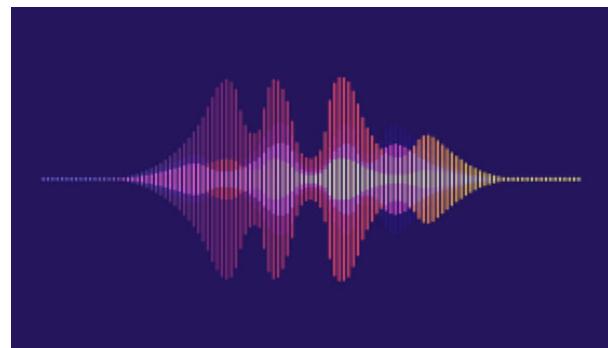


המכללה האקדמית תל חי

הפקולטה למדעי המחשב



דו"ח סיום פרויקט



דוגמאות אדוורסריאליות על מודל של הפרדת ערוצים: יצרה והגנה

מבצעים: נגה ענבי

איגור חמליק

מנחה: אורן ברית

תוכן העניינים

2.....	תוכן העניינים
3.....	רשימת אירורים
4.....	תקציר.....
4.....	Abstract
5.....	1. מבוא
5.....	1.1 מוטיבציה
5.....	1.2 סיכון התרומה העיקרית
6.....	1.3 עבודות קודמות בנושא
6.....	1.4 מבנה העבודה
7.....	2. סקר ספרות.....
7.....	2.1 SPLEETER -
9.....	2.2 DEMUCS -
9.....	2.2.1 DEMUCS V2 -
11.....	2.2.2 DEMUCS V4 -
13.....	2.3 בחרית (ht demucs)
13.....	3. מטריקות הערכתה
14.....	3.1 SDR
15.....	3.2 SIR
15.....	3.3 SAR
17.....	4. סוג התקפות
17.....	4.1 התקפות הרעה
18.....	4.2 התקפות פרטיות
18.....	4.3 התקפות ניצול לרעה
18.....	4.4 התקפות התchmodקот - Evasion Attacks
19.....	4.5 התקפה קופסה לבנה
19.....	4.5.1 התקפה מכוננת
20.....	4.5.2 התקפה לא מכוננת
20.....	4.6 התקפה קופסה שחורה
20.....	4.6.1 Surrogate Model
21.....	4.6.2 אופטימיזציה מבוססת חיפוש - Chilfi -
21.....	Query-based Optimization
21.....	1. מטפס הרים - Hill Climbing
21.....	2. שיטת מונטה קרלו - Monte Carlo Sampling
22.....	3. אלגוריתמים אבולוציוניים - Evolutionary Algorithms
22.....	Score-based or Decision-based
23.....	4.6.3 שיטות ללא תוויות -
25.....	5.ימוש התקפת קופסה לבנה בפרוייקט
26.....	6. אלגוריתם ההגנה
26.....	MUSDB18
27.....	8. רשימת מקורות
27.....	8.1 מאמריהם
28.....	8.2 מקורות נוספים

28	9. תוצאות
28	9.1 הסבר על הגרפים
29	9.2 תוצאות
36	10 פרק סיכום וסיכום
36	10.1 ניתוח תוצאות
37	10.2 מסקנות
37	תודות

רשימת איורים

[איור 1 - טבלה של ביצועי ספליטיר](#)

[איור 2 - דוגמה לספקטrogramma](#)

[איור 3 - תרשימים ארכיטקטורת דימוקס גרסה 2](#)

[איור 4 - טבלת השוואת ביצועי SDR של מודלים להפרדת ערוצים](#)

[איור 5 - תרשימים ארכיטקטורת דימוקס גרסה 4](#)

[איור 6 - טבלת השוואת ביצועי SDR של מודלים להפרדת ערוצים מול הגרסה ההיברידית החדשה של דימוקס](#)

[איור 7 - תרשימים של דרך הפעלת הדוגמה האדוורסיאלית על המודול](#)

[איור 8 - הצגת ביצועי SAR,SIR,SDR של המקור, ההתקפה וההגנה של שיר 1](#)

[איור 9 - הצגת ביצועי SAR,SIR,SDR של המקור, ההתקפה וההגנה של שיר 2](#)

[איור 10 - הצגת ביצועי SAR,SIR,SDR של המקור, ההתקפה וההגנה של שיר 3](#)

[איור 11 - הצגת ביצועי SAR,SIR,SDR של המקור, ההתקפה וההגנה של שיר 4](#)

[איור 12 - הצגת ביצועי SAR,SIR,SDR של המקור, ההתקפה וההגנה של שיר 5](#)

[איור 13 - הצגת ביצועי SAR,SIR,SDR של המקור, ההתקפה וההגנה של שיר 6](#)

[איור 14 - הצגת ביצועי SAR,SIR,SDR של המקור, ההתקפה וההגנה של שיר 7](#)

תקציר

דוגמאות אדוורסריאליות הן סוג של התקפות צדניות על מודלים של למידה عمוקה. התקפות אלו מבוצעות על-ידי שינוי מינימלי של הקלט – שינוי שאינו מורגש על-ידי בני אדם – אך גורם למודל להפיק פלט שונה או שונה לעומת הפלט המקורי. תוקף יכול לנצל מגנון זה כדי לגרום למערכת לבצע פעולות לא נכונות או לקבל החלטות שגויות.

במסגרת פרויקט זה, נבחן האם ניתן לבצע התקפת התחמקות (Evasion Attack) על מודל חדש מסוג Hybrid Transformer Demucs (גרסה 4), אשר נועד להפרדת ערוצים מוזיקליים (כגון שירה, תופים, באס ולויי) מהתוך קובץ שמע מעורב.

התקפה זו בוצעה בסוגנון של קופסה לבנה, כלומר, הייתה גישה מלאה לארכיטקטורת המודל, למשקליו ולקודו.

בנוסף, פיתחנו מגנון הגנה אפקטיבי המוסיף רעש רנדומלי בלתי נשמע לקלט המותקף.

ניסויים על בסיס הנתונים MUSDB18 הראו כי מגנון ההגנה מצליח לנטרל את ההשפעה של ההתקפה ולשפר את איצות הפרדה של המודל לאחר ההתקפה.

הערכת איצות התוצאות נעשתה באמצעות שלוש מטריקות מקובלות בתחום: SIR, SAR, ו-SDR, אשר מאפשרים לבדוק את מידת השיבוש שנגרמה למודל ואת הצלחת ההגנה בתמודדות עמה.

Abstract

Adversarial examples are a type of malicious attack targeting machine learning models, particularly deep learning systems. These attacks involve applying small, imperceptible perturbations to the input data, which can cause the model to produce incorrect or unintended outputs. An attacker can exploit this vulnerability to manipulate the system's behavior in subtle yet potentially harmful ways.

In this project, we investigate whether it is possible to carry out an evasion attack on a state-of-the-art music source separation model, specifically Hybrid Transformer Demucs (version 4). This model is designed to isolate individual musical components (e.g., vocals, drums, bass, and accompaniment) from a mixed audio track. The attack was performed in a white-box setting, meaning the attacker had full access to the model's architecture and weights, and used gradient-based optimization to craft adversarial perturbations that degrade the model's separation quality.

To counter the attack, we developed an effective defense mechanism that adds random, imperceptible noise to the adversarial input. This method successfully disrupts the carefully crafted perturbations, mitigating the impact of the attack and restoring the model's separation performance.

We evaluated the effectiveness of both the attack and the defense using three standard metrics for source separation: SDR (Source-to-Distortion Ratio), SIR (Signal-to-Interference Ratio), and SAR (Signal-to-Artifacts Ratio). Experiments conducted on the MUSDB18 dataset demonstrate

the vulnerability of the model to adversarial perturbations, as well as the potential of lightweight defenses to significantly reduce their effect.

1. מבוא

1.1 מוטיבציה

בשנים האחרונות חלה התפתחות מואצת בתחום הבינה המלאכותית, במיוחד במשימות כמו זיהוי תמונה, זיהוי קול, והבנה של שפה טבעית. לדוגמה, מודלים מתקדמים כמו ChatGPT הפקו לחלק בלתי נפרד מח'ינו, ונעשה בהם שימוש נרחב בתעשייה מגוונת, כולל תחומיים קריטיים כגון ביטחון, רפואי ותחבורה.

למרות הרושם ש모ודלים הללו ולעתים אף יכולים לעלות על ביצועים אנושיים הם עדין סובלים מפגעות חמורות לתקיפות אדוורסריאליות. תקיפות אלו הולכות ונעשה מתחכם יותר, ומהוות איום ממש על אמינות ובטיחות מערכות הבינה המלאכותית [1].

השלכות של תקיפות כאלה עלולות להיות חמורות וękטלניות, במיוחד כאשר מדובר במערכות הפעולות בסביבות רגישות כמו כל אוטונומיים, מערכות רפואיות מאובচנות, או מערכות הגנה ובקרה צבאית. מסיבה זו, חשוב לחקור ולהבין את אופי התקיפות הללו, לפתח דרכי לזייהו והתגוננות ולפתח התקיפות חדשות לטובת הגנה בלבד.

מחקרים נעשו בתחום התקיפות אדוורסריאליות על מודלים של ראייה ממוחשבת אך מעט יחסית נעשו על מודלים של עיבוד אותות. לכן, יש עניין גדול במחקר מעמיק בנושא. יש לבדוק האם אוטם אלגוריתמיים לפתרונות הגנה נגד מתקפות למודלים לשיווג תמונה, יעלים גם למודלים להפרדת ערוצים של ערוצי שמע.

2. סיכום תרומה העיקרית

בפרויקט זה, בוצעה הדגמה של תקיפה אדוורסריאלית מסווג התקומקו (Evasion Attack) על מודל מתקדם להפרדת מקורות מזיקליים – v4 Hybrid Transformer Demucs. התקיפה יושמה במסגרת תರחיש קופסה לבנה, תוך שימוש בגרדיינט של פונקציית ההפסד ביחס לקולט לצורכי יצירת שיבוש אודיו אך אפקטיבי, שלא ניתן לשמוע על ידי אוזן אנושית, אשר גרם לירידה משמעותית באיכות ההפרדה של המודל.

בנוסף, הוצג מנגן הגנה המבוסס על הוספת רעש אקריא מוגבל, אשר הוערך ונמצא שיפור משמעותי את ביצועי המודל לאחר התקפה.

העבודה מהוות תרומה מחקרית חדשה, שכן מרבית המחקרים בתחום התקיפות האדוורסריאליות התקמדו עד כה במודלים לשיווג תמונה, בעוד שבפרויקט זה הוגדרה תקיפה מסווגה על מודל הפועל בתחום האודיו ובמשימת

רגישה רבת-ערוצים. בכך מתחפות כיוון חדש למחקר חולשות של מודלים לעיבוד אותן, כמו גם לגיבוש פתרונות מגוון מתאים.

1.3 עבודות קודמות בנושא

1.Carlini, N., & Wagner, D. (2018). **Audio adversarial examples: Targeted attacks on speech-to-text.** *arXiv preprint arXiv:1801.01944*.

מאמר זה הראה את הבסיס לעבודה שלנו, והיה בין המאמרים הראשונים שעבדו על התקפות אדוורסריאליות של מערכות שמע.

2.Qin, Y., Carlini, N., Goodfellow, I., Cottrell, G., & Raffel, C. (2019). **Imperceptible, robust, and targeted adversarial examples for automatic speech recognition.** In International Conference on Machine Learning (pp. 5231-5240). PMLR.

