

המחלקה להנדסת חשמל

שם הפרויקט: הפרדת כלי נגינה וזמר/ת
מהקלטות של שירים.

Project Name: separation of musical
instruments and singer recordings of
songs.

הגדרת הפרויקט – Statement Of Work

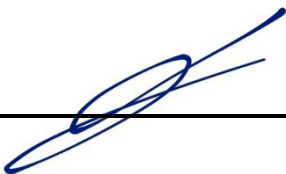
שם הסטודנט: בן ציון צוברי

מספר תעודת זהות:

שם המנחה: מר שגיא הרפז

חתימת המנחה:

תאריך ההגשה:



תוכן עניינים:

3	מבוא
3	מהות הפרויקט
4	מושגים עיקריים
4	מטרת הפרויקט העיקרית והדרך להשגתה
4	לקוחות הפרויקט ומשתמשי הפרויקט
5	מטרות הפרויקט, יעדים ומדדים
5	מטרה מרכזית
5	יעדי הפרויקט
6	סקירת ספרות
7	תרשים בלוקים של המערכת
8	חלופות
10	אמצעים נדרשים לביצוע הפרויקט
10	תוצרי הפרויקט
10	פערים
11	סיכונים עיקריים ודרכי התמודדות
11	תוכנית עבודה ראשונית
12	מקורות

מהות הפרויקט:

קיימת בעיה לבצע הפרדה של כלי נגינה וזמר/ת מהקלטות של שירים. פרויקט זה עונה על בעיה זו, במהלכו תפותח מערכת אשר תקבל בכניסתה קובץ אודיו בפורמט wav אשר מכיל שיר כלשהו, המערכת תהיה מסוגלת להפריד באופן ברור את הצלילים הנשמעים בהקלטת האודיו ל – 4 קבצים נפרדים אשר כל קובץ יכיל קול/כלי נגינה הנפרד. הפרויקט עוסק בתחומי עיבוד האותות ובשילוב טכניקות העיבוד בטכנולוגיית הלמידה העמוקה – deep learning.

Deep Learning או למידה עמוקה פותחה אי שם בשנות ה-60 כאשר המחשבה הייתה שרשת אלמנטים כאלו שמחוברים זה לזה (Feed Forward) תוכל ליצור מערכת מורכבת ועשירה ביכולות, אז בעת ההיא ניסו לתת פתרונות באמצעות רשתות אלו אך ללא הצלחה ולכן עזבו זאת.

בשנות השמונים חזרו לנושא כאשר ניסו ללמד את הרשתות לבצע משימות סיווג והחלטה באמצעות אלגוריתם ה- Back propagation ועוד כלים עשירים וחזקים אך נתקלו בבעיות אחרות ביניהם בעיות ביצועים, תחרות קשה מצד שיטות למידה שונות, מתמטיקה סבוכה וצריכת חישובים גדולה מידי, לכן הנושא נעזב בשנית.

בשנות האלפיים הנושא חזר לכותרות בהובלתם של החוקרים : Yan LeCun, Geoffrey Hinton, Yoshua Bengio, אשר הובילו סדרת ניסויים בתחום לתוצאות מעולות וזאת בזכות שני דברים חשובים שקרו במהלך הזמן:

1. כמות המידע גדלה פלאים – בשל האינטרנט, מצלמות דיגיטליות ויכולת האגירה המשתפרת.
2. מחשבים (בעיקר חומרה) השתפרו מאוד ביכולות החישוב שלהם וכך תוכנות מחשב שדרשו זמן ריצה גבוה בזמנים קודמים היו יכולים להתבצע בזמן קצר יותר באופן דרמטי. וכך, כשחזרו ואימנו רשתות לפי אותם עקרונות הרשתות הלומדות הגיעו לרמת ביצועים מעולות וחסרות תקדים.

