ניקוי הדהודים ורעשי רקע בעזרת מודל למידה עמוקה

מאי טיגר- may.tiger@campus.technion.ac.il

הילה לוי- hilalevi@campus.technion.ac.il

עבודה זו נכתבה במסגרת פרויקט בבינה מלאכותית (236502) סמסטר חורף תשפ"ג בהנחייתו של ד"ר ליאור ארבל

תוכן עניינים

מבוא

1.1. רקע לבעיה- מוטיבציה

למידה דרך סרטונים מוקלטים היא כלי שימושי שהפך נפוץ ואף הכרחי בתקופת הקורונה. החומר הנלמד בהרצאות בטכניון הוא לעיתים קרובות לא פשוט ומצריך סטודנטים רבים לעבור עליו פעם נוספת, ועל כן, סטודנטים רבים פונים להקלטות על מנת לחזק את שליטתם בחומר.
הטכניון השקיע כסף רב בציוד עבור תמיכה בהקלטות ולמידה מרחוק בתקופת הקורונה אך בפועל הקלטות רבות סובלות מאיכות שמע ירודה, עוצמה משתנה ודגש לא שווה בתדרים הנשמעים בה וסטודנטים רבים מצאו קושי בלהבין את דברי המרצה או המתרגל בהקלטות אלו.
למרות ניסיונות שונים לשפר את איכות השמע באמצעות שינוי הוליום או שינוי מהירות הסרטון, הסטודנטים לא הצליחו להפוך את השמע לברור יותר וקטעים רבים מתוך הסרטונים נשארו לא מובנים כלל.
מכאן נולד הרעיון של הפרויקט שלנו. רצינו ליצור פרויקט שמהותו תהיה שיפור ההקלטות בטכניון, במהלך העבודה על הפרויקט התאמנו את הדרישות שלנו לזמן העבודה הנתון ולגודל הפרויקט.

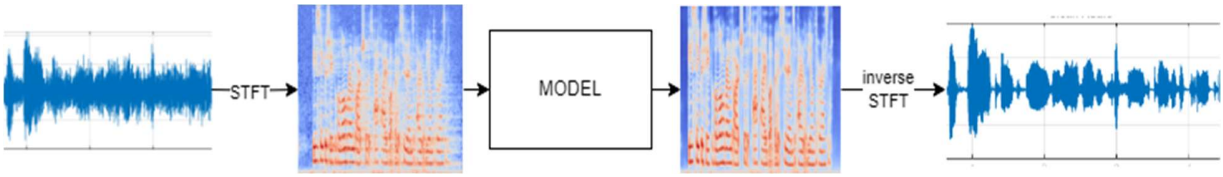
1.2. תיאור הבעיה

לאחר מעבר על מספר רב של סרטונים עם איכות ירודה אשר אותם אספנו מניסיון אישי ותשאול חברים, זיהינו כי אחד המקורות העיקריים לבעיה הוא ההדהוד של החדר בו מוקלטת ההרצאה אשר מעפיל ומטשטש את השמע המקורי (הדיבור הישיר של המרצה).
אומנם מעבר לבעיה זו ישנן בעיות זמניות בהקלטות כגון התנתקויות המיקרופון או רעשים חד פעמיים אך בעיות אלו מפריעות להרצאה באופן רגעי לעומת בעיית ההדהוד בחדר אשר נמשכת לאורך כל זמן ההקלטה ומושפעת באופן ישיר ממאפייני החדר בו היא מתבצעת. על כן, החלטנו להתמקד בבעיה זו שכן אנחנו סבורים כי פתירת בעיית ההדהוד תשפר את איכות השמע בהקלטות ותאפשר לסטודנטים חווית צפייה טובה ומובנת יותר.

פתרון הבעיה

2.1. אופן פתרון הבעיה

על מנת לפתור את הבעיה, החלטנו לנסות להשתמש במודל למידה עמוקה, שיוכל להסיר את ההדהודים מקטעי ההרצאה המוקלטים ולשפר את איכות השמע בהם. אופן ניקוי הרעשים יעשה על ידי חיתוך שמע ההרצאה לקטעים קצרים, הכנסתם כקלטים למודל, וחיבורם חזרה לכדי הקלטה מלאה לאחר שהורדנו את ההדהודים. המודל הבסיסי אותו בחרנו למשימה זו הוא מודל ה LS U-Net (הסבר מורחב בסעיף 4.2) אשר מבצע סגמנטציה (כלומר, מקבל תמונה כקלט ומזהה אובייקטים מרכזיים בה) ולאחר מכן מבצע החסרה בין תמונת המקור לאובייקט המזוהה כך שהתמונה המתקבלת היא של האובייקט המזוהה בלבד.



לאחר שהצלחנו לנקות את הסאונד (ביצוע dereverberation) באמצעות המודל הבסיסי, החלטנו להוסיף למודל יכולת נוספת שתשפר בעיה נוספת שמצאנו בהקלטות ההרצאה. היכולת אותה בחרנו להוסיף למודל היא ניקוי רעשי רקע בהקלטה על ידי מודל מסוג U-net (הסבר מורחב בסעיף 4.1). לבסוף, יצרנו מודל נוסף אשר מובנה מהרכבה של שני מודלים אלו (הסבר מורחב על המודל בסעיף 6.3).

2.2. בדיקות והערכת הפתרון

את בדיקת איכות המודל בחרנו לבצע באמצעות שתי דרכים:
הדרך הראשונה היא שימוש במטריקות הבאות: STOI, CD, LLR, fwSNRseg, SRMR (הסבר מורחב בסעיף 5.3) שיעודן הוא לבחון את הדמיון בין קטעי הקול. לכל אחת מהמטריקות יש אופי השוואה מעט שונה, ולכן שימוש במגוון מטריקות מסייע לנו בהבנה כוללת יותר של איכות המודל. ההשוואה תיעשה בין קטעי דיבור ברורים, לגרסה שונה שלהם- פעם אחת השוואה עם גרסה שעברה הדהוד מלאכותי ובפעם השנייה גרסה שעברה הוספה של רעשי רקע.

הדרך השנייה, אשר תהיה גם השיטה העיקרית לבדיקת איכות המודל, תהיה על ידי הקשבה לקטעי הקול ובדיקה האם חל שיפור ברמת המובנות ובהירות הדיבור של הדובר בהקלטה. כיוון שמדובר בבדיקה סובייקטיבית נערוך סקר בו נבדוק עם אנשים שונים מה דעתם על איכות השמע בהקלטות השונות. באמצעות המענה על הסקר נקבע את שיפור איכות השמע עבור האלגוריתמים שונים בהם השתמשנו.

2.3. קשיים בפתרון הבעיה

הבעיה המרכזית היא מציאת, אימון והתאמת מודל למידה עמוקה על קבוצת הנתונים הרצויה. לאחר שבחרנו את המודל בו השתמשנו, עיקר ההתמודדות לאורך העבודה הייתה שיפור תוצאות האימון בדרכים שונות.
בעיה נוספת בה נתקלנו הייתה יצירת קבוצת הנתונים איתה עבדנו בפרויקט. הרעיון המקורי היה יצירת data set אשר מותאם לחדר ההקלטות הבעייתי בטכניון אך לאחר שהבנו שיצירת data set באופן עצמאי בעייתית עבורנו מסיבות של זמן ויכולת עברנו לאלטרנטיבה של עבודה עם data set מוכן אותו הצלחנו למצוא במספר מקורות שונים באינטרנט.