מחקר המשך שմissor את התוצאות של המחקר הקודם ומשלב טכניקות התקפה מורכבות יותר.

1.4 מבנה העבודה

מהלך העבודה חולק למספר שלבים עוקבים. בשלב הראשון, אשר נמשך במהלך החודשים הראשונים של הפroyskt, בוצע תהליך לימוד עמוק של נושאים תיאורטיים הנוגעים למינית عمוקה, רשתות טרנספורמר, עיבוד אותות דיגיטלי (בהתבסס על חומרה הקורס שנלמדו בעבר), וכן סוגים שונים של התקיפות אדוורסריאליות.

בהמשך, נלמדו לעומק שני מודלים עיקריים בתחום הפרדת מקורות מוזיקליים – Demucs ו-Spleeter על גרסאותיו השונות, תוך בחינה השוואתית של הארכיטקטורות, אופן הפעולה והביטויים בפועל.

בשלב הבא, הוקן והוטמע מאגר הנתונים MUSDB18, אשר שימש כבסיס לכל הניסויים שבוצעו בהמשך. המודלים הותקנו והורכו באופן מקומי, ונערכה סדרת ניסויים כלללה שימוש בדוגמאות מוזיקליות מגוונות במטרה לאתגר מגבלות בביטויי ההפרדה,>ZHYI תרחישים בעיתים, וחקיר חוזקות וחולשות של כל מודל.

לאחר שלב ההיכרות וההכנה, פותח אלגוריתם התקיפה אדוורסריאלית מסווג קווצה לבנה, ואחריו ישמה גם אסטרטגיית הגנה "יעודית".

בשלב הסופי של העבודה, בוצעה הרצה מקיפה של תוכנות הנתונים באמצעות מדדים נמותנים מקובלים (SDR, SIR, SAR).

2. סקר ספרות

במסגרת חקר הספרות נחקר על מודלים שונים של פירוק ערוצים ועל דרך פעולהם. מפורט בהמשך על כל מודל שנחקר ועל המודל הנבחר.

2.1 - SPLEETER

Source code: <https://github.com/deezer/spleeter>

Paper: <https://archives.ismir.net/ismir2019/latebreaking/000036.pdf>

The paper is from 2019

How to run: spleeter separate -o output/audio_output -p spleeter:4stems "song_path"

Music Source Separation Spleeter ב-Deezer פיתחו אותו כדי לאפשר יישומים שונים: קריוקי, רמיקסים, ניתוח מוזיקלי אוטומטי, ועוד.

הרעין המרכזי של המודל:

STFT (Short-Time Fourier Transform) (Transform) במקומם לבודד ישירות על האודיו בזמן, המודל עובד על ייצוג של האודיו במישור התדרים:

כלומר, הקלט למודל הוא ספקטrogramma.

מבנה המודל:

- רשת Net-U שפועלת על המישור הספקטראלי.
- הקלט: ספקטrogramma מורכבת (מודולו ופזה מופרדים - המודל מתמקד במודולו).
- הרשת מסה ללמידה עבור כל מרכיב במיקס (שירה, תופים וכו') את ה-"mask" המתאים - כלומר, אילו תדרים שייכים לאיזה מקור.
- הפלט של הרשת הוא Masks שאוון מכפילים חזרה בספקטrogrammat המקור כדי לקבל את הספרטrogrammat של כל אחד מהמקורות.

ביצועים

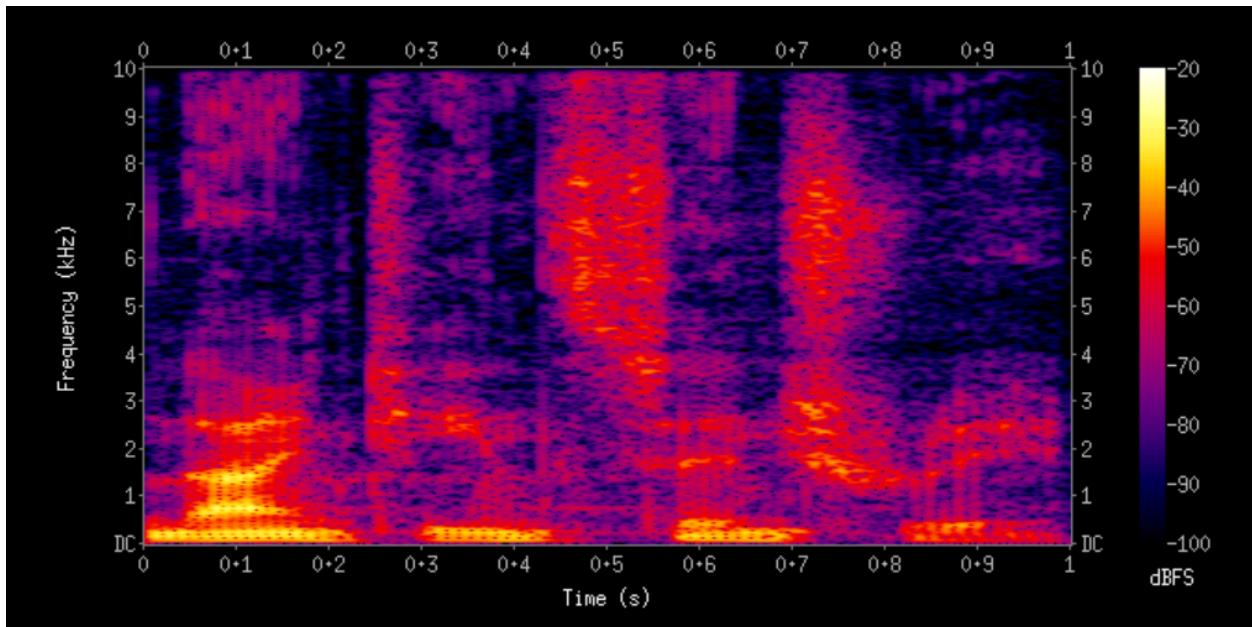
	vocals				bass				drums				other			
	SDR	SIR	SAR	ISR	SDR	SIR	SAR	ISR	SDR	SIR	SAR	ISR	SDR	SIR	SAR	ISR
Spleeter Mask	6.55	15.19	6.44	12.01	5.10	10.01	5.15	9.18	5.93	12.24	5.78	10.50	4.24	7.86	4.63	9.83
Spleeter MWF	6.86	15.86	6.99	11.95	5.51	10.30	5.96	9.61	6.71	13.67	6.54	10.69	4.55	8.16	4.88	9.87
Open-Unmix	6.32	13.33	6.52	11.93	5.23	10.93	6.34	9.23	5.73	11.12	6.02	10.51	4.02	6.59	4.74	9.31

איור 1 - טבלה של ביצועי ספליטיר

מהי ספקטrogramה?

ספקטrogramה (Spectrogram) היא ייצוג חזותי של צליל בשלושה ממדים:

1. ציר X – זמן (מתוך קורה הצליל)
2. ציר Y – תדר (באיזו גובה צליל)
3. צבע/עוצמה – עוצמת הצליל בכל תדר ובכל רגע בזמן



איור 2 - דוגמה לספקטrogramה

איך יוצרים ספקטrogramה?

כדי להפוך את שמע (כמו קובץ WAV) לספקטrogramה, משתמשים ב-STFT (Short-Time Fourier Transform) — כלומר, מחשבים טרנספורם פורייה על מקטעים קצרים מהשמע (במקום על כלו בביטחון). כך אפשר לראות איך התדרים משתנים לאורך זמן.

2.2 - DEMUCS

(קיצור של Deep Extractor for Music Sources Demucs) היא סדרת מודלים להפרדת מקורות מתוך שירים .Vocals, Drums, Bass, Other: (Source Separation)

בעוד ש-Spleeter עובד על ספקטrogramות, Demucs פועל ישירות על האמפליטודה (waveform)

2.2.1 - DEMUCS V2

Code: <https://github.com/facebookresearch/demucs/tree/v2>

Paper: <https://hal.science/hal-02379796/document>

How to run: demucs "song_path"

DEMUCS הוא מודל מבוסס encoder-decoder שמקבל קול מיקס סטראיאו (כלומר שני ערוצים – ימין ושמאל), ומחזיר עבור כל מקור (כמו תופים, באס, שירה) הרצת סטריאו נפרדת.

הארכיטקטורה של DEMUCS V2:

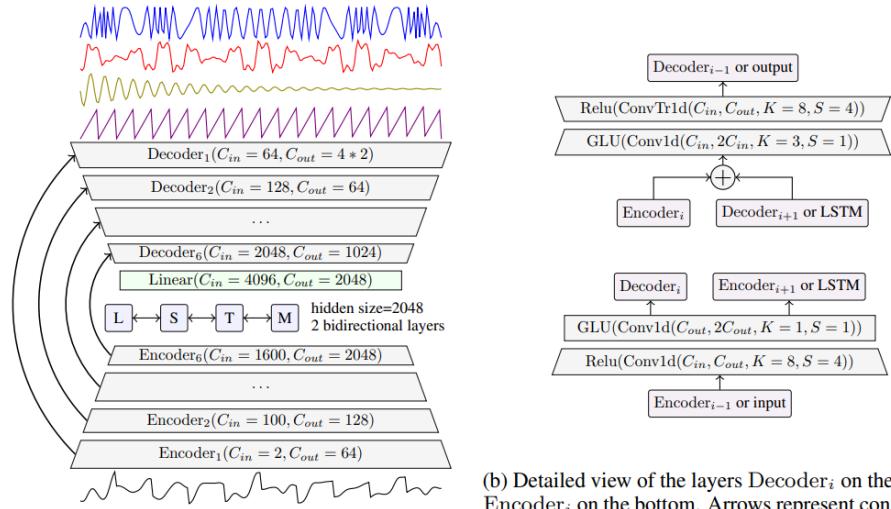
הארכיטקטורה כוללת שלושה שלבים עיקריים:

.1 – ממיר את גל הקול לייצוג מופשט בעומק גבוה. **Convolutional Encoder**

.2 – עיבוד של הרצף לשני הכוונים (קדימה ואחורה בזמן), כדי להבין קשרים ארוכי טווח. **Bidirectional LSTM**

.3 – מפענח בחזרה את ה-wavform עבור כל מקור. **Convolutional Decoder**

בין Encoder ל-Decoder יש skip connections כמו ב-Net-U, שמאפשרים לשמור מידע מפורט משכבות מוקדמות ולהברר אותו בחזרה בשכבות המאוחרות.



(a) Demucs architecture with the mixture waveform as input and the four sources estimates as output. Arrows represents U-Net connections.

(b) Detailed view of the layers $Decoder_i$ on the top and $Encoder_i$ on the bottom. Arrows represent connections to other parts of the model. For convolutions, C_{in} (resp C_{out}) is the number of input channels (resp output), K the kernel size and S the stride.

Figure 2: Demucs complete architecture on the left, with detailed representation of the encoder and decoder layers on the right.