כיום, מערכות מבוססות למידת מכונה חדרו למגוון תחומים מרכזיים בחיינו ואיננו יודעים זאת כמו למשל הפרדת מקורות (השימוש הנעשה בפרויקט זה), נהיגה אוטונומית, ניתוח טקסט ותרגום, אבחון מחלות, זיהוי עצמים בתמונות (רכבים, פנים וכו') ועוד שלל תחומים אחרים.

בפרויקט זה, נרצה לפתח מערכת אשר תקבל קובץ אודיו בפורמט Wav, המכיל שיר כלשהו ותפריד באופן ברור את הזמר/ת ואת כלי הנגינה, המערכת תחזיר 4 קבצי אודיו גם כן בפורמט Wav, קובץ אחד יכיל את הזמר/ת בלבד, השני את התופים בלבד, השלישי את הבס בלבד והרביעי את כלי הנגינה אשר משמיעים צלילים גבוהים.

על מנת לתת פתרון לבעיה זאת נשתמש ביכולותיה של טכנולוגיית ה- Deep Learning של הפרדת מקורות על מנת לתכנן מערכת שתקבל בכניסתה את השיר הרצוי והפלט ממנה תהיה 4 קבצים מופרדים.

מושגים עיקריים שיהיו בשימוש במסמך זה:

- **Machine learning (למידת מכונה)** - שיפור באמצעות ניסון, אלגוריתמי למידת מכונה נבנים על בסיס מודלים מתמטיים של מידע הנקראים "נתוני אימון", המערכת משתמשת בנתונים אלו על מנת לבצע פרדיקציות או החלטות מבלי לתכנת אותה לבצע כך.
- **Deep Learning (למידה עמוקה)** - למידה עמוקה הינה תת נושא של אלגוריתמי machine learning אשר משתמשים במספר שכבות על מנת לחלץ באופן תמידי מידע גולמי מאות כניסה כלשהו.

- **Neural network (רשת נוירונים)** - רשת נוירונים או רשת קשרית הינו מודל מתמטי חישובי הבא לדמות תהליכים קוגניטיביים או מוחיים ומשמשת במסגרת למידת מכונה. רשתות מסוג זה מכילות בדרך כלל מספר רב של יחידות מידע המקושרות זו לזו שלעיתים קרובות עוברות דרך יחידות מידע "חבויות" המקושרות בעוצמות קשר שונות זו מזו ובנוסף מכילות מידע על חוזר הקשר.
- כל "נוירון" בנוי על בסיס יחידת רגרסיה לוגיסטית/לינארית וביחד נבנים שכבות של נוירונים המתארים רשת שלמה בתצורת רשת כלשהי.
- **יחס אות לרעש ; SNR[dB]** - היחס בין עוצמת האות הרצוי (הבעל משמעות), לבין עוצמת הרעש הכולל. היחס מתאר את המידה שבה הרעש הבלתי רצוי "משפיע" על האות הרצוי ומהווה אינדיקציה עבור שימוש באות וניתוחו. באופן מתמטי, יחס אות לרעש מוגדר באופן הבא:

$$SNR = \frac{P_{Signal}}{P_{Noise}} = \left(\frac{A_{Signal}}{A_{Noise}} \right)^2$$

כאשר P הוא ממוצע ההספק של האות או הרעש ו-A הוא שורש ממוצע הריבועים של האות, באופן זה אנו נמדוד את האות ביחידות של דציבלים [dB] כך:

$$SNR_{dB} = 10 \log \left(\frac{P_{Signal}}{P_{Noise}} \right) = 20 \log \left(\frac{A_{Signal}}{A_{Noise}} \right) [dB]$$

- **Mean Opinion Score ; ציון דעה ממוצע** - קנה מידה המתאר איכות של תוצר כלשהו עפ"י הערכת דעות על מספרים תוצאים השייכות לאותה הקטגוריה על פני סקלה קבועה מראש, סכימתם ומיצוע התוצאה לפי מספר ההערכות.

$$MOS = \frac{\sum_{n=1}^N R_n}{N}$$

כאשר: R – הערכה מספרית של תוצאה אחת.
N – מספר ההערכות.

