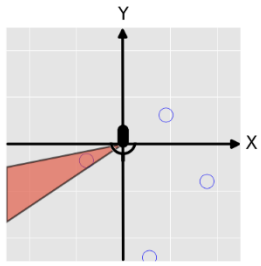


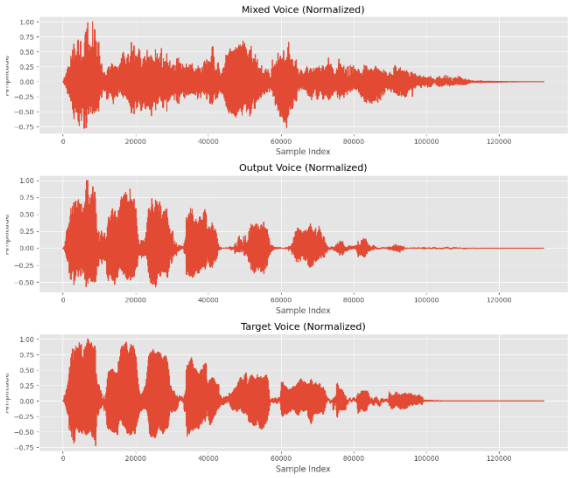


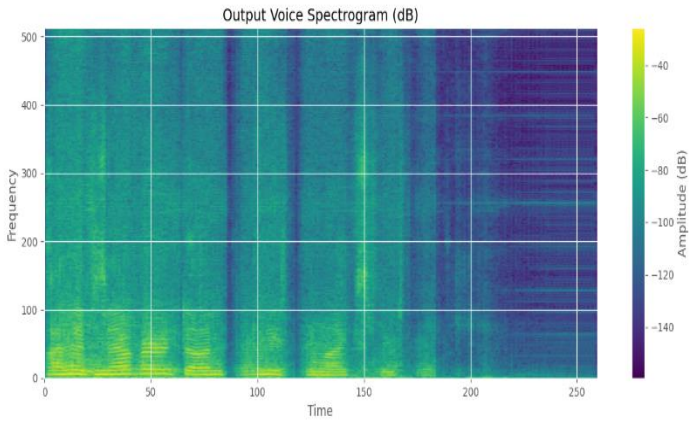


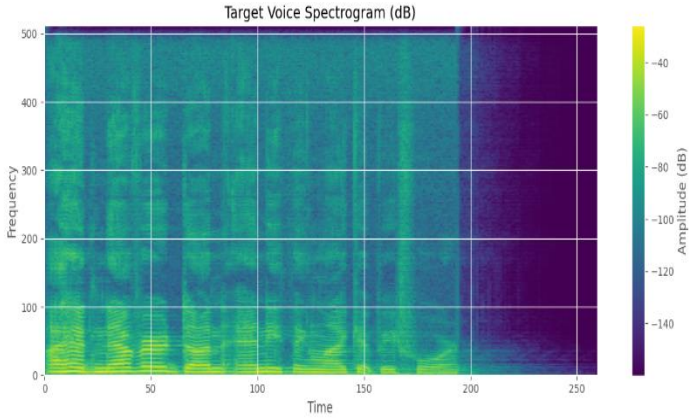


ניסוי 4: (4 דוברים)