מושגי יסוד

3.1. הדהוד

או כפי שנהוג לקרוא למושג זה באנגלית, ריוורברציה reverberation (כמושג פיזיקלי) היא התמדתו של צליל זמן מה לאחר הפקתו. אינטראקציה של גלי קול עם משטחים פיזיים היוצרת צליל נוסף הנשמע בחלל לאחר שצליל המקור הפסיק להשמע. אלפי גלי קול שחוזרים בכיוונים ובזמנים שונים, לאחר שפגעו במשטחים שונים, נעים בחלל לאחר שצליל המקור הושמע ויוצרים הארכה של זמן שמיעת הצליל.

3.2. חזרות מוקדמות ומאוחרות

מרכיבות ומאפיינות הדהוד בחלל כלשהו. חזרות מוקדמות (early reflections) הן חזרות הקול הראשונות אשר מקפצות בחדר או בחלל בו אנחנו נמצאים. החזרות האלה מגיעות לאוזני המאזין תוך 30 עד 50 אלפיות השנייה לכל היותר מרגע היווצרות הצליל הישיר. חזרות אלו עוזרות ליצור תחושת מרחב ומימד בצליל והן יכולות לתרום לבהירות ומובנות של הצליל.
לעומת זאת, חזרות מאוחרות, הן החזרות אשר מקפצות בחלל לפרק זמן ארוך יותר ומגיעות לאוזני המאזין לאחר לפחות 50 אלפיות השנייה. חזרות אלו תורמות לאופי הכללי ולהמשכיות (sustain- אורך מסיום היווצרותו עד דעיכתו) הצליל ועוזרות ליצור תחושה של ריחוק ועומק בצליל.
ביחד, החזרות המוקדמות והמאוחרות יוצרות את הרכב פיזור הצליל המורכב שאותו אנו תופסים כהדהוד. ניתן לכוון ולתמרן את החזרות בדרכים שונות על מנת להשיג אפקטים קוליים שונים בהפקת מוזיקה והנדסת אודיו.

3.3. ייצוג אודיו בצורה דיגיטלית

אודיו מיוצג על ידי רצף של ערכים מספריים, שנדגמו על ידי מכשירי הקלטה או סונתזו באופן מלאכותי. לדגימת השמע והמרתו לרצף מספרים יש שני מאפיינים חשובים, מאפיינים אלו מאפשרים לאותו הרצף להיות מתורגם חזרה לצליל (על ידי מכשירי השמע השונים) ובנוסף למידע כמו מאפייני דחיסה וקוונטיזציה הם קובעים את הפורמט בו נשמר האודיו (למשל flac או WAV):

- תדירות הדגימה- (Sampling rate, נמדד לרוב ב-KHz) מציין את מספר הפעמים בשנייה שהצליל נדגם. ככל שהתדירות תהיה גבוהה יותר כך הצליל יהיה קרוב יותר למקור.
- רזולוציית הדגימה- (נמדד לרוב ב-Kbps) מציין את כמות המידע לשנייה. ככל שכמות המידע תהיה גדולה יותר, כך האיכות של הצליל תהיה גבוהה יותר.

3.4. ספקטרוגרמה

ספקטרוגרמה היא ייצוג חזותי של ספקטרום התדרים של האות והשתנותו עם הזמן. האות, במקרה שלנו הוא הצליל הדגום אשר מומר לתמונה המייצגת את שינוי עוצמת כל תדר לאורך פרק זמן מסוים.
ספקטרוגרמה נותנת ביטוי חזותי מובחן לאלמנטים שונים בצליל כגון הרמוניות בכפולות שלמות של תדר יסוד, חזרה על צליל, גובה הצליל או עוצמתו.
ישנן וריאציות רבות לתבניות של ספקטרוגרמות:
ספקטרוגרמה ראשונה היא ספקטרוגרמה בה ציר ה-Y מייצג את הזמן וציר ה-X מייצג את התדר, ספקטרוגרמה שנייה היא תרשים מפל מים בו המשרעת מיוצגת על ידי גובהו של משטח תלת מימדי. ישנן ספקטרוגרמות בהן צירי התדירות והמשרעת יכולים להיות לינאריים או לוגריתמיים.
אודיו יוצג לרוב עם ציר משרעת לוגריתמי (לרוב בדציבלים, או dB), והתדר יהיה ליניארי או לוגריתמי כאשר תדר ליניארי מדגיש יחסים הרמוניים ואילו תדר לוגריתמי מדגיש קשרים מוזיקליים וטונליים.
בנוסף, ישנה הספקטרוגרמה בה אנו השתמשנו, ספקטרוגרמה זו מיוצגת על ידי תמונה בה ציר ה-X מייצג את הזמן וציר ה-Y מייצג את התדר, בנוסף, הצבע מייצג את משרעת התדר באמצעות סקלה אשר מתחילה מכחול המייצג עוצמה נמוכה ועד אדום המייצג עוצמה גבוהה.
בעבודתנו השתמשנו בעיקר בייצוג של ההדהוד המוקלט הנראה בתור ה"זנבות"/"מריחות" בתמונה, כלומר, האזורים בהם עוברים מצבע אדום וחזק, המייצג תדרים המופק בהם צליל בוליום גבוה (דיבור), לצבע כחול כהה המייצג תדרים שקטים (הפסקת הדיבור).
יצירת ספקטרוגרמה נעשית על ידי תהליך של Short Time Fourier Transform- חלוקת ציר הזמן לקטעים, לרוב עם חפיפה ביניהם. על כל קטע אפשר להפעיל פונקציית חלון כך שעבור כל חלון מחשבים את המשרעות של התדרים בספקטרום על ידי התמרת פורייה והן מיוצגות כקו אנכי.
הקווים האנכיים, מוצבים זה לצד זה ויוצרים תמונה או משטח תלת ממדי, על פי תבנית ההצגה שנבחרה. אותם הפרמטרים מהווים חלק חשוב בהבנת וקריאת הספקטרוגרמה, וכן שחזור הצליל ממנה.
המאפיין העיקרי הוא גודל החלון אשר יכול להיות בטווח רחב של ערכים ולפיו נקבע מספר החלונות. חלון קצר יותר, ייתן תוצאות מדויקות יותר על ציר הזמן, על חשבון הדיוק בתדר, ולהפך. בחירת גודל החלון, מהווה פשרה בין דיוק בזמן לדיוק בתדר.

3.5. Mel scale

התדרים בספקטרוגרמה מוצגים בעזרת סקאלת מל, שמאפשרת לתדרים הנשמעים באוזן האנושית להיות מוצגים במרחק יחסי הזהה לאופן בו הם נשמעים. האוזן האנושית תופסת עוצמה של צלילים בצורה שאינה לינארית במרחק ובתדרים הנשמעים, ולכן על מנת להקל על הבנת מפת התדרים ביחס לשמיעה נשתמש בסקאלת מל.