- COVL – פרדיקציה של האפקט הכללי [3].
- CBAK – פרדיקציה של פולשנות רעשי רקע [3].
- CSIG – פרדיקציה של עיוות הדיגנל המתייחס לדיבור [3].
- PESQ – הערכה תפיסתית של איכות הדיבור [3].
- SSNR – פילח האות לרעש [3].

מטרת הפרויקט העיקרית והדרך להשגתה:

מטרת פרויקט זה היא לתכנן מערכת אשר בכניסתה תקבל שיר, בתור קובץ אודיו מסוג WAV, ותדע להפרידו ל-4 מקורות נפרדים – זמר/ת, תופים, בס, צלילים גבוהים באמצעות שימוש באלגוריתמי למידה עמוקה.

באופן זה ניתן להראות את יכולתם של מערכות למידה עמוקה להתמודד עם בעיות מורכבות שמערכות למידת מכונה לא מצליחות לפתור.

לקוחות הפרויקט ומשתמשי הפרויקט:

תוצר הפרויקט רלוונטי עבור כל אדם אשר עוסק בעריכת וידאו עמו, די-ג'יי, מפיק, מוזיקאי וכו'. כיום קיימות בשוק תוכנות ייעודיות לעריכת וידאו אשר מבצעות את ההפרדת המקורות (כמו XTRAX STEMS של Audionamix), בדרך כלל יקרות לשימוש חודשי ויש ללמוד כיצד לעבור איתן.

מטרות הפרויקט, יעדים ומדדים:

מטרה מרכזית:

מטרת פרויקט זה היא לתכנן מערכת אשר בכניסתה תקבל שיר, בתור קובץ אודיו מסוג WAV, ותדע להפרידו ל-4 מקורות נפרדים – זמר/ת, תופים, בס, צלילים גבוהים.

יעדי הפרויקט:

1. **היעד:** יכולת הפרדה ברזולוציה גבוהה, על מנת להבחין באופן ברור שקיימת הפרדה בין השיר לסיגנל הרצוי.

המדד: נדרוש יחס אות לרעש (SNR) של לפחות 20[dB] על מנת לקבל את ההפרדה הרצויה.

כדי לחלץ את ה-SNR של הסיגנל המופרד נשתמש בתיאור הספקטרלי שלו, בתיאור זה יהיו קיימים כל התדרים הרצויים לעומת התדרים הלא רצויים בעוצמות שונות ובאמצעות חישוב היחסים באופן הבא: $SNR = \frac{P_{signal}}{P_{noise}} [dB]$ נוודא שהוא נמצא בתחום הרצוי לקבלת רזולוציה טובה.

2. **היעד:** הסיגנל ב-Output יהיה איכותי לאחר ההפרדה למען המשתמש.

המדד: על מנת לבצע מדד על איכות הסיגנל (מציאת מדד איכותי) נשתמש בשיטת MOS (Mean Opinion Score), אשר משתמשת בדירוגים של התוצאות השונות שיתקבלו בסקלה מסויימת ועל פי הדירוג הממוצע של כולם, התוצאה תצביע על איכות ההפרדה של האלגוריתם.

סקירת ספרות:

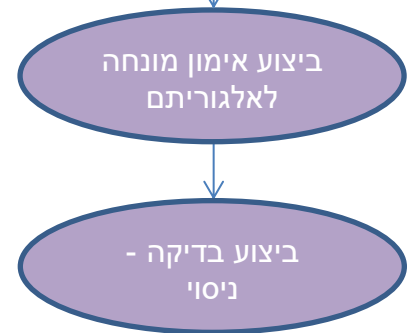
המאמר עוסק בחקר הפוטנציאל לתכנון מערכת מפרידת מקורות מקצה לקצה באופן מלא, פירוט הפתרון המוצע עבור מערכת מסוג זה והשוואתו למערכות אחרות מאותו התחום. עפ"י החקר שהתבצע הוצע הפתרון של רשת ה-Wave-U-Net אשר מפרידה מקורות באופן ישיר בתחום הזמן ומעבירה כמות מידע גבוהה. המערכת המתוכננת בפרויקט זה תתבסס על פי הרשת המוצעת במאמר [1].