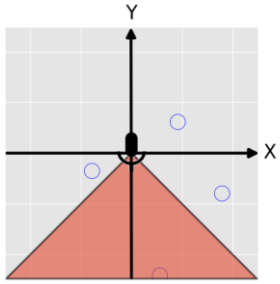


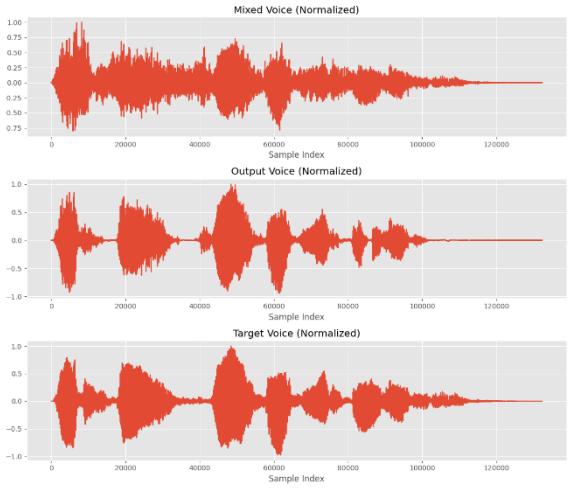


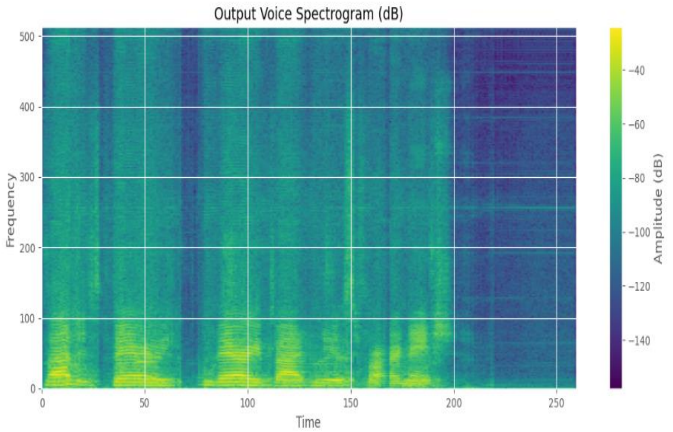


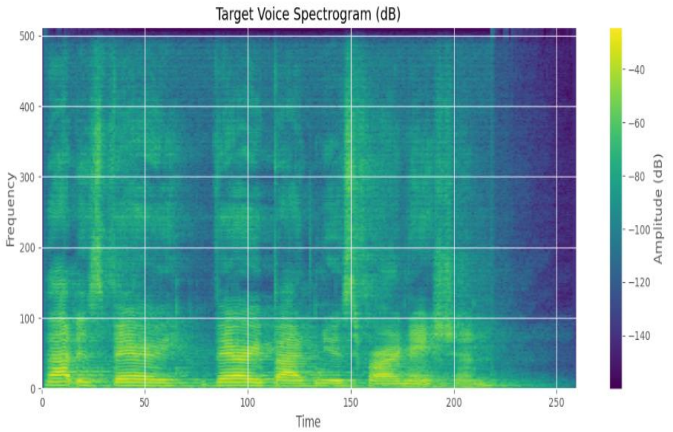


ניסוי 5: (4 דוברים)