איך 3 - תרשימים ארכיטקטורת דימוקס גרסה 2

ביצועים:

Architecture	Wav?	Extra?	Test SDR in dB				
			All	Drums	Bass	Other	Vocals
IRM oracle	✗	N/A	8.22	8.45	7.12	7.85	9.43
Wave-U-Net	✓	✗	3.23	4.22	3.21	2.25	3.25
Open-Unmix	✗	✗	5.33	5.73	5.23	4.02	6.32
Meta-Tasnet	✓	✗	5.52	5.91	5.58	4.19	6.40
Conv-Tasnet [†]	✓	✗	$5.73 \pm .10$	$6.02 \pm .08$	$6.20 \pm .15$	$4.27 \pm .03$	$6.43 \pm .16$
DPRNN	✓	✗	5.82	6.15	5.88	4.32	6.92
D3Net	✗	✗	6.01	7.01	5.25	4.53	7.24
Demucs [†]	✓	✗	$6.28 \pm .03$	$6.86 \pm .05$	$7.01 \pm .19$	$4.42 \pm .06$	$6.84 \pm .10$
Spleeter	✗	~ 25k*	5.91	6.71	5.51	4.55	6.86
TasNet	✓	~ 2.5k	6.01	7.01	5.25	4.53	7.24
MMDenseLSTM	✗	804	6.04	6.81	5.40	4.80	7.16
Conv-Tasnet ^{††}	✓	150	$6.32 \pm .04$	$7.11 \pm .13$	$7.00 \pm .05$	$4.44 \pm .03$	$6.74 \pm .06$
D3Net	✗	1.5k	6.68	7.36	6.20	5.37	7.80
Demucs [†]	✓	150	$6.79 \pm .02$	$7.58 \pm .02$	$7.60 \pm .13$	$4.69 \pm .04$	$7.29 \pm .06$

איור 4 - טבלת השוואת ביצועי SDR של מודלים להפרדת ערוצים

2.2.2 - DEMUCS V4

Transformer הוא שדרוג של המודל Hybrid Demucs המקורי, בשילוב עם Cross-domain Transformer (Cross-domain Transformer).

Code: <https://github.com/adefossez/demucs>

Paper: <https://arxiv.org/pdf/2211.08553>

The paper is from 2022

How to run: demucs -n htdemucs_ft "song_path"

מבנה המודל

המבנה בניו מ-2 מסלולים עיקריים:

1. מסלול הזמן (Waveform domain) – עם קונבולוציות זמניות.

2. מסלול הספקטרום (Spectrogram domain) – עם קונבולוציות בציר התדר.

לכל מסלול יש:

- 5 שכבות Encoder

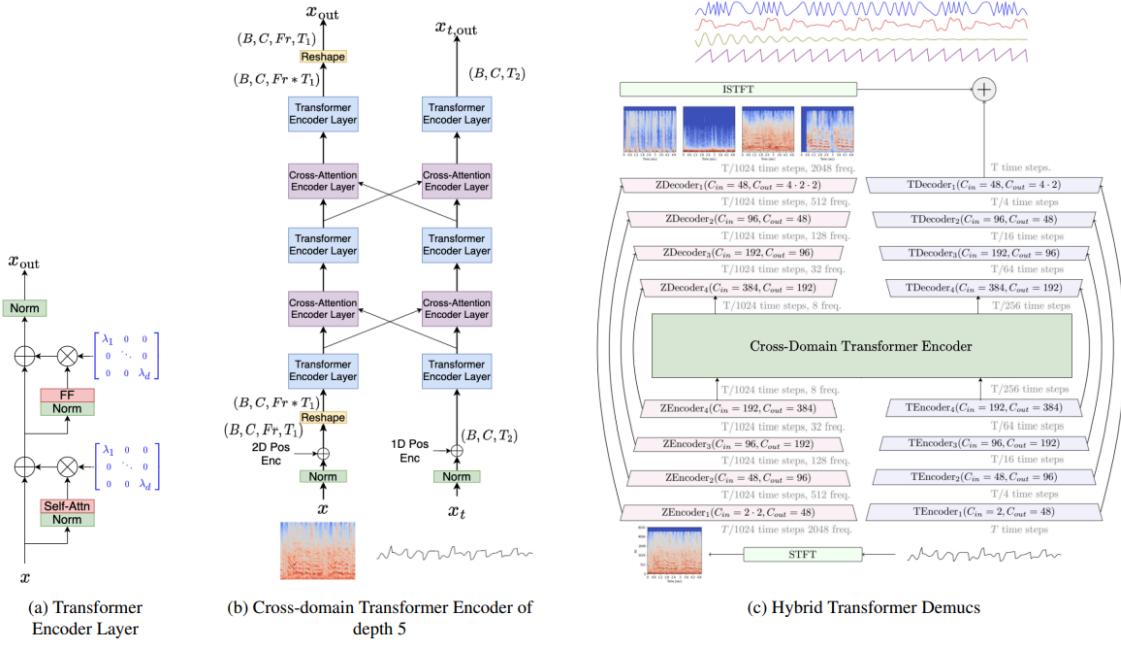
- 5 שכבות Decoder

לאחר שכבת Encoder החמישית בשני המסלולים – היצוגים מיושרים ונחברים, ואז מוזרמים יחד ל-שכבה ששית משותפת.

חידוש המרכזי – Cross-domain Transformer

ה-Transformer פועל גם על הספקטרום (2D) וגם על גל הזמן (1D) במקביל.

כולל שילוב של Self-Attention (spectrogram / waveform) ו-Cross-Attention (waveform / waveform) בין התחומיים – כמו למשל, תדרים וזמן "מדוברים" זה עם זה.



איור 5 - תרשימים ארכיטקטורת דימוקס גרסא 4

ביצועים

Architecture	Extra?	Test SDR in dB				
		All	Drums	Bass	Other	Vocals
IRM oracle	N/A	8.22	8.45	7.12	7.85	9.43
KUIELAB-MDX-Net [17]	✗	7.54	7.33	7.86	5.95	9.00
Hybrid Demucs [2]	✗	7.64	8.12	8.43	5.65	8.35
Band-Split RNN [14]	✗	8.24	9.01	7.22	6.70	10.01
HT Demucs	✗	7.52	7.94	8.48	5.72	7.93
Spleeter* [19]	25k	5.91	6.71	5.51	4.55	6.86
D3Net* [12]	1.5k	6.68	7.36	6.20	5.37	7.80
Demucs v2* [10]	150	6.79	7.58	7.60	4.69	7.29
Hybrid Demucs [2]	800	8.34	9.31	9.13	6.18	8.75
Band-Split RNN [14]	1750 [†]	8.97	10.15	8.16	7.08	10.47
HT Demucs	150	8.49	9.51	9.76	6.13	8.56
HT Demucs	800	8.80	10.05	9.78	6.42	8.93
HT Demucs (fine tuned)	800	9.00	10.08	10.39	6.32	9.20
Sparse HT Demucs (fine tuned)	800	9.20	10.83	10.47	6.41	9.37

אייר 6 - טבלת השוואת ביצועי SDR של מודלים להפרדת ערוצים מול הגרסה ההיברידית החדשה של דימוקס

2.3 בחירת (ht demucs)

החליט לבחור במודל Demucs HT כבסיס למערכת ההגנה שלנו מפני מתכונות אדוורסריאליות, מספר סיבות מרכזיות:

1. **נגישות ונוחות שימוש:**
המודל זמין רקוד פתוח, קל להריצה, וכל אפשרות להורדה של גרסה מאומנת מראש. כך ניתן לבצע ניסויים ללא צורך באימון ממושך על מכונות יקרות.

2. **ביצועים גבוהים:**
HT Demucs הגיע לתוצאות מובילות במדד הפרדת מקורות בהשוואה למודלים אחרים שנבדקו, לרבות על נתונים סטנדרטיים כגון MUSDB.

3. **תיעוד ומחקר מקיים:**
המודל מלאה במאמרים עדכניים וברורים, המתארים את הארכיטקטורה והעקרונות שעומדים מאחוריה, מה שמקל על הבנה, התאמת ופיתוח נוספים לצרכים שלנו.

3. מטריקות הערכתה

הסבר על מטריקות הערכתה
על פי איזה קритריונים ימדד מהו פירוק טוב?

1. שימור הערך המקורי רצינו למדרג (כמה שומעים טוב את התופים לצורך העניין)
2. מידור הערוצים האחרים (כמה לא שומעים את שאר הערוצים)
3. נוכחות רעשנים מלאכותיים
- 4.aicות כוללת של הערך עצמו

3.1 SDR

Source-to-Distortion Ratio

מה זה SDR?

SDR measures the power ratio between the intended source and the distortion artifacts introduced during separation.

איך SDR עובד?

המדד משווה בין:

1. האות המקורי של מה שרצית להפריד (לדוגמה, רק השירה).
2. האות שהמודול שלך הפיק (השירה שהצליחה להוציא מהמיקס).

הוא בעצם מודד את היחס בין האנרגיה של האות המקורי (המטרה) לבין הטיעות הכוללת (הרעש, ההפרעות והעיוותים שהמודול יצר).

SDR גבוה אומר שה הפרדה טובה

SDR נמוך אומר שה תוצאה לא קרובה מאוד למקור

$$SDR = 10 * \log_{10}(E_s / E_d)$$

Where:

- E_s is the energy of the original or reference signal.
- E_d is the energy of the distortion or noise in the degraded signal.

איך להבין את המספר?

- SDR גבוה (לדוגמה, 15-20): המודול הצליח מאד, ההפרדה מדוקת.
- SDR נמוך (לדוגמה, 5-10): יש הרבה הפרעות בתוצאה, ההפרדה לא טובה.

3.2 SIR

Signal-to-Interference Ratio

מודד את איכות הפירוק של הקרקע המופרד (למשל ערז של כלי נגינה מסוימים) בהשוואה לרעש (interference) ש מגיע ממוקורות אחרים שנמצאים באותו הקרקע.

הוא מחושב על ידי השוואת הסיגナル המקורי (שאנו רצים לשמר) לעומת הסיגナル שנמצא בשאר הערכות (שהוא למעשה רעש או הפרעה). אם ה-SIR גבוהה, המשמעות היא שהפירוק של העוז טוב ויש פחות רעש או הפרעה. אם ה-SIR נמוך, המשמעות היא שהפירוק פחות טוב ויש יותר רעש.

המטריקה נמדדת בדציבלים (dB), כאשר ערכים גבוהים יותר מצביעים על תוצאה טובה יותר של הפירוק.

$$\text{SIR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}$$

- e_{interf} : component of the estimate that aligns with *other* true sources (interference).
- Zero interference $\Rightarrow \text{SIR} \rightarrow +\infty$.

Interpretation:

- **Higher SIR \Rightarrow better separation (less bleed-through).**
- SIR only looks at cross-talk, not at artifacts or noise.

הבדל בין SIR לSDR:

SDR מודד את איכות הסיגナル הסופי נגיד אם הגיטרה נשמעת מתכנית או לא אמיתית או נגיד יש מקטעים שהווים לא אחיד לדוגמה
SIR בודק את ההפרדה מהכלים האחרים. כלומר אם שומעים קצת תופים בערך של הגיטרה אז SIR היה נמוך

מה הטווחים של SIR ?

SIR גבוה = טוב מאוד (יותר מ-20 dB).
SIR בינוני = סביר (dB 10-20).
SIR נמוך = לא טוב (פחות מ-10 dB)

3.3 SAR

Signal-to-Artifacts Ratio

What it measures:

Amount of “unnatural” noise or processing artifacts introduced by the separation algorithm.

Definition:

$$\text{SAR} = 10 \log_{10} \frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2}$$

- e_{artif} : residual error after removing both the target component and the interference component from the estimate.