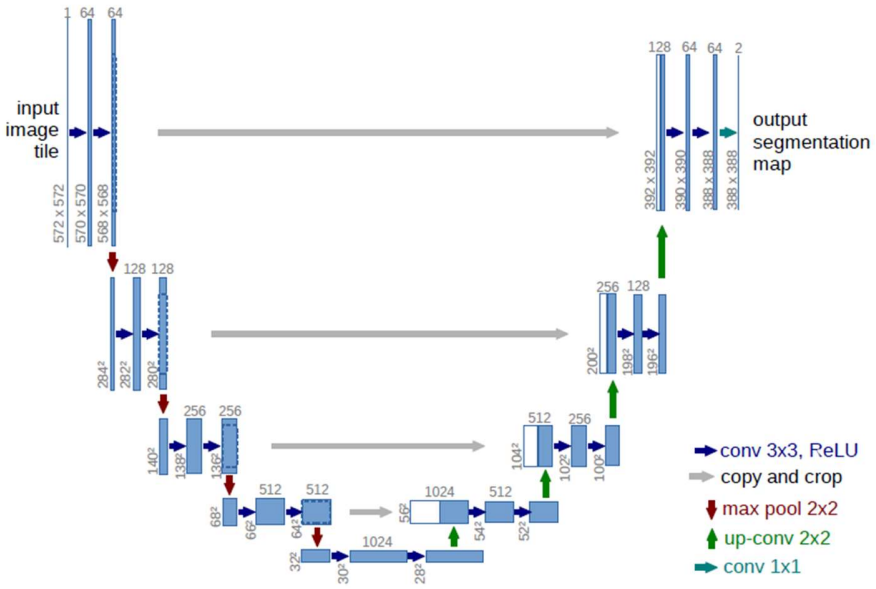
מודלים

U-Net .4.1

השימוש בארכיטקטורת U-Net על מנת להפריד את מרכיבי התמונה העיקריים ולנקות את השאר, יחד עם ההצגה של קטעי הקול על ידי ספקטרוגרמות, היוו את הבסיס לעבודה עליה ביססנו את הפרויקט בהמשך.
ה-U-Net היא ארכיטקטורת רשת עצבית קונבולוציונית עמוקה המיועדת לצורך סגמנטציה של תמונה ביו-רפואית, שהוצגה במאמר:
"U-Net: Convolutional Networks for Biomedical Image Segmentation" בשנת 2015.
ארכיטקטורת U-Net מורכבת משני נתיבים מרכזיים:

- <u>נתיב הצמצום-</u> מיישם סדרה של שכבות קונבולוציוניות ופעולות max pooling כדי לחלץ ולמפות מאפיינים ברמה גבוהה מתמונות הקלט (החלק היורד בתרשים המופיע מטה).
- <u>הנתיב המתרחב-</u> מופיע מיד לאחר נתיב הצמצום ומשתמש בשכבות קונבולוציוניות משוחלפות כדי לבצע upsample עם מיפוי המאפיינים ולשחזר את הרזולוציה המקורית של התמונה (החלק העולה בתרשים המופיע מטה).

בנוסף לשכבות הקונבולוציוניות הסטנדרטיות, ה-U-Net כולל גם דילוג בין השלבים השונים ובין הנתיבים השונים. חיבורים אלו מאפשרים לרשת לשמר מידע עם רמה נמוכה מתמונות הקלט, דבר שיכול להיות חשוב לסגמנטציה מדויקת, ושחזור תמונה מדויק (החיצים האפורים בתרשים).



4.2. LS U-Net

המאמר אשר מציג את המודל עליו ביססנו את עבודתינו מציג מספר רב של מודלים המשתמשים או מרחיבים באופן ישיר או עקיף את הארכיטקטורה הנ"ל. בחרנו לשפר את המודל הטוב ביותר לפי ההערכות והתוצאות שהוצגו במאמר ועל כן, המודל שבחרנו הוא Late Suppression U-Net. על פי המאמר מודל זה מבוסס ישירות על ארכיטקטורת ה U-Net, ומטרתו היא להחליש את ההדודים (רוורברציות מאוחרות), הנשמעים בקטעי קול.

מודל ה- LS U-net שונה ממודל ה- U-Net המקורי בכך שבארכיטקטורה זו נוסף skip connection בין הקלט (התמונה המקורית והמהודהדת) לבין השכבה האחרונה ברשת, וביצוע פעולת הסרה ביניהם.

הכנת הקלט לעבודה עם המודלים

5.1. עיבוד קטעי הקול

הדאטה סט עליו אומן המודל מתחלק לשתי קטגוריות מרכזיות:

- מידע מוקלט- קטעי שמע ודיבור שהוקלטו בחדרים שונים על ידי מספר מיקרופונים הממוקמים צמוד לדובר, על מנת לדגום הקלטה נקייה וברורה, ורחוק ממנו, על מנת לדגום הקלטה רוויה בהדהוד.
- מידע מסונתז- קטעי שמע נקיים ש"לוכלכו" באמצעות תהליך פשוט של קונבולוציה עם תגובות הלם של חדרים (מידע שנשמר בצורה של אודיו, המתאר את מאפייני ההדהוד של חדר מסוים), ומכך התקבלו הקלטות מהודהדות. בשיטה זו מתאפשר לקבל אופי הדהוד שונה במקצת מזה של ההקלטה הטבעית, ואף לשלוט בזמן ועוצמת ההדהוד.

בעזרת ריבוי של סוגי הנתונים, נוכל לקבל מגוון רחב של הדהודים שונים אותם נרצה לאמן את המודל, לזהות ולנקות.

דגימות השמע היו בקצב דגימה של kHz16.

5.2. מעבר לספקטרוגרמות

הפרמטרים לטרנספורמציה היו:

גודל החלון- 2048 דגימות

מרחק הקפיצה לחלון הבא- 512 דגימות

מספר הפריימים אותם דוגמים- 340

תדר מינימלי שמוצג- Hz20

תדר מקסימלי שמוצג- Hz8300

5.3. הערכת איכות המודלים

על מנת להעריך את איכות הספקטרוגרמות החדשות שפולט המודל, השתמשנו במספר מטריקות מוכרות עבור השוואת סיגנלים ובדיקת איכות שמע ודיבור:

STOI: Short Term Objective Intelligibility

CD: Cepstral Distance

LLR: Log-Likelihood Ratio

fwSNRseg: Frequency Weighted Segmental SNR

SRMR: Speech to Reverberation Modulation Energy Ratio

כאשר ארבעת המטריקות הראשונות עובדות על עיקרון של השוואת הסיגנל המקורי עם הסיגנל המנוקה, ומביאות הערכה במונחים של קרבה/דמיון בין הסיגנלים.
לעומתם, המטריקה האחרונה שמה דגש על דיבור ורהיטות, ונותנת הערכה, על ידי דימוי והבנה אקוסטית של האוזן האנושית, לאיכות הדיבור והסיגנל כשלעצמו.

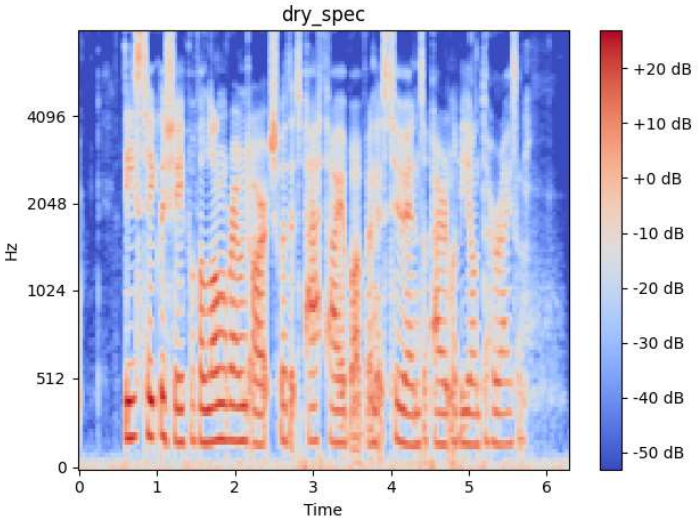
תהליך העבודה

בחלק זה אנו מסבירים על הקלטים והתוצרים של המודלים לאחר אימונם. במהלך חלק זה נציג ספקטרוגרמות וקבצי שמע אשר מייצגים את אותו קטע קול עם שינויים שונים, כגון, הוספת הדהוד או רעשי רקע.