מאמר המתמקד בהפרדת 4 סוגי סיגנלים שונים: זמר/ת, תופים, בס ואחרים מתקליטים מז'אנרים שונים בעזרת אלגוריתמי Deep Learning (בעיקר בתחום ה-CNN – convolutional neural networks). הכיוון במאמר מוצג כאלגוריתמי State-Of-The-Art (העדכניים ביותר) שבאמצעותם ניגשים לפתרון הבעיה המוצגת [2].

מאמר זה חוקר את הפתרון המוצג Wave-U-Net [1], סוקר פתרון זה ומציג את יישומיו ועורך השוואה ע"י ניסויים מול אלגוריתמים אחרים שעוסקים בבעיה (Noisy, Wiener, SEGAN) על בסיס מטריקות שונות (PESQ, CSIG, CBAK, COVL, SSNR) [3].

תרשים בלוקים של המערכת:

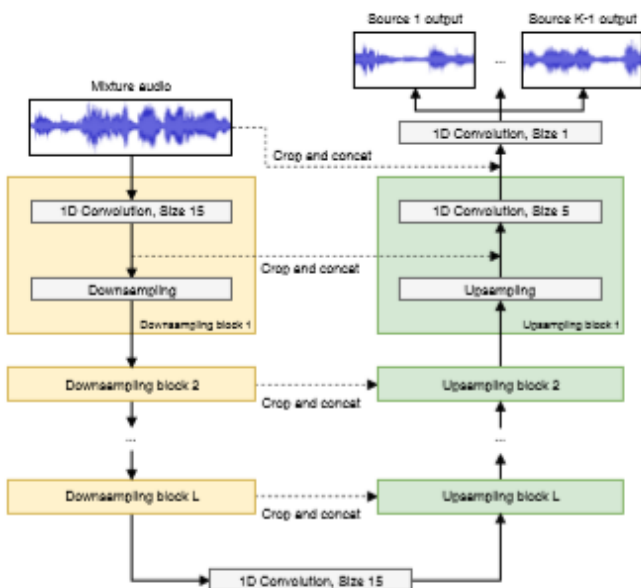
בניית אלגוריתם מבוסס למידה עמוקה אשר יקבל שיר כקלט ובמוצאו יתקבלו 4 סיגנלים שונים.



הכנסת קלט – קובץ אודיו
המערכת תקבל בכניסתה קובץ אודיו (WAV) אשר מכיל את השיר אותו אנו רוצים להפריד לסיגנלים שונים.

פלט – הפרדת סיגנלים.
הבלע 4 סיגנלים מורדים במוצא המערכת לפי ההפרדה הבאה: זמר/ת. תופים. בס. גבוהים.

להלן תרשים בלוקים עקרוני של רשת Wave-U – Net [1]:



- האלגוריתם מקבל בכניסתו את השיר המורכב מכל הסיגנלים אותם נרצה להפריד.
- הסיגנל עובר בשכבות הרשת – עבור המערכת שלנו תהינה 4 שכבות כאשר כל שכבה מתמקדת בסיגנל מסויים.
- K – מייצג את מספר הסיגנלים במוצא.
- L – מייצג את מספר שכבות הרשת.

חלופות:

חלופה מערכתית:

תוכנת עריכת אודיו אל מול מערכת AI המבצעת את אותה הפעולה.

משקל	מערכת AI	תוכנת עריכת אודיו	מדד / חלופה
0.3	אוטומטי (5)	ידני (3)	ידני/אוטומטי
0.15	דיוק של האלגוריתם (3)	תלוי תוכנת אודיו (4)	דיוק
0.25	מתבצע באופן אוטומטי ומהיר (5)	דורש זמן רב מהמשתמש (2)	מהירות
0.3	(2)	תלוי תוכנת אודיו (4)	איכות
	3.8	3.2	ציון משוקלל

חלופה נבחרת: מערכת AI.

