תמונה שמכילה גופן, לוגו, סמל, גרפיקה

התיאור נוצר באופן אוטומטי

**ספר פרויקט גמר**

**The Cone of Silence: Speech Separation by Localization**

**איליה ליפובן תומר להב**

**תוכן עניינים:**

**1. הצגת הבעיה............................................................עמ2**

**2. פתרון הבעיה............................................................עמ2**

**3. בניית הDATA...........................................................עמ3**

**4. בניית הDATASET.....................................................עמ5**

**5.הקדמה לרשת העמוקה..............................................**

**6.Cone Of Silence Network......................................**

1. **הצגת הבעיה:**

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

קשה להסיק סטטיסטיקה על הדוברים שהיא ידוע 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 יותר, כגון : Wave-U-Net,Tasnet,Conv-Tasnet.

1. **פתרון הבעיה:**

בבסיס הפתרון לבעיה אנו משתמשים ברשת עמוקה אשר מעוצבת בדומה לWave-U-net רשת במישור הזמן שמפרידה בין הדוברים, בהינתן חלון וזווית לרשת כאשר מדובר בחלון ברוחב , אשר רוכב סביב זווית .השרת מאומנת על ידי למידה Supervised.

על ידי חלוקה אקספוננציאלית של גודל החלון אנחנו יכולים לעשות חיפוש בינארי סביב מעגל במישור האזימוטי ולמקם את הדוברים על ידי שימוש ברשת בכל חלוקה וגם להפריד את הדוברים בזמן לוגריתמי.

תמונה שמכילה שרטוט, ציור, תרשים, עיגול

התיאור נוצר באופן אוטומטי

*מכיוון שהרשת שלנו משתמשת במישור הזמן נוכל להשתמש בטכניקות beamforming כדי לשנות את הפוקוס בין הדוברים השונים במישור הזמן, על ידי שימוש בהזזה בזמן ובכך לכלול את זווית הדובר בתוך הקול עצמו ולהעביר לרשת חלון ואות בלבד.*

**סימונים לאורך ההסברים בספר:**

M- מספר המיקרופונים

T- אורך הסיגנל

X- הסיגנל, ווקטור בין מינוס אחד לאחד באורך T

– הזווית של הדובר/סיגנל

W גודל חלון

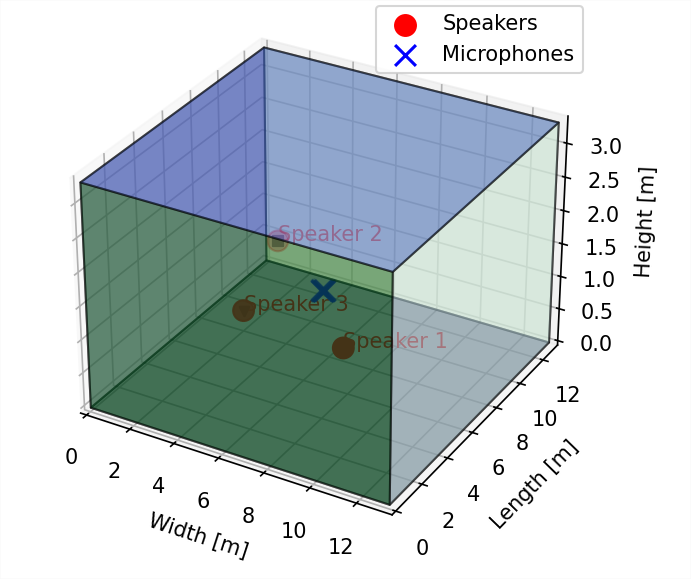
1. **יצירת הDATA**

**3.1 הסבר עד הרעיון והתוצאה הרצוייה:**

נרצה ליצור סימולציה של חדר של דוברים שמדברים סביב מערך מעגלי של מיקרופונים, כך שהדוברים לא יהיו קרובים מידי אחד לשני , למערך של המיקרופונים, לקירות. וליצור רעש דיפוזי. השתמשנו בהקלטות מספריית LibriSpeech , וכדי לסמלץ את הדיבור בחדר השתמשנו בספריית PyRoomAcoustics כדי לקבל Room Impulse Response שאנחנו מעבירים אותו דרך קונבולוציה.