Interpretation:

- Higher SAR \Rightarrow fewer artifacts (cleaner processing).
- SAR focuses strictly on algorithmic distortions, ignoring leakage.

$$e_{\text{artif}} = \hat{s}_i - \left(\underbrace{\frac{\langle \hat{s}_i, s_i \rangle}{\langle s_i, s_i \rangle} s_i}_{s_{\text{target}}} + \underbrace{\sum_{j \neq i} \frac{\langle \hat{s}_i, s_j \rangle}{\langle s_j, s_j \rangle} s_j}_{e_{\text{interf}}} \right)$$

where:

- \hat{s}_i is the estimated i th source.
- s_i is the true i th source (the “target”).
- s_j for $j \neq i$ are the other true sources.
- $\langle x, y \rangle = \sum_n x[n] y[n]$ is the inner product.

So you subtract from your estimate both:

1. its projection onto the true target s_i , and
2. its projections onto all other sources s_j ,

and the remainder is the “artifact” component e_{artif} .

4. סוגי התקפות

ישנן כמה סוג התקפות שנitinן לבצע על מודלים. בפרק זה נסביר בקצרה על כמה מהם.

4.1 התקפות הרעלת

התקפות הרעלת מתבצעות בשלב האימון של המודל, כאשר התקוף מחדיר אליו דוגמאות קטל מזיקות או שגויות במטרה לשבש את תהליכי הלמידה. מטרת התקפה זו היא לפגוע באמינות המודל, כך שבמהלך השימוש בו בעתיד הוא יניב תוצאות שגויות או לא צפויות.

התקפה מסווגה זו עלולה להתבצע, למשל, על ידי שחקן עין כמו חברה מתחרה, אך תיתכן גם באופן בלתי מכוון – כאשר מפתח המודל משתמש במאגר נתונים שכולל שגיאות, הטויות או דגימות פגומות, מבליל להיות מודעים לכך.

4.2 התקפות פרטיות

בדרכ כל התקפה זו נעשית על צ'אט בוטים ומטרתה להפיק מידע מהמודל אודוות המודל עצמו או כל מידע רגיש צזה או אחר או לחlopen נקודות תורפה של המודל שהמודל עלול לגלות. לדוגמה, אם המודל אומן על מידע רגיש התקוף עלול לחושף מידע זה.

4.3 התקפות ניצול לרעה

התקפה נוספת שנפוצה במודלים של שפה טבעיות. במקרה זה, התקוף משנה או עורק מקור שנחשב כאמין על ידי המודל בזמן הפעלה של המודל (למשל ויקיפדיה). התקפה זו יכולה לשמש למטרות פרוגנדיה, ממשלה או ארגונים שונים.

4.4 התקפות התchmodקות - Evasion Attacks

התקפות התchmodקות הן סוג של התקיפות אדוורסריאליות המתבצעות בשלב הפעלה של המודל (ולא בשלב האימון), ובهن התקוף משנה את הקלט המקורי באופן מזרע או מבוקר, כך שהשינוי אינו ניתן לזיהוי בעין או באוזן האנושית, אך מספיק כדי לגרום למודל להפיק פלט שגוי או שונה מהצפוי.

במקרים מסוימים, מטרת התוקף היא לגרום למודל להפיק פלט שגוי כללי, ובמקרים אחרים – להפיק פלט מסוים ומוגדר מראש, בהתאם לאינטראס של התוקף.

סוג זה של התקפה עשוי להוביל לתוצאות חמורות, במיוחד כאשר המערכת נמצאת בשימוש רגיש.

זהו סוג ההתקפה שיזום בפרויקט הנוכחי, כחלק מבדיקה עמידות של מודל להפרדת מקורות מזיקליים מפניהם שינוי אדוורסרילים צעירים בקלט.

דוגמאות למערכות רגישות שיכלות להיפגע מהתקפות אלו:

1. **מערכות דיזי פקודות קוליות:**

במערכות המקבלות קלט קולי לצורך הפעלת פקודות (כגון עוזרות קוליות או מכשירים חכמים), תוקף יכול לשלב פקודה מסוימת בקובץ אודי תמים – למשל Shir המתנגן ברקע – כך שהמערכת תזהה פקודה כלשהי ותבצע פעולה לא רצiosa.

2. **רכב אוטונומי:**

רכב אוטונומי מסתמך על מערכת ראייה מוחשבת ליזי תמורים. תוקף עשוי לשנות תמרור (באופן שלא ניתן להבחין בו בעין אנושית) כך שהמערכת תזהה אותו כתמרור שונה לחלוtin – למשל, להפוך תמרור עצור לתמרור נסיעה חופשית.

3. **סימנים סמיים על הכביש:**

הוספה של דפוסים גיאומטריים או צבעים שאינם מורגשים לעין האנושית עשויה לשבש את מערכת הניווט של רכב אוטונומי ולגרום לו לסתות ממסלו.

4. **מערכות סיווג תמונה:**

תוקף יכול לשנות קובץ תמונה בשינויים מינימליים כך שמערכת סיווג תזהה את התמונה כקטgorיה שגיה, מה שעלול לשבש תפקוד של מערכות אבטחה, ניטור רפואי, בקרת איכות תעשייתית או כל מערכת שמשתמשת במודל לסיווג תמונה.

בתוך ההתקפה זו ישנן שתי תת-התקפות עיקריות:

4.5 התקפה קופsea לבנה

מתבצעת כאשר הקוד, הארכיטקטורה והמשקלים של המודל נתונים לתוקף במלואם.

התוקף עלול להשתמש בגזירה לאחר (backpropagation) בשבייל לחשב את את נגזרת פונקציית ההפסד (הגרדיינט) ביחס לקלט (בניגוד לחישוב נגזרת פונקציית ההפסד ביחס למשקלים בעת אימון המודל).

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{y})$$

כאשר \mathbf{x} הוא הקלט, \mathbf{y} הוא הפלט ו- L היא פונקציית ההפסד

בהתבסס על גרדיאנט זה, התוקף מבצע סדרת איטרציות, כאשר בכל איטרציה הוא משנה את היקלט בשינויים קטנים מאוד (שלא ניתנים לזיהוי בעין או באוזן אנושית), כך שהפלט של המודל יוסט באופן משמעותי מהתוצאה צפויה.

באופן זה, ניתן לגרום למודל להיכשל במשימתו, או אפילו להפיק פלט מסויים שהוגדר מראש – כל זאת מבלי לשנות את היקלט באופן מורגש.

ישנן שתי גישות להתקפה זו:

4.5.1 התקפה מכונה

*בהסבר הבא תהיה התיחסות למודלים של סיוג ולא מודלים גנרטיביים (Generative Models)

במקרה זה, התוקף מעוניין בפלט ספציפי שהוגדר מראש, אך הוא יעשה את חישוב פונקציית ההפסד ביחס לפלט שהוא רוצה, ובכל איטרציה הוא ישנה מעט את היקלט בעזרת הגרדיאנט עד שפונקציית ההפסד תראה מספיק נמוכה ביחס לפלט זה.

דוגמה:

נניח כי ברשותנו מודל למידת מכונה מסווג בין תמונות של כלבים לחתולים. בהתקפה אדוורסריאלית, ניתן לעבד תמונה של כלב באמצעות שינויים מינוריים ובלתי נראים לעין האנושית, וכך לאחר ההתקפה, המודל יסוג אותה כתמונה של חתול – למרות שבענין המתבונן האנושי, התמונה נותרת בבירור תמונה כלב.

הסבר מתמטי:

Given model $f(x)$ where x is input, Y is the true label and T is the desired label

$$Y \neq T$$

Find $\hat{x} = x + \delta$ where's δ is the perbutation that gives $f(\hat{x}) = T$

4.5.2 התקפה לא מכונה

כאן המטרה היא לא לקבל פלט ספציפי אלא פלט שגוי כלשהו במטרה לשבש את המודל ולהתקבל תוצאות לא צפויות. כאן התוקף ישתמש בגרדיאנט ביחס לפלט הצפי, אבל הוא ישנה את היקלט כך שפונקציית השגיאה תהיה מקסימלית במטרה לא לקבל את הפלט הצפי.

זהות ההתקפה אשר תשומש בפרויקט.

בשתי הgisות, היקלט החדש שנוצר נקרא דוגמה אדוורסריאלית (adversarial example)

כמו כן, התוקף יגביר את הדוגמה האדוורסריאלית בכך שהשוני בין הדוגמה והיקלט המקורי לא יהיה ניתנים להבחנה על ידי בן אדם.

4.6 התקפה קופסה שחורה

במתפקיד אדוורסריאליות מסווג התקפה קופסה שחורה (Black-box Attacks), התקוף אינו מחזיק בגישה ישירה למודל הלמידה — כלומר, אין לו ידע על מבנה הרשות, משקלים פנימיים, או גזירת גרדיאנטים. התקוף יכול להפעיל את המודל קופסה שחורה בלבד: להזין קלטים ולקבל את הפלט (למשל, תווית סיווג או ערך החיזוי), בדומה לגישה חיצונית למשק API.

במתפקיד מסווג זה, התקוף לרוב משתמש באחת מהשיטות הבאות:

4.6.1 התקפה באמצעות חיזוי תחלפי - Surrogate Model

התקוף מאמין מודל חלופי על בסיס דוגמאות קלט-פלט שהתקבלו מהמטרה, ובמצע עליו התקפה לבנה (White-box). לאחר מכן, משתמש באותו רושי התקפה כדי לתקוף את המודל המקורי. שיטה זו מנצלת את תופעת *transferability* — יכולת של דוגמאות אדוורסריאליות לפעול על מספר מודלים שונים.

4.6.2 אופטימיזציה מבוססת חיפוש - Query-based Optimization

בשיטת אלו, התקוף שואל את המודל שוב ושוב עם גרסאות שונות של הקלט, ובודק אילו שינויים גורמים לשינוי בפלט. לדוגמה, אלגוריתמים מבוססי אבולוציה, חיפוש אקראי, או אלגוריתםolution NES (Natural Evolution Strategies) משמשים למציאת שינוי קטן בקלט שייגרום לשגיאה משמעותית במודל.

ישן כמה שיטות לאופטימיזציה מבוססת חיפוש:

1. מטפס הרים - Hill Climbing

טכניקת אופטימיזציה שבה מתחילה מנקודת התחלת כלשהי (למשל קלט אודיו מקורי), ומבצעים שינויים קטנים ("צעדים") בקלט, כך שבכל שלב בוחרים את השינוי שմangler את התוצאה (למשל, מגדיל את הסיכוי לטעות של המודל).

בקשר אדוורסרילי:

התקוף מוסיף רעש קטן לקלט, בודק אם המודל שינה את הפלט (למשל, פירק את הערכתם בצורה פחות מדויקת), ואם כן — ממשיך באותו כיוון. אם לא, מנסה כיוון אחר. זהו תהליך איטרטיבי שדורש פניות חוזרות

למודל. החיסרון הוא שהמערכת שלו יכולה בקלה ליתקע במקסימום מקומי – כלומר, לא תמיד ימצא את ההפרעה האופטימלית.

2. שיטת מונטה קרלו - Monte Carlo Sampling

טכניתה הסתברותית שבה מבצעים דגימה אקראית מטווח מרחב האפשרויות, כדי לאמוד או לגלו את הפתרון הטוב ביותר.

בקשר אדוורסרילי:

התוקף יוצר הרבה גרסאות רועשות של הקלט המקורי (noise perturbations), ושולח אותן למודל. על סמך התגובה של המודל (למשל ירידה ב-SDR), הוא מזהה אילו שינויים גורמים להפרעה מקסימלית, ובוחר להעמיק באזוריים אלו של מרחב הקלט.

יתרון: לא דורש גרדיאנט, מתאים למצבים בהם יש רק גישה לפלט.