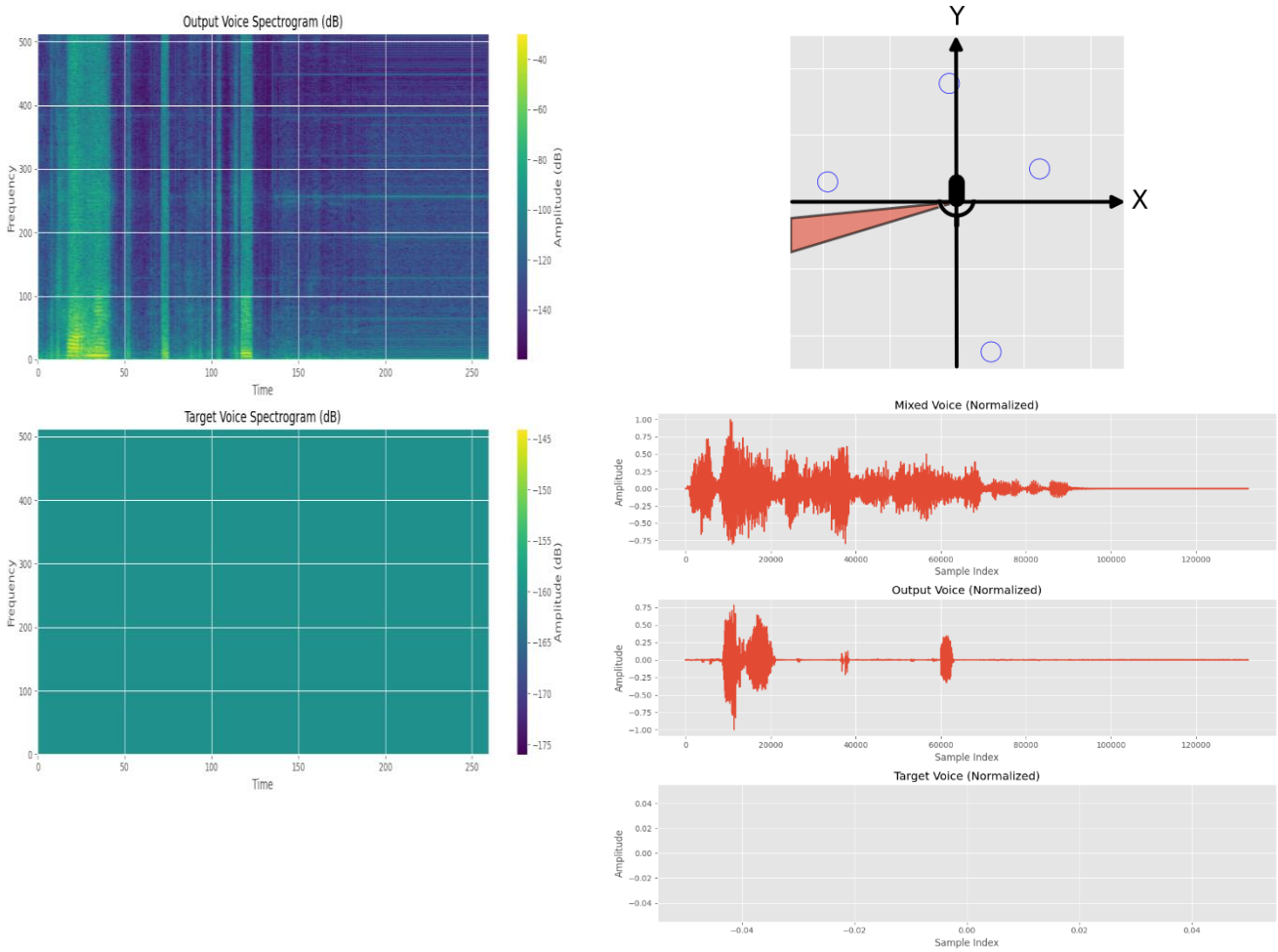




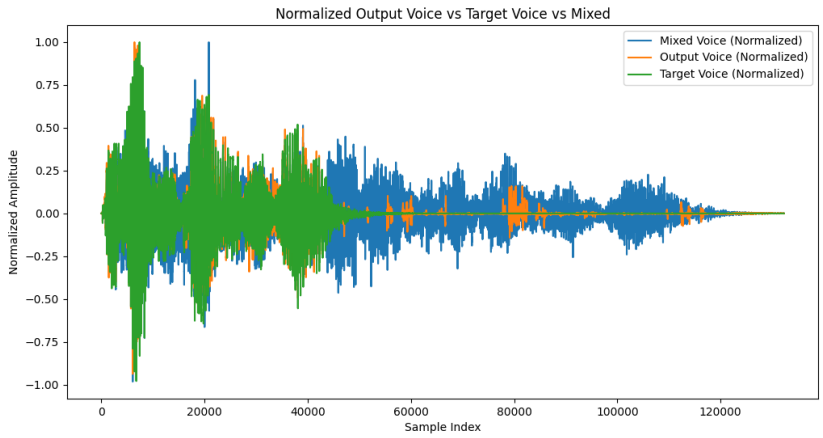




ניסוי 6: (4 דוברים)



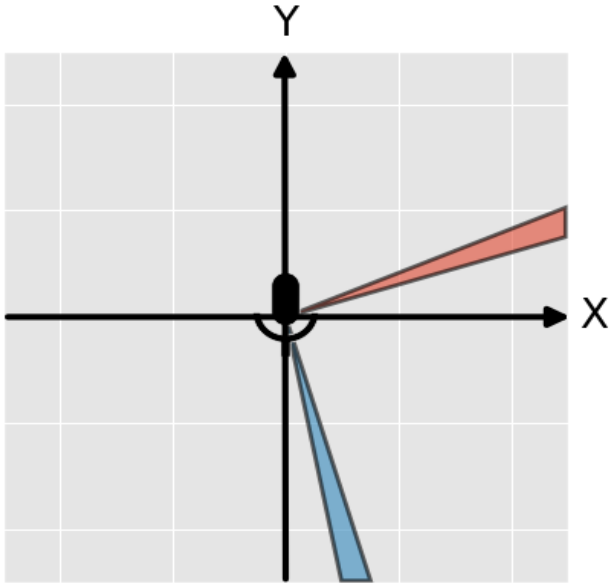
תוצאת המודל בפרדה מחדר עם 3 דוברים, חלון המכיל דובר אחד: (ירוק מקור, כתום תוצאה)



נספח ג' – מיקום דוברים לפי TEST DATA:

מיקום של דוברים מהטסט אשר נמצא במטאדאטא –

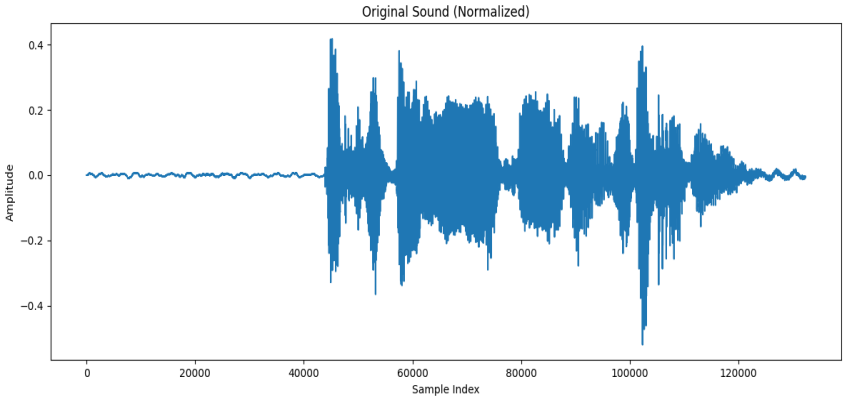
```
root@a87a88c956f8:/app/gannot-lab-UG/ilya_tomer/OUTPUTS/OUTPUTS_BASIC_TRAIN/TESTS_1311/output00004# cat metadata.json
}
"    voice00} :"
"        Position] :"
,1.678231063177929-
,0.24634647240655028
1.45619829420323
,[
"        speaker_id": "p297"
,{
"    voice01} :"
"        Position] :"
,0.7962293789336705-
,0.8301825752545327-
1.4957048204674657
,[
"        speaker_id": "p247"
,{
"    bg} :"
"        Position] :"
]
,14.08588110287385-
,5.577052658240859
1.4309471850596625
[
[
{
{r
```

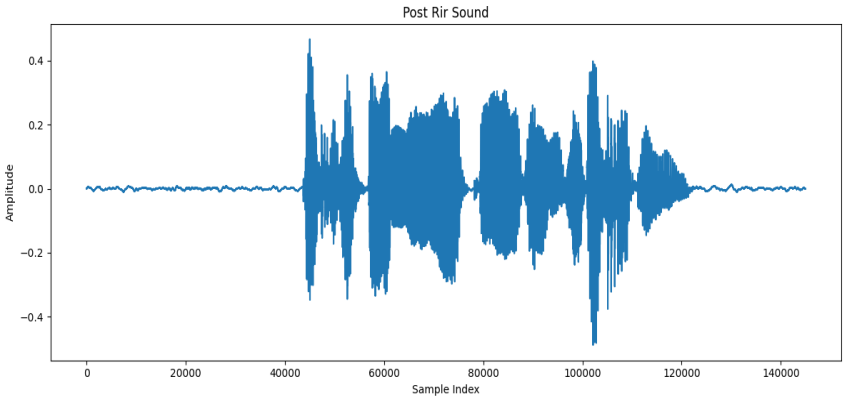


נספח ד' – תוצאות ROOM GENERATION:

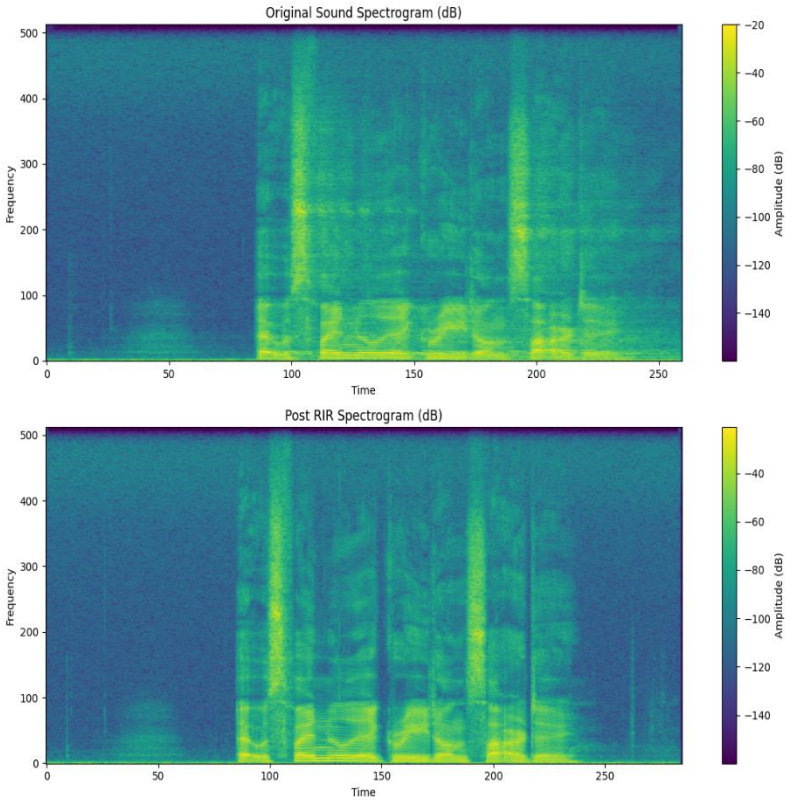
מצורפות דוגמאות של אותות לפני ואחרי שעברו RIR:

מימד הזמן:





מימד התדר:


Original Sound Spectrogram (dB)
Frequency
Time
Amplitude (dB)
Post RIR Spectrogram (dB)
Frequency
Time
Amplitude (dB)