בסוף אנחנו מקבלים שלכל חדר שאנחנו מסמלצים יהיה  *.*

הסבר: לכל ערוץ, אנחנו רוצים לסמלץ דיבור של כל אחד מהדוברים בחדר, ולכן  *, לכל ערוץ אנחנו רוצים גם לסמלץ בנפרד את הרעש הדיפוזי בנפרד, וגם את כל הדוברים ביחד.*

*איור חדר לדוגמא, ללא ויזואליזציה של דוברי רעש דיפוזי:*תמונה שמכילה טקסט, צילום מסך, תרשים, מלבן

התיאור נוצר באופן אוטומטי

**3.2.1 אופן יצירת הDATA**

יצרנו אותות דיבור שמסמלצים דיבור בחדר מול מערך של מיקרופונים, אנחנו מגרילים גודל חדר ברוחב ואורך בין 12 ל15 וגובה בין 3 ל4, מגרילים בין 2 ל4 דוברים בחדר, כך שהדוברים נמצאים במרחק של 1 עד 4 ממרכז מערך המיקרופונים וכל דובר נמצא במרחק לפחות של חצי מטר מהשני.

ההקלטות אורכו/קוצרו ל5 שניות.

נחשב RIR לכל אחד מהדוברים ביחס לכל אחד מהמיקרופונים ונקבל את הרצוי, RIR לדוגמא:

שמים 4 דוברים מוגרלים ב4 צדדי החדר עם רווח של 0.2 מטר כדי לסמלץ רעש דיפוזי ושמים מערך מיקרופונים מעגלי כך ש מיקרופונים במרכז החדר , יש אפשרות לשינוי מיקום ומספר המיקרופונים.

רעש דיפיוזי : 4 קולות שממוקמים ב4 צידי החדר : לאחר מעבר דרך 6 מיקרופונים נקבל לכל רעש RIR:

*נחסיר את ה50 מילישניות הראשונות כדי לסמלץ RIR של רעש ולחסר מהרפלקציות הראשוניות של הקירות.*

ולכל מיקרופון : *.*

*קולות דיבור:*

*יהי N מספר הדוברים, , נמצא את הRIR של כל אחד מהדוברים ביחס לכל אחד מהמיקרופונים*

*Reverb :*

*אומר לנו כמה זמן ביחס לאחוז לוקח לקול לדעוך. נשתמש ב שנסמן גם בrt60\_tgt , אנחנו קובעים דעיכה של 60 אחוז בצליל לכלל המערכת על ידי Inverse Sabine , פונקציה שנותנת את מקדמי הספיגה של החדר שיתן דעיכה כזאת, ומקדם max order שנותן כמות החזרות מקסימלית בין הקירות. משתמשים במקדמים האלה לפרמטרים של החדר בסימולציה של הRIR בPyroomAcoustics. הגדרנו*

**3.2.2 PIPELINE התאמת יחס בין קול לרעש**

*נקבל את הסיגנלים, את הרעש ואת הMIX אחרי סימולציה של החדר נרצה להנחית את הרעש ביחס, אנחנו מבצעים מספר שינויים על הסיגנלים:*

*עבור , נגריל תחילה דובר ולכל דובר ביחס לדובר המוגרל הזה נגריל SIR בין 0ל5, נעשה את זה כדי למנוע מצב שלדוברים באותו מקום יהיו אמפליטודות שונות מידי. לאחר מכן נעביר את כל הדוברים בRIR של החדר .*