חיסרון: דורש הרבה מאוד דגימות (queries) – יקר בזמן/משאבים.

3. אלגוריתמים אבולוציוניים - Evolutionary Algorithms

שיטה בהשראת האבולוציה הבילוגית. מגדרים "אוכלוסיה" של פתרונות (גרסאות שונות של הקלט), מודדים לכל אחת את איכותה (fitness), ובוחרים את הפתרונות הטובים ביותר ל"התרבות" – שילוב ומוטציה יוצרים דור חדש, וחוזר חלילה.

בקשר אדוורסרילי:

ה"אוכלוסיה" היא קבוצה של קלטים עם רעש שונה. כל אחד נשלח למודל, נמדדת "כמות ההטעיה" שלו (כמה המודל טועה). בוחרים את הדוגמאות הטובות ביותר ביותר, יוצרים מהן גרסאות חדשות עם מוטציה, ווחזרים על התהילה.

יתרון: יכול למצוא פתרונות טובים גם במרחב חיפוש מורכב.

חיסרון: איטי, דורש הרבה פניות למודל וDIRG של כל דוגמה.

4.6.3 שיטות ללא תוויות - Score-based or Decision-based

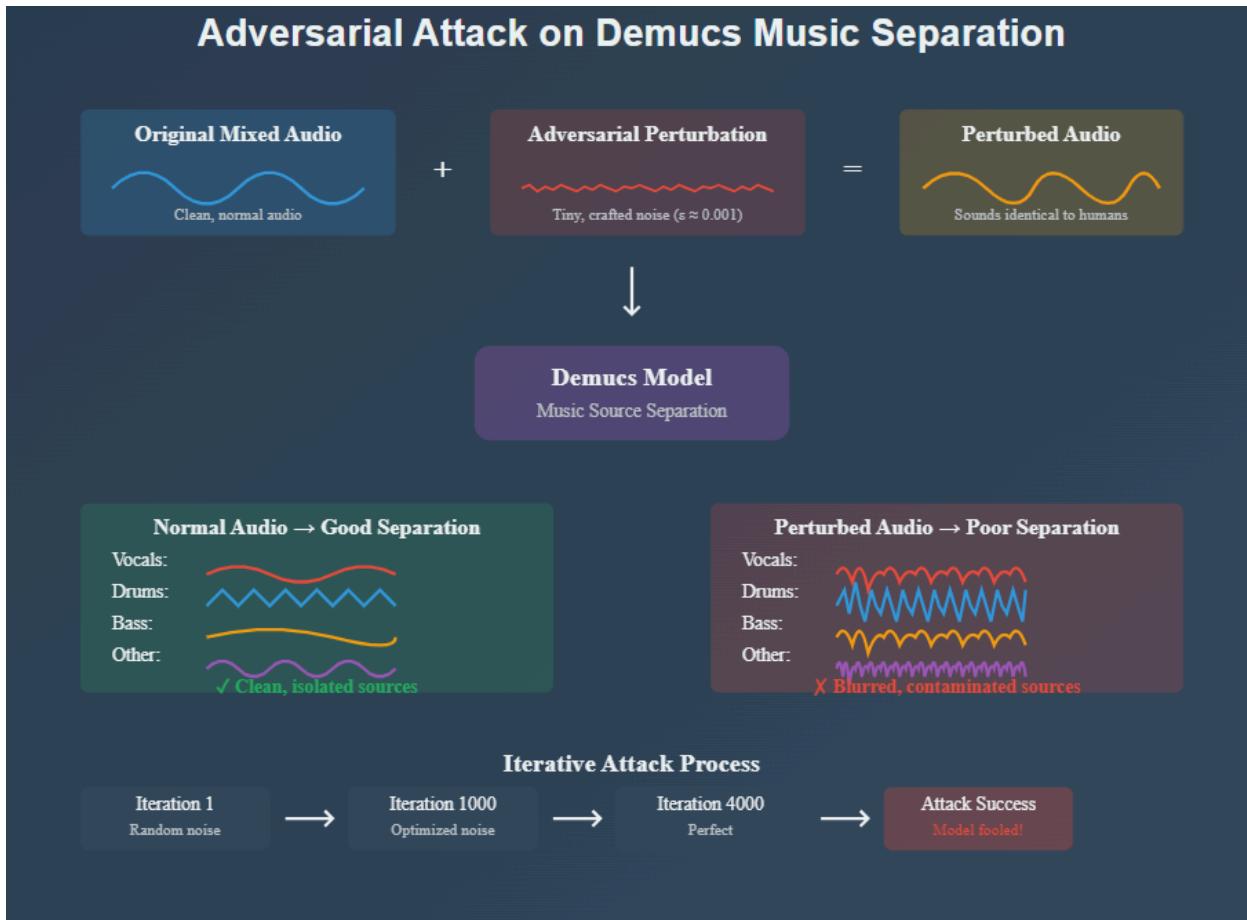
כאשר המודל מחזיר רק את התווית הסופית (ולא את הסבירות או פונקציית ההפסד), משתמשים בשיטות כמו OPT-Sign, Boundary Attack או Sign-OPT, שמחפשות את הגבול בין הקלאסים על ידי דגימה אקראית והתקדמות הדרגתית.

שלא כמו בהתקפות קופסה לבנה, בהן ניתן להשתמש בಗזירת גרדיאנטים ישירה, התקפות קופסה שחורה דורשות בדרך כלל יותר פניות (queries) למודל, והן איטיות יותר, אך מתאימות למצבים מציאותיים יותר – כמו למשל במערכות בענן או שירותים סגורים, שבהם המשתמש לא נחשף למודל הפנימי.

5. מימוש התקפת קופסה לבנה בפרויקט

כמו שצינו, בפרויקט מומשה התקפה אדוורסריאלית בסגנון קופסה לבנה לא מכוונת על מודל של הפרדת ערכאים של Shir.

עיקר ההתקפה מtabסת על מאמר של [3] carlini and wagner שהתמקד בהתקפה מכוונת על מודל של דיבור לטקסט, כמו בהינתן דיבור כקלט המודל מפרש את זה לטקסט כפלט.



איור 7 - תרשים של דרך הפעלת הדוגמה האדווורסריאלית על המודל

התקפה נעשית באמצעות חישוב הגדריאנט ביחס לקלט באופן איטרטיבי, כלומר בכל צעד אנו מושנים מעט את השיבוש שנוסף לקלט המקורי בכיוון של הגדריאנט ובכך מייעלים את התקפה.

דבר זה נעשה על ידי בניית אלגוריתם שմבצע:

א. מוסיף רעש קטן לקלט

ב. מעביר את הקלט דרך המודל של דימוקס

ג. מחשב שהגיאת הפוכה, כלומר פונקציית השגיאה דומה לפונקציית השגיאה של דימוקס אבל מוכפלת במינוס אחד.(המטרה זה למקסם את השגיאה המקורי = למזער את השגיאה הפוכה).

ד. מחשב את הגדריאנט ביחס לשגיאה ועשה צעד בכיוון, כלומר משנה את הרעש שנוסף לקלט.

ה. אם יש צורך, מגביל את הרעש שלא יהיה חזק מדי באמצעות אפסילון כלשהו.

ו. חזרה לא'.

הגבלת הפרמטרים

ישנן שיטות שונות וمتוחכבות להגבלת הפרמטרים בין אם בתמונה או בשמע.

בפרויקט יש שימוש בשיטה פשוטה שנקראת `epsilon` $\leq \|\text{perturbation}\|_i \leq \epsilon$.
כלומר, בהינתן אפסילון נעשה "חיתוך" ערכים בערך מוחלט שקטנים מepsilון שלו כדי לכל היותר אפסילון או מינוס אפסילון.

פרמטרים

איטרציות: 5000, אבל גם 4000 יכול לעבוד.

אפסילון: 0.0005

קצב למידה (learning rate alpha) 0.00005

6. אלגוריתם ההגנה

מנגנון ההגנה נועד לנטרל את ההשפעה של הדוגמה האדודורסיאלית על המודל. הרעיון המרכזי במנגנון זה הוא להוסיף רעש רנדומלי קטן לקלט שהותקף, כך שהרעש יהיה בלתי נשמע לאוזן האנושית ועדיין ישמש את הדוגמה האדודורסיאלית שהוספה לאות על ידי מודל ההתקפה ובכך להחזיר את הפלט של המודל לתוצאה תקינה יותר.

איך האלגוריתם עובד?

1. הוספת רעש מוגבל לפי נורמת אינסוף (L^∞):

עבור כל קובץ מותקף, נוצרת מטריצת רעש אקרואית בגודל זהה לקובץ השמע. הרעש מנורמל לטווח [−1,1], ואז מותאם לגבול קבוע מראש מראש epsilon (למשל 0.003), כך מובטח שהרעש לא יחרוג מהטווח הרצוי בכלל נקודה.

2. חיבור הרעש לקובץ המותקף:

קובץ האודיו מועתק עם תוספת הרעש והתוכאה נשמרת תחת שם חדש בתיקיית הפלט.

ההיגיון מאחורי זה:

ההתקפה האדוורסיאלית מחושבת בצורה מאוד מדויקת.
כל שינוי קטן בקלט נבחר בכונה לפי ציון הגראדיאנט – כדי למקסם את השגיאה של המודל.
כלומר, מדובר בשימוש מאד ספציפי, מאוד רגיש, ולא אקראי.
כשמוסיפים רעש אקראי – אפילו קטן, הוא "שובר" את האיזון הזה בכך שהוא מערבב את הקלט מחדש,
והופך את השימוש הקודם ללא אפקטיבי,
בלי לפגוע באופן מורגש בשמע המקורי (כי הרעש מאד קטן).

7. מסד הנתונים MUSDB18

שותש במסד הנתונים MUSDB18:

<https://paperswithcode.com/dataset/musdb18>

מסד זה כולל 150 שירים המולוקים ל-100 שירי אימון, ו-50 שירי בדיקה
המודל של Demucs אותו נבחר לשפר, אומן על דאטאסת זה
מסד נתונים זה נבחר משתי סיבות עיקריות:

1. לצורך בחינה מדעית של השפעת התקיפה והאגנה, יש לבצע את הניסויים על שירים הדומים באופיים
לאלו ששימושו לאימון מודל Demucs. שימוש בדוגמאות שאינן מייצגות את תחום האימון של המודל עלול
להוביל להפרעות או שיבושים בתוצאות, אשר מקורם לא ניתן לזרוי חד-משמעות – האם נובעים
מהתקיפה עצמה או מהתאמתה לקויה בין הדטה למודל.

לדוגמה, במידה שתיבחר דוגמת שמע מז'אנר מוזיקלי שונה (כגון רוק כבד), והמפריד יכשל בפירוק תקין
של הערכאים, לא יהיה ניתן לייחס את הכישלון בהכרח להתקפה, שכן יתכן והמודל מלכתחילה לא
הותאם לד'אנר זה במהלך האימון, והביצועים הנומכים צפויים גם ללא כל שיבוש יזום.