תוכנת עריכת אודיו היא ידנית ולא אוטומטית, לא כל אחד יודע להשתמש בתוכנות אלו וזה נסיון שצריך לרכוש ודורש זמן רב, גם לאחר רכישת הניסיון לוקח זמן לבצע את ההפרדה בפועל.

לעומת זאת מערכת AI פועלת באופן אוטומטי ואינה דורשת נסיון בהפעלתה, קיים Input וכך גם Output בהתאם, היא פועלת באופן מהיר וחוסכת זמן לעומת זאת היא פחות מדויקת מתוכנת עריכה קיימת, בהתאם לכך נבחרה חלופת מערכת AI.

בחירת GPU:

משקל	NVIDIA GTX 1070TI	NVIDIA RTX2070	NVIDIA-RTX2060	מדד / חלופה
0.3	~2700 (2)	~2000 (4)	~1700 (5)	מחיר [ש"ח] (ציון)
0.4	8 (4)	8 (4)	6 (2)	RAM [GB] (ציון)
0.1	1607 (5)	1410 (4)	1365 (3)	base clock [Mhz] (ציון)
0.2	2432 (5)	2304 (4)	1920 (3)	Num of cores. (ציון)
	3.7	4	3.2	ציון משוקלל

חלופה נבחרת: Nvidia RTX2070.

יתרונותיה של החלופה הנבחרת לעומת האחרות בעיקר מתבטאת ביחס עלות מול תועלת מיטבי עבור תכנון והרצת אלגוריתמי אימון למערכת הלמידה בעמוקה, 8GB של זכרון RAM מתאים מאוד למערכות למידה עמוקה מתחילות ובהתאם לגודל הזכרון כך יאפשר להכניס יותר מידע לאימון באיטרציה אחת, כך גם כמות הליבות הקיימות בכל GPU, ובהתאם למחיר של החלופה – Nvidia RTX2070 נבחרה.

האמצעים/הכלים הנדרשים לביצוע הפרויקט:

כמו שנאמר במסמך זה, פרויקט זה עוסק בכל תחום הלמידה העמוקה, תחום זה לא צלח בעבר מכיוון שלמחשבים של פעם לא היה את כח החישוב הנדרש על מנת להריץ אלגוריתמים של למידה עמוקה, לכן על מנת ליצור מערכת שכזו יש צורך בחומרה המתאימה לכך.

- GPU (Graphical Processing Unit) – אלגוריתמי Deep Learning דורשים מהמחשב כח חישוב רב מאוד בביצוע פעולות מתמטיות רבות מאוד בהרצת הקוד / אימון המערכת, על מנת לבצע את התהליכים הנ"ל באופן מהיר ומקבילי נדרשת חומרה חזקה שתוכל לבצע את הכמות חישובים הנדרשת בזמן סביר ו- כרטיס GPU עונה על בעיה זו, לדוגמא, אלגוריתם VGG16 (אלגוריתם של רשת נוירונים המבוסס על 16 שכבות חביות וקונבולוציות ביינהן) יש כ- 140 מליון פרמטרים (משקלים לאימון המערכת) שאיתם יש לבצע פעולות מתמטיות לאימון המערכת.

הכרטיס GPU הנבחר הינו GeForce RTX2070, שוויו עומד על כ- 2000 ₪ וביצועיו ברמה גבוהה מאוד, חשוב לציין שזכרון ה-RAM של כרטיס זה עומד על 8GB אשר מספק לכל מודל סטנדרטי של Deep Learning מבחינת גודל זכרון, בנוסף ניתן לבצע אימון ב- 16bit וכך "להכפיל" את הזכרון הנצרך בפועל.
(<https://timdettmers.com/2019/04/03/which-gpu-for-deep-learning/>)

- RAM – על מנת להמנע מריצה איטית של הקוד יש לדרוש כחלק מהחומרה לגודל זכרון RAM, ככל שיהיה יותר זכרון כך הקוד ירוץ יותר מהר וימנע מתקיעות מיותרות.