*MIX:*

*לכל ,*

*נגדיר*

*כאשר*

*נחשב את הSNR כך שיהיה בין 7 ל10 (ניתן להגדיר פרמטרית)*

*ולאחר מכן נכפיל את הרעש:*

*בדיאגרמה זה יראה כך:*

תמונה שמכילה תרשים, קו, צילום מסך, טקסט

התיאור נוצר באופן אוטומטיתמונה שמכילה תרשים, תוכנית, שרטוט טכני, עיגול

התיאור נוצר באופן אוטומטי

***3.3 שמירת הDATA:****שמרנו 10000 חדרים שונים מהTRAIN כך שכל ההקלטות בתדר דגימה של 44100 כדי לקבל דיוק טוב יותר באימון, הצעד הזה יתבאר בהמשך, זה התבצע על ידי העלאת תדר דגימה באמצעות SCIPY RESAMPLE. שמרנו לכל חדר שסימלצנו את המידע בנפרד כקובץ JSON בפורמט הבא, הוא יכלול מידע שנחוץ לאימון המודל:*

{

    "voicei": {

        "Position": [

            radius,

            ,

            z

        ],

        "speaker\_id": "id\_from\_librispeach"

    },

"bgj": {

        "position": [

            radius,

            ,

            z

        ]

    },

*כמו כן שמרנו גם פרמטרים נוספים בקובץ CSV שלא נחוצים לאימון עצמו אבל נחוצים לאבחנה בביצועים ובDEBUG של המודל באופן הבא:*

Parameter,Value

rt60,0.7

room dimentions,"14.8654,13.3937,3.4128"

background speakers,"\5338\24640\5338-24640-0000.wav,\3853\163249\3853-163249-0002.wav,\8297\275154\8297-275154-0026.wav,\6295\244435\6295-244435-0000.wav"

snr,8

**4. בניית הDATASET**

כדי לאמן את המודל שלנו ולתת לו דוגמאות Supervised מהמודל, השתמשנו במחלקת dataset של Pytorch עם שינויים. המידע שאנחנו מביאים למודל כולל את התוצאה הרצוייה,

הרשת שלנו היא SUPERVISED ,מקבלת 2 קלטים X' וW , המטרה שלנו תהיה לתת לרשת דוגמאות שליליות וחיוביות. המטרה הסופית היא שהדוגמאות החיוביות יתנו LOSS נמוך והדוגמאות השליליות יתנו לנו LOSS גבוהה.

ראשית, נסביר על הLOSS במודל שלנו:

פונקציית LOSS היא פונקציה שמודדת את המרחק בין התוצאה הרצוייה לתוצאה מהמודל בפועל. מטרתינו היא לקבל מהמודל את הסיגנל הנכון במידה והסיגנל בזווית הנתונה נמצא בתוך החלון הנתון אחת נצפה לשקט. אם כן , הLOSS מוגדר באופן הבא:

תמונה שמכילה טקסט, גופן, קו, כתב יד

התיאור נוצר באופן אוטומטי

התוצאה : שהיא כוללת את שיערוך הקולות שנמצאים בסביבת ה במידה ואין קולות אז השיערוך אמור להיות ריק אחרת אמור לכלול את הצלילים שבאזור. לכן נשתמש באינדיקטור, שלוקח לפי זווית המטרה סביב לחלון את הסיגנל הרצוי ומחסר עם המצוי.

הפונקציה היא L1 LOSS משמע, פחות רגישה לדוגמאות שהן חריגות ביחס לL2 שמחשבת הפרשים ריבועיים.

דוגמא שלילית כוללת סיגנל אשר נמצא מחוץ לחלון ודוגמא חיובית תכלול סיגנל שנמצא בחלון.

על מנת לייצר את הדאטא הנ"ל נשתמש בפונקציית SHIFT שפועלת כך:

בהנתן מערך דו מימדי  *וזווית הפונקצייה תבצע הזזה לכל ערוץ באופן שהולך לתת התמרה חח"ע ועל בין הזווית להקלטות ותפלוט: .*

*הדרך שבה היא מבצעת את זה :  
נבחר מיקרופון שיחס אליו נחשב את הנוסחה ונקרא לו mic0, נחשב דילי בין כל אחד מהערוצים של הדובר על ידי Time difference of arrival:*

*ועל כל אחת מהערוצים נבצע הזזה לפי ההפרש הגעה בין mic0 לבין mici וריפוד באפסים חד צדדי:*

דוגמא למודל תיהיה מהצורה הבאה:

דוגמא חיובית:

דוגמא שצריכה לכלול את הדוברים הנכונים ביחס לזווית וגודל החלון.