2. המסד זהה מספק נתוני פירוק אמייתיים של כל הנגינה. ככלומר עבור כל שיר במסד הנתונים זהה יש לנו:

- a. Mixture - השיר המקורי
- b. Drums - פירוק ערזע התופים, כולל כל הקשה ואפקטי מצלבים אלקטרוניים
- c. Bass - קונטרא בס או כל בס שונה הבסיס המופרד, יכול להיות גיטרה בס או בס אלקטרוני
- d. Vocals - ערזע הקולות המופרד, כל מה שנקלט כקול אנושי
- e. Other - כל שאר כל הנגינה והאפקטים שלא נכנסו לאחד מהערוצים המופרדים לעיל

הפורמט

כל שיר מיוצג ע"י קובץ בפורמט mp4.song_name.stem

הפורמט זהה מכיל את חמישה הקבצים לעיל

נכתב קוד שיפרק את קובץ stem לחמשת ערזעיו השמע הרצויים בפורמט wav (באיכות גבוהה)

8. רשיימת מקורות

8.1 מאמרים

[1]

Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2025). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2e2025). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-2e2025>

[2]

Rouard, S., Massa, F., & Défossez, A. (2023). Hybrid transformers for music source separation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[3]

Carlini, N., & Wagner, D. (2018). Audio adversarial examples: Targeted attacks on speech-to-text. In Proceedings of the 2018 IEEE Symposium on Security and Privacy (SP) (pp. 1-15). IEEE.

[4]

Alexandre Défossez, Nicolas Usunier, Léon Bottou, Francis Bach
(2021). Music Source Separation in the Waveform Domain

[5]

Romain Hennequin, Anis Khelif, Felix Voituret, Manuel Moussallam
(2019). SPLEETER: A FAST AND STATE-OF-THE ART MUSIC SOURCE
SEPARATION TOOL WITH PRE-TRAINED MODELS

[6]

Simon Rouard, Francisco Massa, Alexandre Defossez
(2022). HYBRID TRANSFORMERS FOR MUSIC SOURCE SEPARATION

8.2 מקורות נוספים

מוד הנתונים :MUSDB18

<https://paperswithcode.com/dataset/musdb18>

Evaluation Metrics for Speech(Audio) Signal Processing:

<https://medium.com/@poudelnipriyanka/audio-metrics-their-importance-and-their-necessity-417950b0d848>

9. תוצאות

9.1 הסבר על הגרפים

Original Separation = evaluate the original channels compared to the demucs estimation separation

Attack Effect = comparing the demucs operation on the original song to the demucs operation on the attacked song

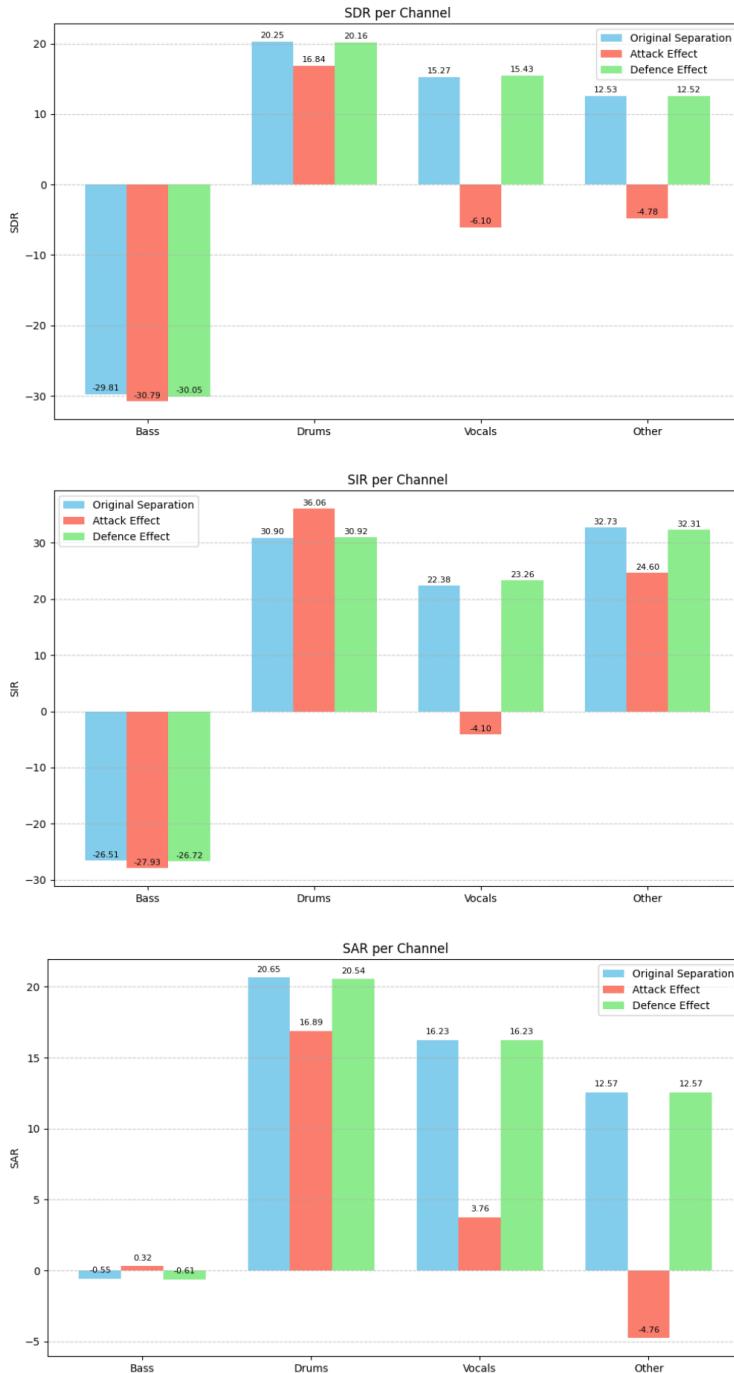
Defense effect = comparing the original channels to the demucs operation on a song that has been attacked and defended

לקחנו 7 שירים מתוך מסד הנתונים MUSDB18 מתוך תיוקית הבדיקה ועליהם ביצענו:

- .1. התקפה
- .2. הגנה
- .3. הערכה

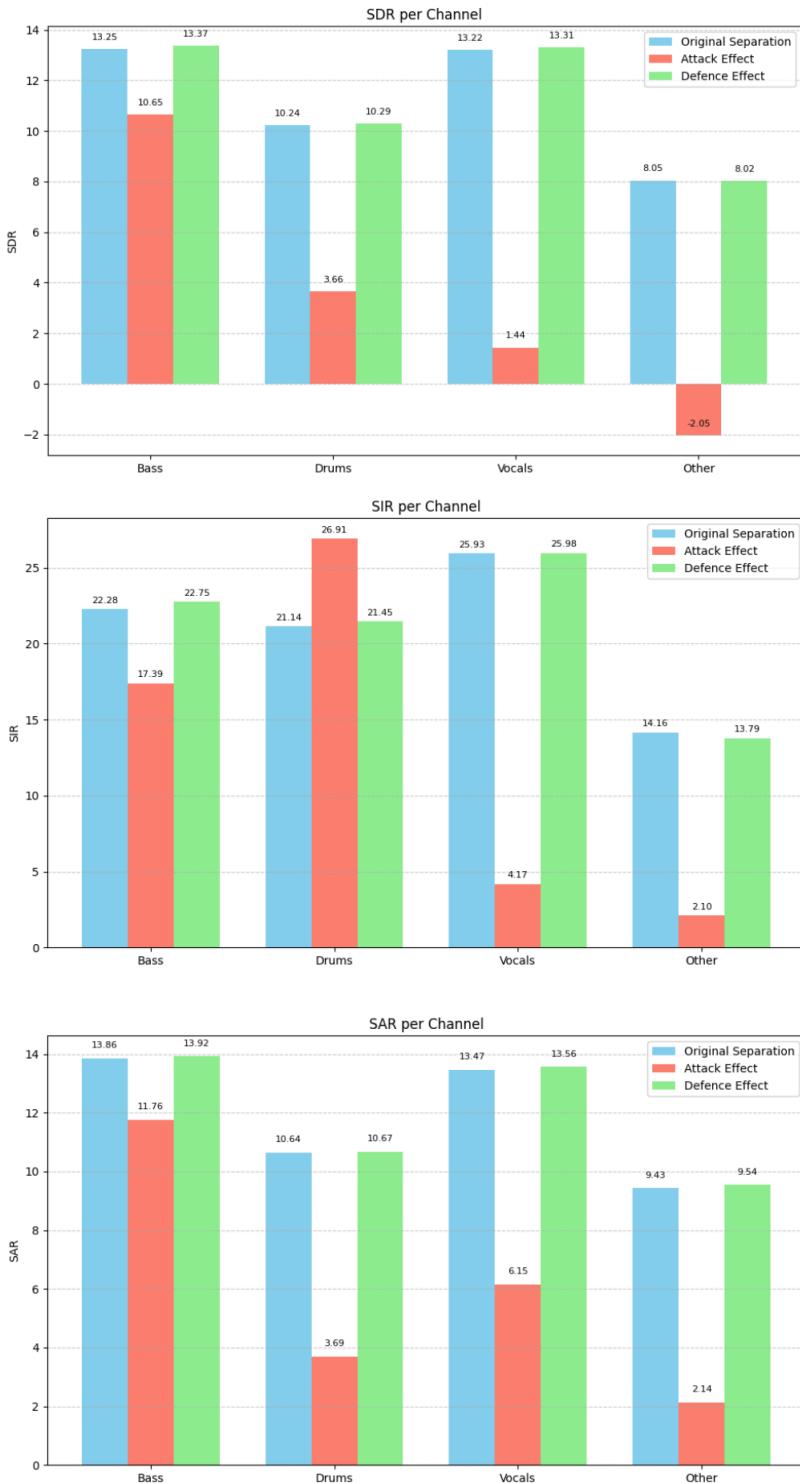
לכל שיר מודפסים שלושה גרפים, כל גרפ מיצג מטריקת הערכה שונה (SAR,SIR,SDR)