(עבור כל שלב בעבודה עליו אנחנו מפרטים בחלק זה, הוספנו בקובץ הZIP את קובץ השמע הרלוונטי לאותו שלב. שמות קבצי השמע הרלוונטיים יצויינו בטקסט ירוק בסמוך לספקטוגרמה המתאימה לו).

ספקטרוגרמה של קטע הקול המקורי (דיבור נקי ללא הדהוד או רעשי רקע):

(קטע קול מקורי original_sound.wav)



6.1. המודל הראשון- ניקוי הדהוד

6.1.1. תהליך האימון ושגיאת ה-LOSS

הגדרת הבעיה הראשונית:

$x[t]$- סינגל מקורי

$y[t]$- סינגל מהודהד

$h[t]$- פונקציית הקונבולוציה המהדהדת את הסינגל בעזרת תגובת הלם

$eta[t]$- רעש הנוסף לסינגל המהודהד

$y[t] = (x * h)[t] + eta[t]$

(כאשר $*$ היא פעולת הקונבולוציה)

על מנת לשים את הדגש על ההדהודים המאוחרים, נרחיב את הבעיה:

$$y[t] = \left(x * h_{early}\right)[t] + (x * h_{late})[t] + eta[t] = y_{early}[t] + y_{late}[t]$$

נכונות המודל מבוססת על כך שהסיגנל המקורי מופיע גם הוא בסיגנל המהודהד, ולכן דילוג או ויתור על הפרדת ההדהודים המוקדמים המופיעים בסמוך לסיגנל המקורי, יעזור למודל לשים את הדגש על ההדהודים המאוחרים ובכך לנקות בצורה טובה יותר את הספקטרוגרמה.
סוג השגיאה המוגדרת לאימון המודל על הספקטרוגרמות היא MSE שהיא טעות ריבועית ממוצעת בין שתי התמונות.

6.1.2. הסבר כללי

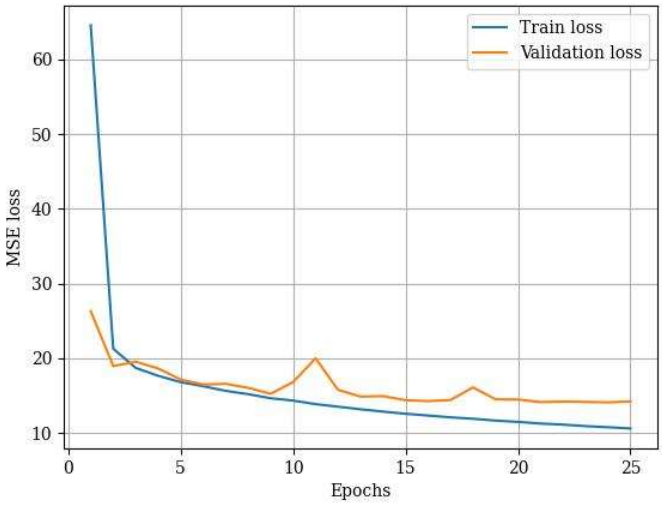
המודל אומן על 20,000 קטעי קול שונים, נקיים ומלוכלכים, הנלקחו מ-3 סיפריות שונות.
פרמטרים:

loss function	MSE
optimizer	ADAM
learning rate	0.0002
beta 1	0.5
beta 2	0.999
learning rate decay	0.97
decay rate	2

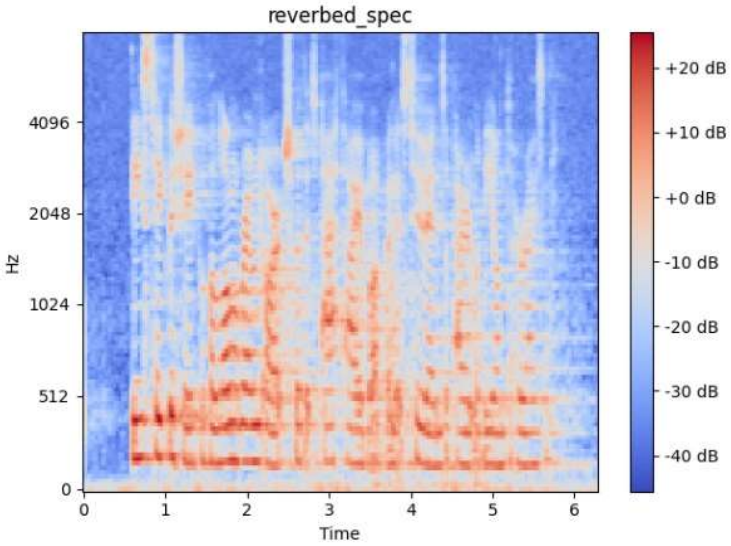
כאשר מתקיים העדכון הבא עבור האופטימייזר:

```
if (epoch % decay_rate == 1):
    optimizer.param_groups[0]['lr'] *= lr_decay
```

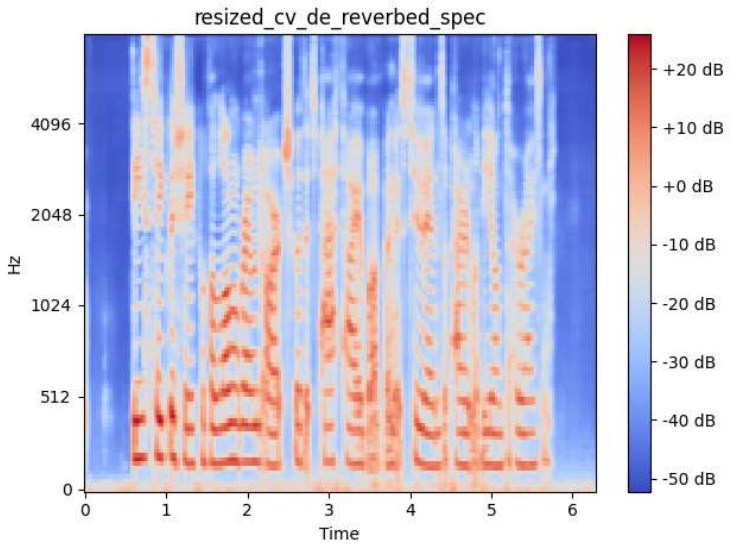
6.1.3. תוצאות האימון



לאחר כ-25 אפוקים קיבלנו שגיאת אימון של 10.794 ושגיאת אימות של 14.476.
פלטים לדוגמה:
ייצוג קטע הקול המהודהד הנכנס כקלט המודל:
(קטע קול אליו התווסף הדהוד Reverbed_6_1.wav)



ייצוג קטע הקול אותו פלט המודל:
(קטע הקול הקודם לאחר שהמודל ניקה ממנו את ההדהוד DeReverbed_6_1_3.wav)



6.1.4. תוצאות הבדיקה EVALUATION

<u>Reverberant signal:</u>

STOI: 0.75
LLR: 0.9
CD: 5.65
fwSNRseg: 6.78
SRMR: 3.13

<u>Dereverberated signal:</u>

STOI: 0.85
LLR: 0.44
CD: 3.67
fwSNRseg: 9.99
SRMR: 6.07

6.2. המודל השני- ניקוי רעשים

החלטנו לנסות לשפר את המודל הראשון, על ידי הוספת התכונה ניקוי רעש הנוצר ממספר גורמים שונים כמו איכות המיקרופון, דיבור של דוברים נוספים ברקע או רעשים סטטים בחלל (מזגן, מאוררים וכדומה).
תחילה חקרנו במקורות שונים כיצד לנקות רעשים מסוגים שונים מקטעי קול ומצאנו כי רשת ה-U-Net משמשת גם עבור פתרון בעיה זו.
שימוש במודל ה-U-Net הפשוט, ללא דילוג המבצע שינוי בפלט ומסיר את החלקים המרווחים בתמונה שהם ההדהודים שהיו הבעיה הקודמת.
כעת, אנו מעוניינים במודל המבצע סגמנטציה לחלקי הדיבור המרכזיים אל מול רעשי הרקע ולכן מודל זה יעבוד עבור מטרה זאת.