לשם כך ניקח זווית רנדומית, ניקח את הצליל שהכי קרוב אליה ואם הוא נימצא בחלון נכלול את הצליל וכך נבנה את הGroundTruth

דוגמא שלילית:

נרצה לקחת צליל שלא תואם לזווית ולחלון.

1.ניקח זווית רנדומית 2.נגריל מערך הסתברויות  *שיסמן לנו את הזוויות בהתאם לחלון על מעגל היחידה 3. נגריל הסתברות גבוהה לזווית קרובה לזווית המוגרלת ורחוקה לרחוקה ואפס לזווית עצמה כדי שלא ניקח זווית תואמת. 4.בנוסף המודל בודק את הזוויות לפי הזזות לכן נבדוק שאכן* הזווית שבחרנו שונה לפי ההתמרה של הSHIFT , ואם תואמת אז ז"א שהזווית קרובה מידי כדי שתהיה שגויה אז נחזור על התהליך.

בסוף נחזיר במערך הGroundTruth ,שכולל רק את הצליל בפועל מהזווית השלילית/חיובית ואת הMIX שכולל את הכל ואת הוקטור המקודד של החלון.

**הרשת:**

הרשת שלנו תהיה דומה לU-NET וספציפית תתבסס על ארכיטקטורת Demucs , בפרט המשקלים של הרשת יאותחלו מרשת הDemucs. דיאגרמת בלוקים של ארכיטקטורת הרשת:

תמונה שמכילה טקסט, תרשים, צילום מסך, תוכנית

התיאור נוצר באופן אוטומטי

הסבר על הפעולות ברשת:

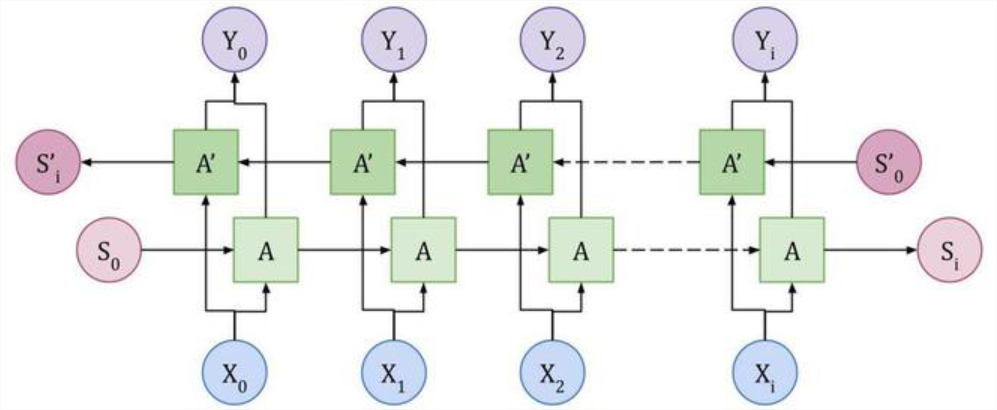
Conv1D:

*פעולה המבצעת הכפלה של ווקטורים כאשר ווקטור אחד הוא ווקטור הINPUT ווקטורי ה kernel כאשר הOUTPUT יהיה הכפלה של הווקטור בכל אחד מהKERNEL ו*

ReLU/GLU:

Bi-LSTM:

שכבה

 תמונה שמכילה תרשים, עיצוב

התיאור נוצר באופן אוטומטי

הארכיטקטורה:

הרשת תקבל בINPUT וh , שהוא ווקטור שמייצג לנו את גודל החלון. תוציא בOUTPUT  *.כאשר X גל זה התוצאה של הרשת לאחר שהיא מנסה להנחית את הדוברים שנמצאים מחוץ לזווית ולגודל החלון שאנחנו נותנים לרשת.*