Song 1: AM Contra - Heart Peripheral



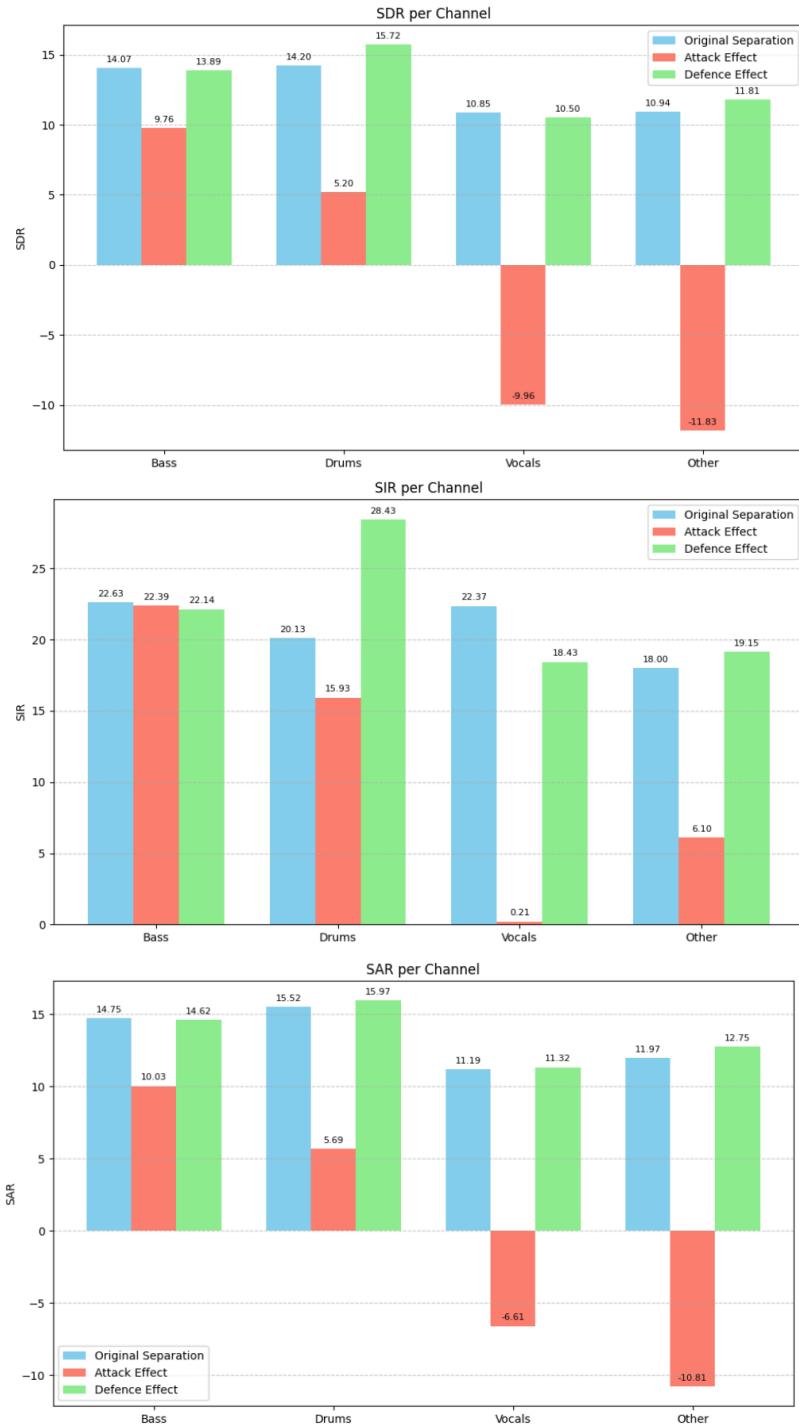
איור 8 - הציג ביצועי SDR,SIR,SAR של המקור, ההתקפה וההגנה של שיר 1

Song 2: Angels In Amplifiers - I'm Alright



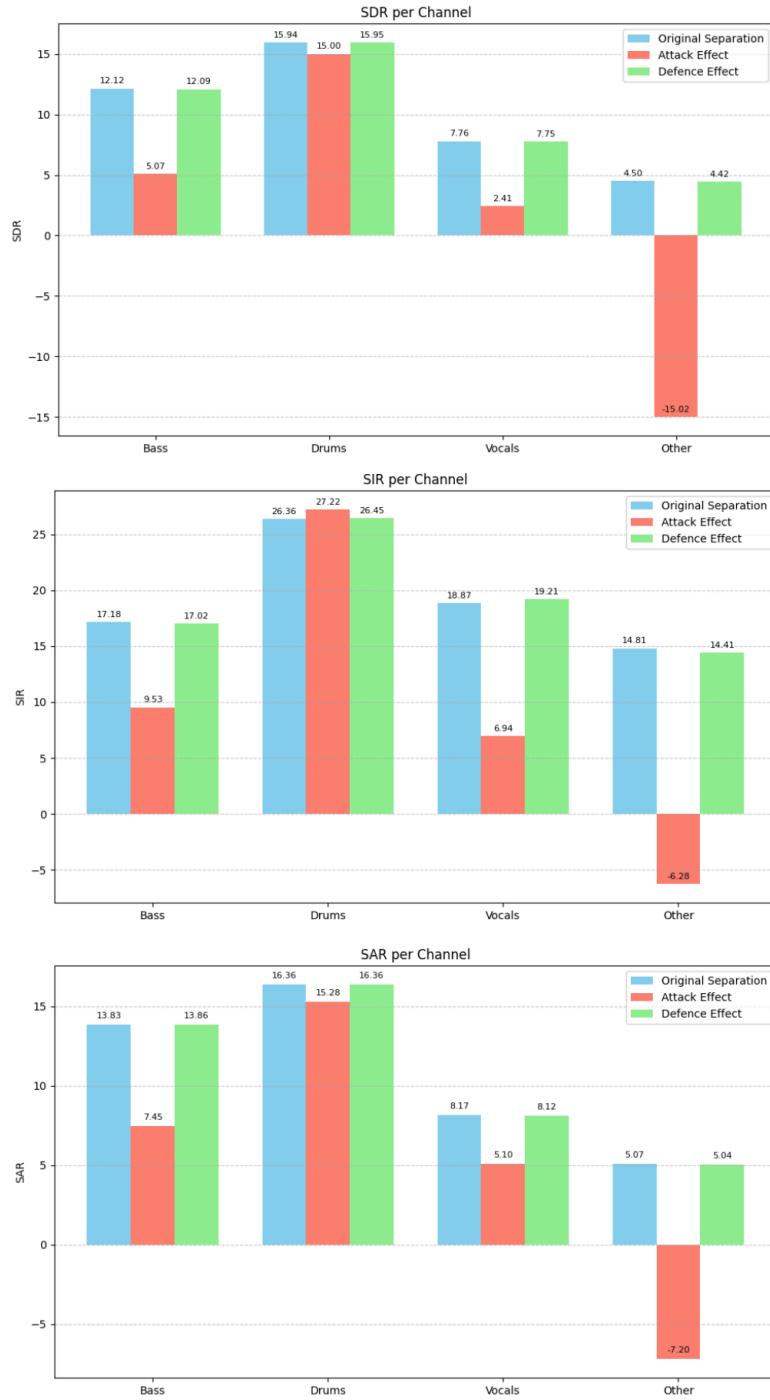
איור 9 - הציגת ביצועי SDR,SIR,SAR של המקור, ההתקפה וההגנה של שיר 2

Song 3: Ben Carrigan - Well Talk About It All Tonight



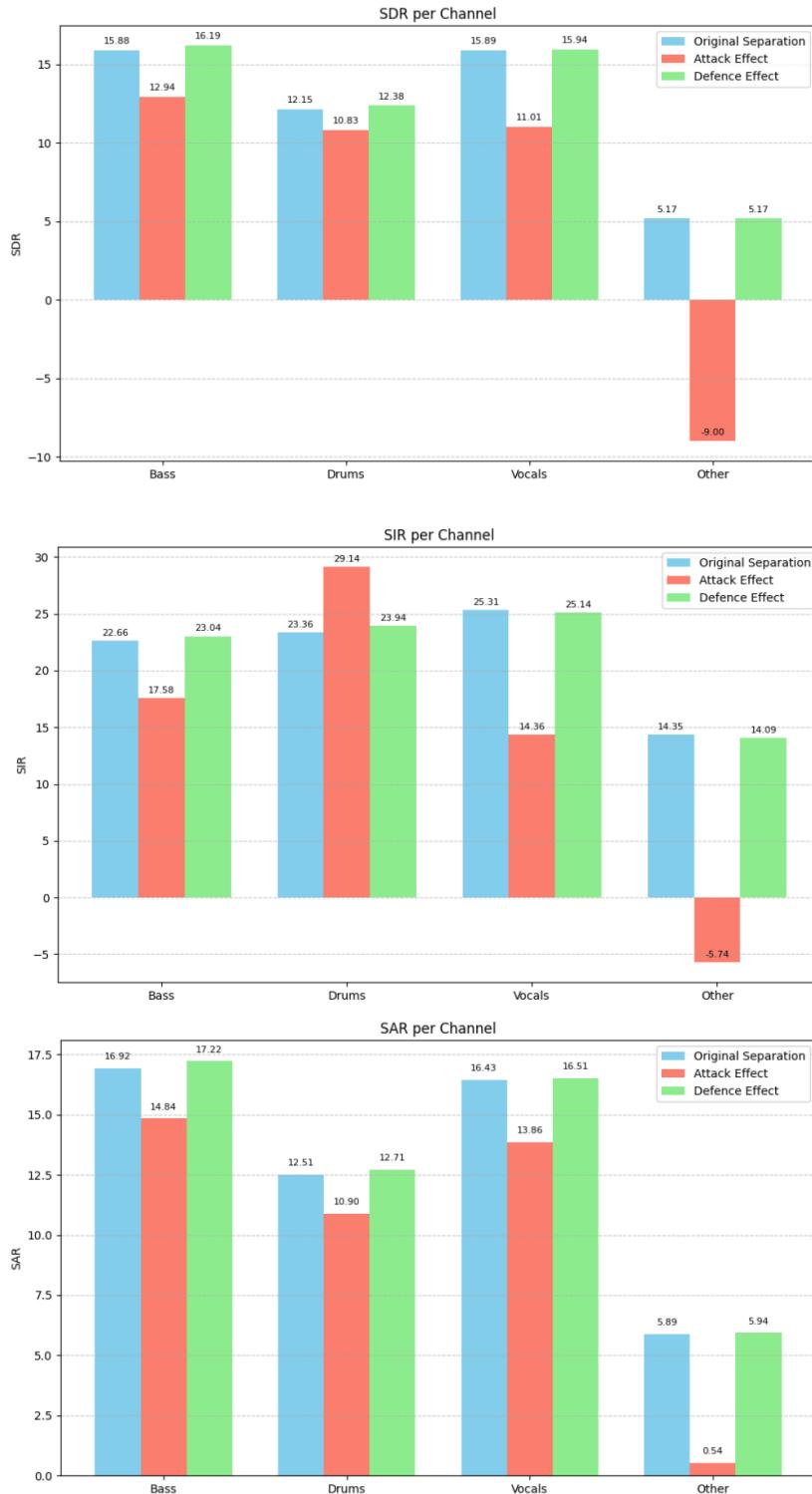
איור 10 - הציגת ביצועי SDR,SIR,SAR של המקור, ההתקפה וההגנה של שיר 3