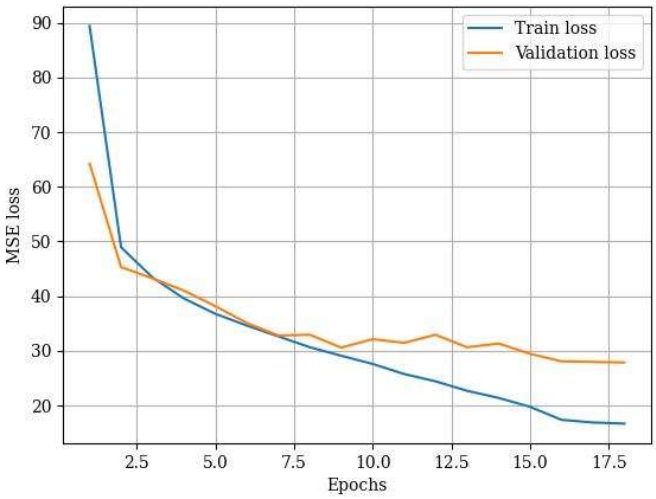
6.2.1. שינוי בקבוצת הנתונים

על מנת לאמן את המודל על התכונה החדשה, השתמשנו בקטעי קול שונים מאלו שהשתמשנו בהם לאימון המודל הראשון, בעלי אותם מאפיינים כמו אורך, עוצמה וקצב דגימה.
קטעי הקול ה"מלוכלכים" היו מעט מגוונים, כלומר, תצורת הלכלוך הייתה שונה.
בקבוצת הנתונים בה השתמשנו לחלק מהקטעים התווספו קולות רקע של שיחה בין אנשים, לחלק התווספו רעשים סטטים כמו מים זורמים, ולחלק התווסף רעש לבן אחיד בעוצמה נמוכה יחסית.

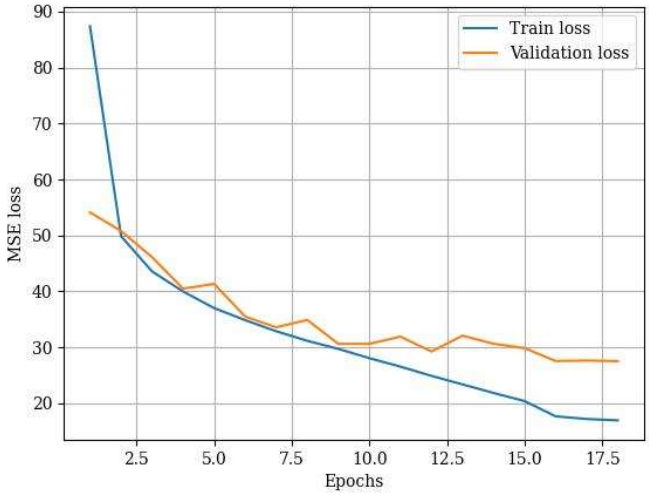
6.2.2. אימון המודל

שגיאת ה Loss זהה לזאת של המודל הראשון, ותהליך האימון בוצע באותו האופן, אך קיבלנו תוצאות פחות טובות ושגיאות אימון וולידציה גבוהות משל המודל הראשון. על מנת לשפר את איכות המודל, בחנו מספר פרמטרים באימון המודל.

6.2.3. תוצאות האימון

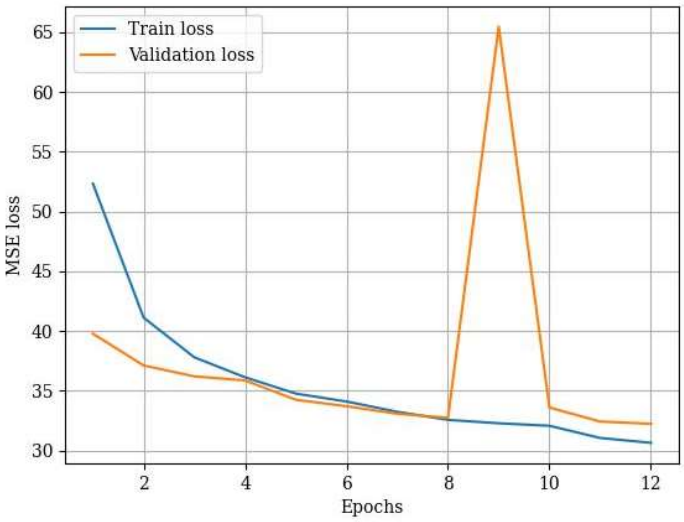


לאחר שראינו כי תוצאת האימון הראשונה לא הייתה דומה בטיבה לזו של המודל הראשון, החלטנו לבחון שינויים של ההיפר-פרמטרים של אימון המודל הנוכחי. לאחר מספר הרצות ובדיקות, ועל ידי בחינת עקומת הלמידה, החלטנו לאתחל את תהליך האימון עם קצב למידה של 0.002, ולשנותו לאחר כ-15 אפוקים להיות 0.0002.
תוצאת האימון הטובה ביותר שהשגנו עבור קבוצת הנתונים הנוכחית לאחר 18 אפוקים הייתה שגיאת אימון של 18.044 ושגיאת אימות של 28.895.



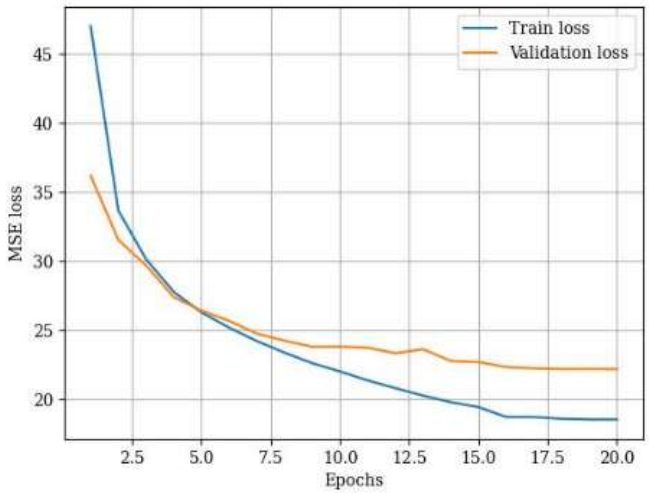
לאחר מכן החלטנו לנסות את האפשרות של הגדלת קבוצת הנתונים איתה המודל מתאמן, משום שמספר קטעי הקול איתם אימנו את המודל השני היה 13475, לעומת 20,000 שבו אומן המודל הראשון על רשת דומה. תחילה בחנו את האפשרות של הוספת רעש לבן לקטעי הקול, אך נתקלנו במספר קשיים שלאחר מכן אף סייעו לנו ללמוד ולתחום בצורה טובה יותר את גבול יכולת המודל.
תוצאת האימון עבור קטעי קול שהתווסף להם רעש לבן היו שגיאת אימון של 30.943 ושגיאת אימות של 32.792.
לאחר בחינה של פלט המודל על קטעי הקול שהוספנו להם רעש לבן, גילינו כי ניקוי רעש לבן היא משימה מורכבת משתי סיבות:

- רעש לבן חולש על כל תחום התדרים הנשמע, ובפרט אלו המיוצגים בספקטרוגרמות עבור קלטי המודל, בשונה ממשימת האימון הקודמת על רעשי רקע רגעיים/מוגבלים בתדר.
- רעש רקע מתמשך מופיע לכל אורך ההקלטה ובעצם מצמצם את השוני בין ייצוג קטע הדיבור לרגעים השקטים בהקלטה, מה שמקשה מאוד על מודל סגמנטציה כמו שלנו למתוח את הגבול בין הדיבור לשקט בספקטרוגרמה.

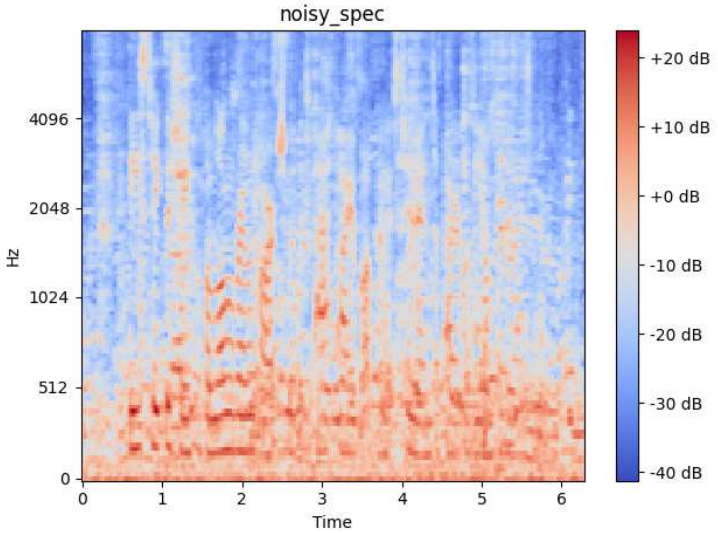


בחרנו להסיר את עדכון משקולות המודל עבור האימון עם הרעש הלבן, ולייצר קבוצת נתונים משלנו אשר דומה לקבוצת הנתונים הראשונה: השתמשנו בספרייה המכילה 10,000 קטעי דיבור על ידי דוברים שונים, ובנוסף השתמשנו בספרייה של רעשי רקע שונים (מסעדה רועשת, מכוניות חולפות, נביחות כלבים וכו'). הוספנו את רעשי הרקע לקטעי הקול בעוצמה נמוכה ובצורה רנדומלי כך שלבסוף קיבלנו סט נוסף של קטעי

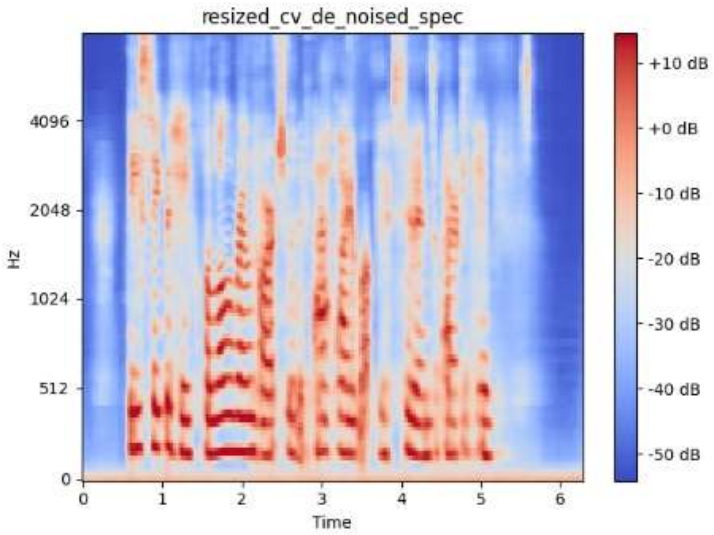
דיבור עם רעשי רקע מגוונים בעוצמה דומה לזו של קטעי הקול מסט האימון הראשון של מודל זה. ביצענו את האימון לאחר טעינת משקולות האימון הראשון ולאחר 20 אפוקים קיבלנו שגיאת אימון של 17.424 ושגיאת אימות של 20.355.



פלטים לדוגמה:
ייצוג קטע הקול לאחר הוספת רעש רקע (בדוגמה זאת רעש רקע של מסעדה):
(קטע קול אליו התווסף לכלוך Noisy_6_2_1.wav)



ייצוג קטע הקול אותו פלט המודל:
(קטע קול הקודם לאחר שהמודל ניקה ממנו את הרעש המוסף
DeNoised_6_2_3.wav)



6.2.4. תוצאות הבדיקה EVALUATION

<u>Reverberant signal:</u>

0.06 :STOI
0.1 :LLR
1.84 :CD
11.41 :fwSNRseg
0.21 :SRMR

<u>Dereverberated signal:</u>

0.03 :STOI
0.1 :LLR
1.84 :CD
11.37 :fwSNRseg
0.24 :SRMR

6.3. המודל השלישי- המודל הסופי

המודל הסופי מורכב משני המודלים הקודמים עליהם פרטנו בסעיפים 6.1 ו-6.2.

6.3.1. הרכב המודל

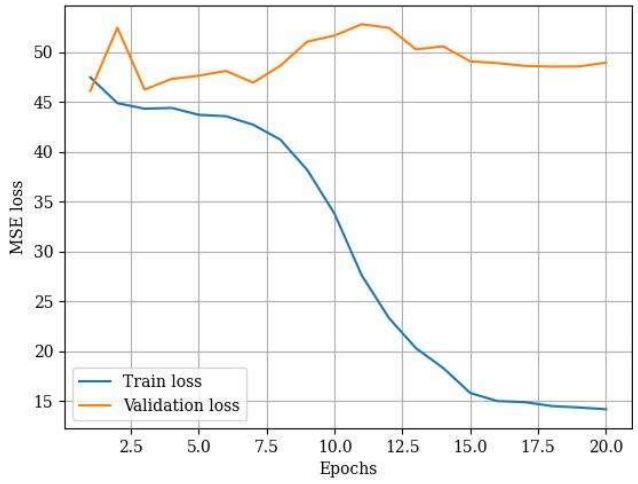
על מנת להשתמש בשני המודלים ולהציג שיפור אל מול החלקים הנפרדים, בנינו מודל סופי משרשור של שני המודלים הקודמים. בחרנו להסיר תחילה את ההדהוד ולאחר מכן את להסיר את הרעש, מתוך אינטואיציה לכך שניקוי ה"מריחות" מהספקטרוגרמה תקל על החלק של זיהוי וניקוי רעשי הרקע המוצגים בה.