גודל הזכרון המומלץ עבור אלגוריתמי למידה עמוקה הוא 32GB, עבור פרויקט זה נסתפק ב- 16GB.

תוצרי הפרויקט:

תוצר פרויקט זה יכיל ערכת תוכנה מותקנת PC.

פערים:

פערי ידע קיימים עבור פרויקט זה שכן תחום ה- Deep Learning נלמד באופן עצמאי וכולל בתוכו נושאים שלא נלמדו כלל כמו Python, מבני רשתות נוירונים, אלגוריתמיקה של למידה עמוקה וכו'. פערים אלו ניתן להשלים באמצעות ספרות מתאימה והאינטרנט, Python הינה שפה קלה ללמידה כאשר קיים ידע בסיסי בתכנות (כמו c, c++, matlab וכו) וניתנת להשלמה במהלך הפרויקט, אמנם נושא הלמידה העמוקה יותר קשה להבנה וללמידה ולכן אעזר בקורס אינטרנטי בשם Data Science: Deep Learning in Python אשר כולל בתוכו את כל החומר הפרקטי והרקע המתאים ליישום בניית התוכנה המתאימה לפרויקט.

פער נוסף הוא הפער החומרתי, כמו שנאמר לעיל יש להצטייד המחשב עם GPU מתאים על מנת לאמן את הרשת, GPU שכזה ניתן להשיג בכל חנות מחשבים בבניית מחשב נייח או כהזמנתו כרכיב בודד.

סיכונים עיקריים ודרכי התמודדות:

- כמות המידע המועברת באלגוריתם גדולה מידיי לכדי שה-GPU יוכל להתמודד איתה ולכן ייקח זמן רב לאמן את המערכת, כדי להתמודד עם סיכון זה נבחר רכיב GPU מומלץ להרצת הקוד כך שהסיכון יורד משמעותית.
- השלמת פערים – לימוד של חומר חדש באופן עצמאי והבנתו לעומק, ניתן סיוע מאנשים העוסקים בנושא ובנוסף קורס אינטרני המלמד את הבסיס לתכנות Deep Learning ב-Python.

תוכנית עבודה ראשונית:

תאור מטלה	זמן ביצוע משוער (ימים)	תאריך סיום ביצוע	תאריך הגשת מטלה
SOW	בחירת הפרויקט	1	22.2.20
	פגישה עם המנחה	1	4.3.20
	השלמות פערים וחקר	ביצוע שוטף	
	פגישה עם המנחה	1	1.7.20
	כתיבת SOW	7	13.07.20
דוח התקדמות	בניית הקוד	30	10.8.20
	כתיבת דוח ההתקדמות	10	20.8.20
	פגישה עם המנחה	1	24.8.20
	השלמות ותיקונים לדוח	5	29.8.20
דוח ביניים	המשך בניית הקוד	45	15.12.20
	ניתוח תוצאות ראשוני	5	20.12.20
	כתיבת דוח הביניים	10	30.12.20
	פגישה עם המנחה	1	3.1.21
	השלמות ותיקונים לדוח	5	8.1.21
דוח סופי ומסירת הפרויקט	שדרוג ואופטימיזציה לקוד	45	14.3.21
	כתיבת הדוח הסופי	15	10.4.21
	פגישה עם המנחה	1	13.4.21
	השלמות ותיקונים לדוח	10	23.4.21
	מסירת הפרויקט	1	25.5.21

- [1] Tim Dettmers, Which GPU(s) to get for deep learning: My experience and advice for using GPU's in deep learning, <https://timdettmers.com/2019/04/03/which-gpu-for-deep-learning/>, published at 2019-04-03.
- [2] Daniel Stoller, Sebastian Ewert, Simon Dixon, Wave-U-Net: a multi-scale neural network for end-to-end audio source separation, submitted at 2018-06-08.
- [3] Craig Macartney, Tillman Weyde, Improved Speech Enhancement with the Wave-U-Net, Submitted at 2018-11-27.
- [4] Pritish Chandna, Audio Source Separation Using Deep Neural Networks, submitted at 2014.