Song 4: BKS - Bulldozer



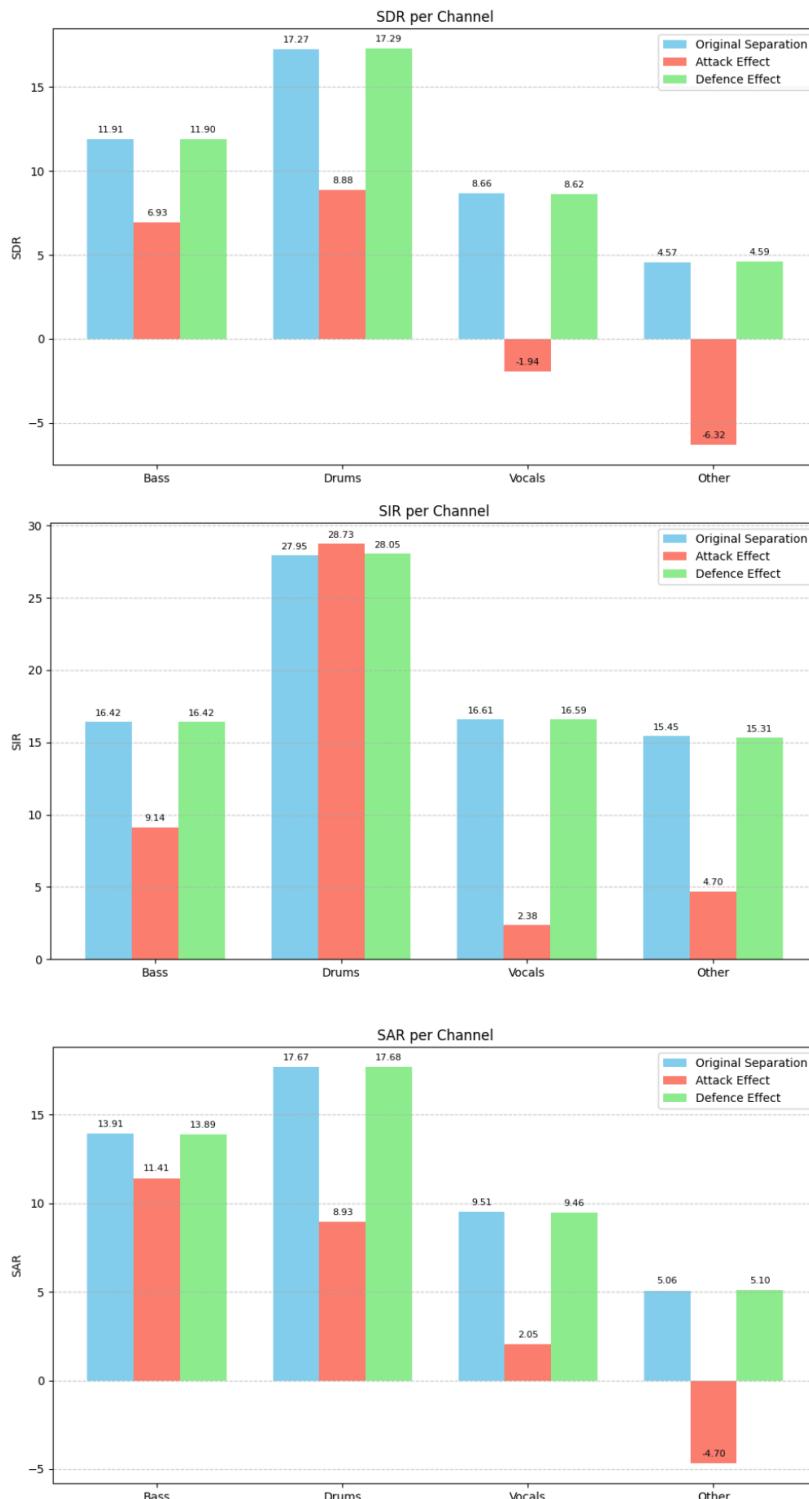
איור 11 - הציגת ביצועי SDR,SIR,SAR של המקור, ההתקפה וההגנה של שיר 4

Song 5: BKS - Too Much



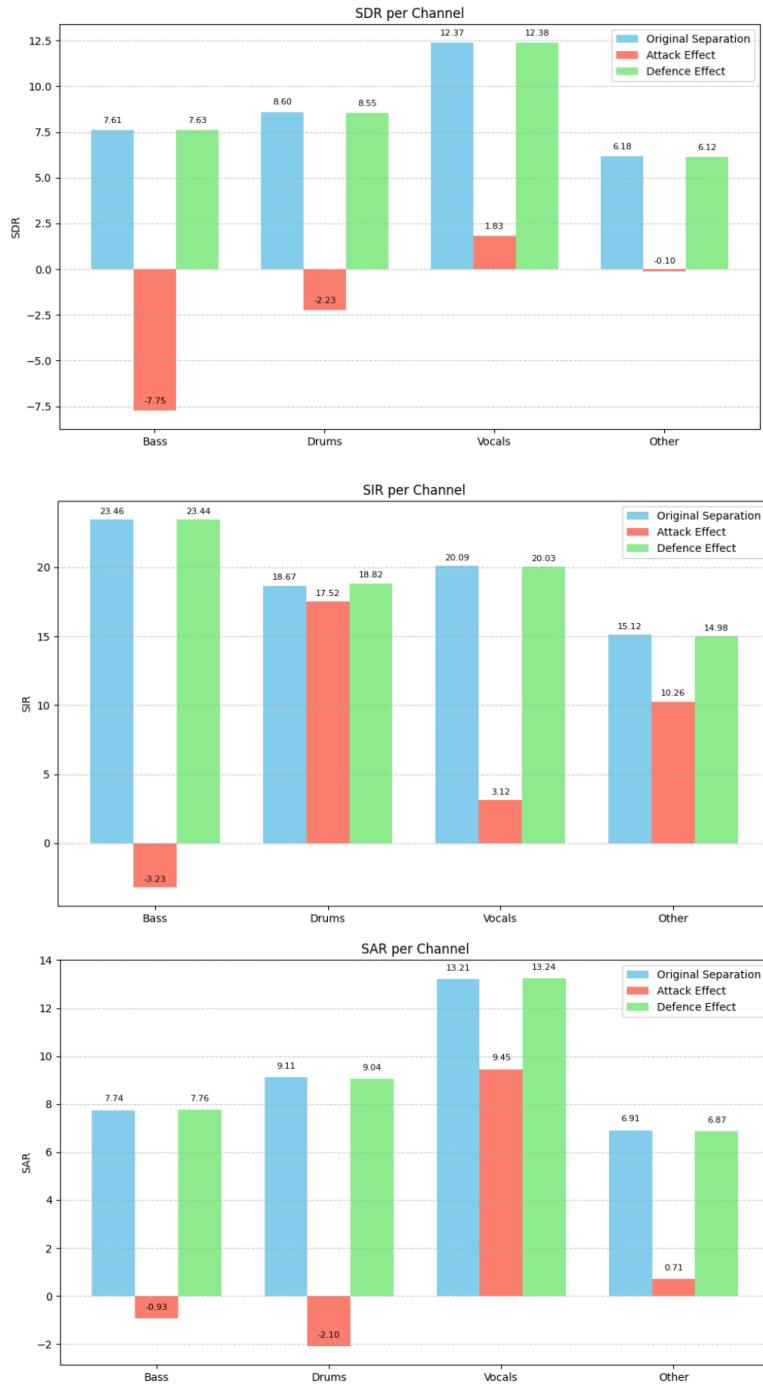
איור 12 – הציגת ביצועי SDR,S/R,SAR של המקור, ההתקפה וההגנה של שיר 5

Song 6: BKS - Buitraker - Revo X



איור 13 - הציגת ביצועי SDR,SIR,SAR של המקור, ההתקפה וההגנה של שיר 6

Song 7: Carlos Gonzalez - A Place For Us



אייר 14 - הציגת ביצועי SDR,SIR,SAR של המקור, ההתקפה וההגנה של שיר 7

10 פרק סיכום ומסקנות

10.1 ניתוח תוצאות

השוואת תוצאות הניסויים עולה בבירור כי התקיפה האדוורסריאלית משפיעה באופן נרחב על איות הפירוק שביצעו המודל. ירידת משמעות נצפתה בערכי המדדים SIR, SIR ו-SAR לאחר התקיפה, דבר המעיד על פגעה כוללת באיות ההפרדה, עלייה בהפרעות בין הערכותם, והופעת עיוויתם ועיבוד לא טבעי באות המפוץול.

לאחר הפעלת מגמת שיפור בהורה בכל המדדים – כאשר במרבית השירים, ערכי SIR, SDR, SAR התקרבו מאוד לערכים שהיו לפני התקיפה, ולעתים אף חזרו כמעט לגמרי לרמה המקורית. הדבר מצביע על יעילות גבוהה של ההגנה בנטרול השפעת השיבוש.

ניתן לראות בשירים 6-1 במטריקת ה SIR בפירוק של ערוץ other Vocals של ערוץ 1,3 ו-6 במטריקת ה SIR בפירוק של ערוץ 3 שהערך המקורי והערך לאחר ההגנה הוא חיובי, ולאחר המתקפה הערך של ה SIR הוא שלילי. מה זה אומר?

כאשר $0 < \text{SIR}$, כלומר שלילי, זה אומר שה:

- ההפרעה חזקה יותר מהסיגナル המקורי.
- המודל הוציא ערוץ שבו רוב מה שימושים כלל לא שיר לערז זה.
- **יתכן שהערוץ כמעט ריק מהתוכן הרצוי, ומלא בשאריות של ערוצים אחרים (למשל: שימושים שירה בתופים).**

במילים אחרות:

המודל נכשל בהפרדה – הוא זיהה בצורה גרואה מה שיר לערז זהה, עד שהוא "זילגא" דומיננטי יותר מהמקור.

תוצאה מעניינת ולא צפואה עלתה במדד SIR עבור ערוץ התופים, בשירים 1 עד 6: נמצא כי לאחר התקיפה – לפני הפעלת ההגנה – נשמהה אליה במדד SIR, כלומר לכארה חלה הפחתה בהפרעות מערוצים אחרים דווקא בעקבות התקיפה. תופעה זו לא הופיעה בשיר 7.

תוצאה זו עשויה להצביע על כך שההיבש האדוורסרילי, שבמכוון נועד למסם את השגיאה הכלולת, גרם למודל "להתמקד" בתדרים מסוימים באופן שהקטין את הדילפה מערוצים אחרים דווקא בערז התופים.

ichert ההשערות האפשרות להסביר התופעה היא ש-ערוץ התופים מכיל תדרים חזקים ודומיננטיים במיוחד ביחס לשאר הכלים. מאפיין זה עשוי לגרום לכך שכאר מוזרים שיבוש אדוורסרילי לפחות, הפגעה היחסית בתופים נמוכה יותר, או שהמודול "מתרבלב" לטובת הדגשת רכיבי הקצב, ובכך מופחתת זילגה מערוצים אחרים. לחלוויין, ניתן שההיבש פוגע בכל אחרים ומחליש את ההשפעות שלהם על התופים, דבר שmobiel ל-SIR גבוה – גם אם איות הצליל הכללי נפגעה.

יש לציין שמדד SIR בלבד אינו מספק את כל האспектים של איות ההפרדה, ויש לבדוק אותו בשילוב עם SDR ו-SAR על מנת להבין את מלא המשמעות של התוצאה.

תופעה זו מצביעה על חשיבות בחינה רב-מדנית של איות ההפרדה, ועל הצורך בזיהירות בפרשנות של מדד אחד בודד ללא הקשר רחב יותר.

10.2 מסקנות

מהממצאים שהתקבלו עולה המודל **Hybrid Transformer Demucs**, רגיש באופן מובהק לתקיפות אדוורסיאליות מסווג התחמקות (**Evasion Attack**).

התקפה זו, שפותחה במקור עבור מודלים בתחום ראיית מכונה (כגון סיוג תמונות), הותאמת במסגרת פרויקט זה לפועל גם על אוטות שמע רציפים, והוכחה **כיעילה במיוחד** בהשפעתה – הן על יכולת ההפרדה והן על מידת **הшибוש בתוצרי הפלט**.

המסקנה המרכזית היא כי **הפגיעה למתכונים איננה ייחודית למודלים וייזואליים**, אלא קיימת גם בתחוםים מובוסי שמע. לפיכך, קיימת חשיבות מהותית בהמשך מחקר בנושא זה, ובפיתוח מנגןני הגנה "יעודיים" גם למודלים שפועלים מחוץ לעולמות התמונה והטקסט.

תודות

ברצוננו להודות למר אורי בריט, שהנחה אותנו לאורח הפרויקט במסירות, סבלנות וראייה רחבה. תחת הדרכתו קיבלנו לא רק הכוונה מדעית מדויקת, אלא גם חופש לחקר, להתנסות ולפתח רעיונות באופן עצמאי וביקורת. התמיכה והזמיןות שלו היו עבורנו עוגן משמעותי בתהילך הלמידה, ואפשרו לנו להתמודד עם אתגרים מורכבים הן בrama הטכנית והן בrama המחקרית.

הפרויקט לא היה נראה כר' ללא הליווי המקצועי והאנושי שלו, ועל כך תודתנו העמוקה.