6.3.2. שינוי בקבוצת הנתונים

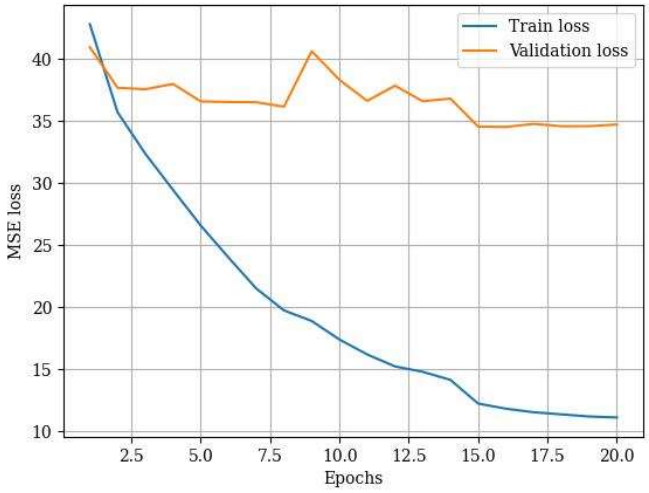
גם עבור מודל זה השתמשנו בקטעי קול שונים מאלו שהשתמשנו בהם לאימון המודלים הקודמים, בעלי אותם מאפיינים כמו אורך, עוצמה וקצב דגימה.
כיוון שלא הצלחנו למצוא קבוצת נתונים של קטעי דיבור מלוכלכים ומהודהדים יחדיו, החלטנו ליצור אותה בעצמנו, על ידי הוספת הדהוד לקטעי דיבור עם רעש רקע.
הוספנו את ההדהוד על ידי פעולת קונבולוציה פשוטה בין קטעי הקול הרועשים, לתגובות הלם שונות, וחיתוך הקטעים לאורך אותו מצפה המודל לקבל.

6.3.3. אימון המודל

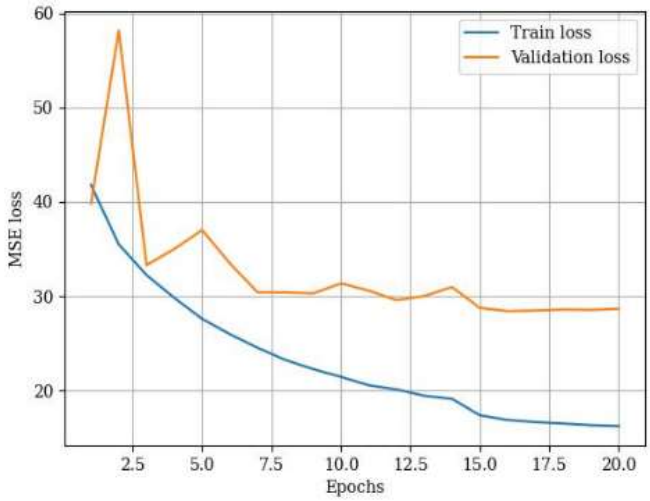
בדומה לשוני בין המודל הראשון לשני, גם כאן נתקלנו בקבלת תוצאות פחות טובות מאשר אלו שנראו עבור שני המודלים הקודמים.
ציפינו לכך שמשימת המודל הסופי תהיה מורכבת וקשה יותר מאשר המודלים האחרים מהם הוא מורכב, אך יחד עם זאת ציפינו לתוצאות טובות יותר מאלו שקיבלנו ולכן ניסינו לשפר את תוצאות האימון ואיכות המודל.
בהתחלה, ניסינו להרכיב את המודל הסופי כך שיבצע הסרה של רעש הרקע ולאחר מכן יסיר את ההדהוד, וקיבלנו את גרף התוצאות הבא:



לאחר מכן, ביצענו את אותו אימון זהה עם הבדל יחיד לפיו המודל מסיר תחילה את ההדהוד ולאחר מכן את רעש הרקע, וקיבלנו את גרף התוצאות הבא:



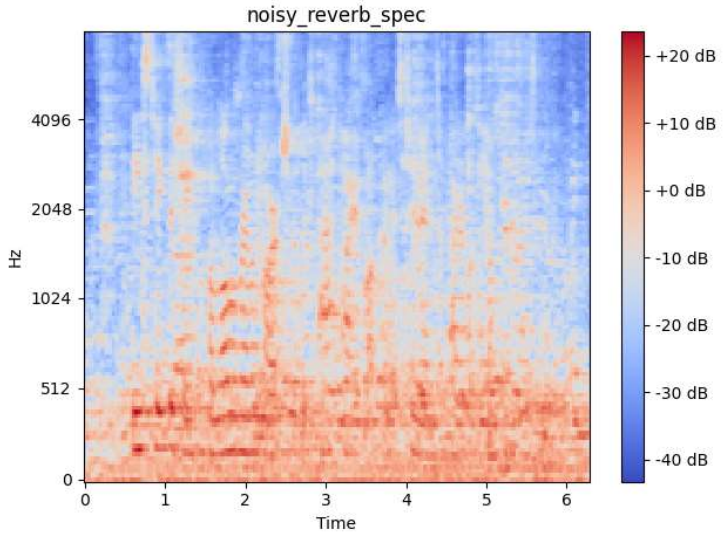
כלומר, כפי שציפינו, הרכבת המודל הסופי בתצורה השנייה הניבה תוצאות בסיס טובות יותר, ולכן החלטנו להמשיך עם התצורה הזאת, ולנסות לשפר את תוצאות האימון. לאחר טעינת המשקולות המעודכנות מהאימון האחרון, ביצענו סבב אימון נוסף על קבוצת נתונים נוספת שיצרנו, של קטעי דיבור רועשים ומהודהדים. קיבלנו לאחר 20 אפוקים שגיאת אימון של 24.692 ושגיאת אימות של 35.468.



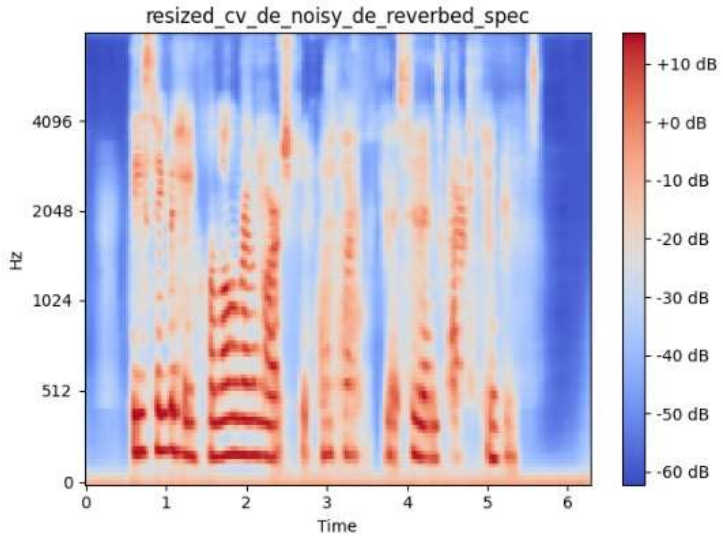
פלטים לדוגמה:

ייצוג קטע הקול לאחר הוספת רעש רקע (רעש רקע של מסעדה + הדהוד):

(קטע קול אליו הוספנו רעש ולאחר מכן הדהוד NoisyReverbed_6_3_2.wav)



ייצוג קטע הקול אותו פלט המודל:
(קטע הקול הקודם לאחר שהמודל ניקה ממנו את ההדהוד ואת הרעש
(De_noisy_de_reverbed_6_3_3.wav



6.3.4. תוצאות הבדיקה EVALUATION

<u>Reverberant signal:</u>
0.46 :STOI
0.13 :LLR
2.05 :CD
10.79 :fwSNRseg
1.32 :SRMR

<u>Dereverberated signal:</u>
0.45 :STOI
0.13 :LLR
2.08 :CD
10.73 :fwSNRseg
1.36 :SRMR

6.3.5. סקר איכות המודל

על מנת לבדוק באופן אובייקטיבי את איכות המודל הסופי, ביצענו סקר double blind בו השמענו קטע מוקלט של דיבור נקי ולאחריו השמענו קטעי קול עם תוספת רעש או הדהוד או שניהם על קטע הקול הנקי ולאחר מכן השמענו קטעי קול לאחר הניקוי שביצעו המודלים השונים. ביקשנו מהמאזינים לדרג את מידת הדמיון בין הקטע המקורי לבין כל אחד מהקטעים האחרים.

<u>תוצאות הסקר-</u>

תוצאות הסקר הראו כי המשיבים על הסקר חושבים כי קטע הקול עם תוספת הרעש וההדהוד הוא קטע הקול שהכי לא דומה לקטע הקול הנקי. בנוסף, ניתן לראות כי המשיבים חושבים שקטע הקול עם תוספת ההדהוד יותר דומה לקטע הנקי מאשר קטע הקול עם תוספת הרעש.

תוצאות הסקר בנוסף לכך משקפות שהמשיבים על הסקר חושבים שקטע הקול אשר נוקה על ידי המודל השלישי (הסרת רעש והדהוד) הוא קטע הקול הדומה ביותר לקטע הקול הנקי ובאופן מפתיע המשיבים חושבים שקטע הקול אשר נוקה על ידי המודל הראשון (הסרת הדהוד) דומה יותר לקטע הקול הנקי מאשר קטע הקול אשר נוקה על ידי המודל השני (הסרת רעש) וזאת על אף שחשבו כי קטע הקול עם תוספת ההדהוד יותר דומה לקטע הנקי מאשר קטע הקול עם תוספת הרעש.

מכך אנו למדים כי המודל שיצרנו עונה על הציפיות שלנו והוא אכן משפר את שני המודלים האחרים איתם עבדנו בעבודתנו.

סיכום

בפרויקט זה חקרנו את אחד מהמקורות המרכזיים של איכות הקלטה ירודה- הדהוד והחזרי קול המרצה המוקלטים, בנוסף, הצגנו מודל למידה עמוקה שמטרתו לנקות קטעי קול מהודהדים ושיפור יכולת ההבנה של הדובר.
בתחילה, גיששנו ובדקנו מספר סוגים שונים של גישות ומודלי למידה, ולבסוף לאחר מחקר ובדיקה של הנושא החלטנו להשתמש ברשת מסוג U-Net לסגמנטציה של תמונות, ולהפעיל אותה על ספקטרוגרמות של קטעי קול. השתמשנו במימוש הקיים למודל, והבחנו כי קטעי הקול הנפלטים אכן נוקו מההדהוד ונשמעו ברורים יותר.
החלטנו להרחיב את המשימה ולהביא לשיפור המודל ולהוסיף את האפשרות לניקוי רעשי רקע מקטעי קול.
תחילה חשבנו לנסות לבצע fine-tuning על המודל הקיים, אך לאחר מחקר בנושא, הבנו כי משימת הסגמנטציה הספציפית של המודל הקיים אינה מתאימה, והחלטנו לבצע אנסמבל ולבנות מודל המבצע שרשור של 2 מודלים כך שלכל אחד מהם משימה שונה: הראשון ניקוי ההדהודים והשני ניקוי רעשי הרקע.
המודל הראשון שמטרתו ניקוי ההדהוד נעשה כפי שהוצג במאמר, והמודל השני שמטרתו ניקוי רעשי רקע, הוא מודל ה U-Net המקורי.
ביצענו מספר איטרציות של אימון עבור המודל השני, עד שהגענו לתוצאות אימון מספקות, ולאחר מכן ביצענו אימון על מודל האנסמבל אשר בדומה לקודמו דרש מספר איטרציות, אופטימיזציות ושינויים עד שהגיע לתוצאות מספקות.
לבסוף, ביצענו הערכה של כלל המודלים על סוגי הפלטים המתאימים להם והשוואנו עם תוצאות הערכה הנראו במאמר, ביצענו סקר קצר על מנת לבדוק את טיב השימוש "האמיתי" של המודל הסופי ואת איכות ניקוי הרעשים וההדהוד לפי בדיקה סובייקטיבית.

שיפור העבודה והמשך מחקר

קיימים מספר כיוונים שונים בהם ניתן לצעוד על מנת לשפר את תוצאות המודל הסופי בעזרת ההרכב הנוכחי שלו.
מספר רעיונות אפשריים לשיפור:

- לשפר את איכות המודל הראשון והשני המרכיבים את מודל האנסמבל, על ידי איטרציות נוספות של אימון.
- להגדיל את קבוצות הנתונים השונות ולקבל הכללה טובה יותר עבור המודלים.
- לבצע שינוי בהרכב המודל הסופי, כמו הוספת שכבות ייחודיות, או קישור המודלים המרכיבים אותו בצורה שונה משרשור.

ועוד.
בנוסף, למודלים הנוכחיים, כיוון נוסף לשיפור העבודה למטרת ניקוי הדהוד ורעשי רקע הוא שינוי המודלים מבסיסם, כלומר, התנסות עם מודלים נוספים העוסקים בסגמנטציה, או בפעולה אחרת אשר יוכלו לקבל ספקטרוגרמות כקלט, ולפלוט אותם לאחר ניקוי מרעש או הדהוד. לדוגמא, החלפת המודל השני מארכיטקטורת U-Net, לארכיטקטורה שונה המבצעת סגמנטציה- SegFormer, DeepLab או Mask R-CNN.

כיוון מחקר נוסף הוא שימוש במודל אותו בנינו עבור משימות שונות בעיבוד אודיו בעזרת ספקטרוגרמה.
לדוגמה:

- ביצוע פילטור לקטעי קול, על ידי "החלשת/TONE DOWN" של חלקי הדיבור בספקטרוגרמה בתדרים מסויימים (טווח Y בתמונה).
- De-Essing: החלשת האותיות/ הברות הצורמות הנשמעות בקטע הקול שהדגש שלהן הוא בתדרים הגבוהים (ס, צ, ש, ט/ ת) על ידי מציאת חלקי דיבור הנמצאים בחלק העליון של הספקטרוגרמה והחלשתם במידה והם מיוצגים בעוצמה גבוהה (איזו שהיא עוצמת סף).
- הפרדת דוברים במידה והם עומדים בתנאים מסוימים כמו למשל הקלטה המורכבת משיחה בה לדוברים יש הפרדה ברורה בטווח התדרים של הדיבור (גבר ואישה, אמא ובת וכו').

מקורות

1. D. Le´on and F. A. Tobar, “Late reverberation suppression using u-nets,” 2021.
2. Repo: https://github.com/DiegoLeon96/Neural-Speech-Dereverberation
3. Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “Unet: Convolutional networks for biomedical image segmentation,” in Proc. of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015
4. Kun Han, Yuxuan Wang, DeLiang Wang, William S Woods, Ivo Merks, and Tao Zhang, “Learning spectral mapping for speech dereverberation and denoising,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 6, pp. 982–992, 2015.

קישור לקוד המקור של הפרויקט:

https://github.com/dogomen11/AIProject.git