Speech To Sing

ספר פרויקט - פרויקט מסכם תואר, הנדסת חשמל משולב עם מוזיקה

מאת: אהרון טאוב , עמית אליאב

מנחים: פרופ' שרון גנות, רננה אופוצ'נסקי
הפקולטה להנדסה, בר אילן

מספר קורס 83-401
שנת 2020-2021

תוכן עניינים

1. תקציר
2. תודות
3. הצגת הבעיה:
   - 3.1. דוגמה - החלפת סגנון לתמונה
   - 3.2. האתגרים בפתרון בעיה
4. רקע עיוני:
   - 4.1. טון בסיס (F0) והרמוניות באותות אקוסטיים
   - 4.2. פורמנטים
   - 4.3. מקדמי LPC
   - 4.4. הבדלים בין דיבור ושירה, וכיצד הם משתקפים בפורמנטים ובספקטוגרמות.
   - 4.5. יצירת וקטור embedding לדובר
5. הכלים הקיימים טרם העבודה על הפרויקט:
   - 5.1. כליי text-to-speech
   - 5.2. כליי החלפות דוברים
   - 5.3. כלים של DeepFake
6. תיאור של חלופות לפתרון:
   - 6.1. חלופות לפתרון - שיטות **קלאסיות**:
     - 6.1.1. חלופות לפתרון - שיטות קלאסיות - חלופה א - סינתזה בעזרת מקדמי LPC
     - 6.1.2. חלופות לפתרון - שיטות קלאסיות - חלופה ב - Formant Domain Adaptation
   - 6.2. חלופות לפתרון - שיטות של **Deep Learning**
     - 6.2.1. חיפוש Dataset מתאים
     - 6.2.2. חלופות לפתרון - שיטות של Deep Learning - חלופה ג - Phoneme Learning
7. **תיאור מפורט של השיטה הנבחרת והסיבות לבחירתה.**
   - 7.1. הקדמה
   - 7.2. הכנת הדאטה
   - 7.3. תיאור השיטה
   - 7.4. מבנה ה-encoder-decoder
   - 7.5. אופן הלמידה של הרשת
   - 7.6. הסיבות שבגללן בחרנו לבצע שינויים ברשת
   - **7.7. השינויים שלנו ברשת המתוארת במאמר:**
     - 7.7.1. החלפת ה-encoder של הסגנון - ה-embedding:
     - 7.7.2. החלפת ה-vocoder
     - 7.7.3. שימוש ב-stft ולא ב-mel-spectrogram
     - 7.7.4. אימון על Datasets נוספים:
       - 7.7.4.1. אימון על מאגרים קיימים
       - 7.7.4.2. יצירת מאגר שירים חדש ואימון מכונה
   - 7.8. שינוי ובדיקת היפר-פרמטרים, ותצורות שונות לאימון הרשת:
     - 7.8.1. בדיקת מימדים שונים לצוואר הבקבוק:
     - 7.8.2. גדלים שונים של ספקטוגרמות
     - 7.8.3. סוגי דאטה שונים, והצלבות ביניהם
   - 7.9. ניתוח התוצאות שלנו
8. מסקנות וסיכום
9. רעיונות להמשך
10. ביבליוגרפיה
11. נספחים

1. **תקציר:**

פרויקט זה הינו הראשון מסוגו, פרויקט המשלב עקרונות וכלים הנדסיים יחד עם מאפיינים ושימושים מוסיקליים.
מטרת פרויקט זה : "Speech-to-Sing" הוא ליצור מערכת שלמה, מהקצה לקצה, אשר תוכל לקבל שיר מקורי, והקלטה של דובר חדש מדבר במשך פרק זמן קצר, כך שהמערכת תוציא קובץ שבו נשמע הדובר החדש שר את השיר המקורי.
בספר פרויקט זה נציג רקע לבעיה שאנו מנסים לפתור - שיטות קיימות לפתירת בעיות דומות וכן התקדמויות שנעשו אף בניסיון פתירת הבעיה הזו. נציג גם את האתגרים שהיו המכשולים העיקריים שלנו במהלך פתרון הבעיה . לאחר מכן, נפרט את השיטות הראשונות והכיוונים בהם בחרנו בתחילת הדרך - כיצד המחקר וההתעסקות בכל אחד מפתרונות הביניים קידם אותנו בדרך אל היעד. לבסוף נציג את הפיתרון הנבחר שלנו.
נתעמק בו ונראה כיצד גם לאחר בחירת השיטה הנוכחית נאלצנו להמשיך לחקור ולבדוק את צעדינו על מנת לוודא שאנו מתקדמים בכיוון הנכון. נציג תוצאות ונשווה לשיטות חלופיות הקיימים היום.
נסכם את ספר הפרויקט עם הממצאים שלנו ועם המסקנות שהסקנו במהלך העבודה על הפרויקט, וכן תובנות לעתיד.

2. **תודות:**

ראשית, אני חייבים תודה למנחים המקצועיים אשר ליוו אותנו במהלך כל השנה האחרונה: פרופ' שרון גנות, וגב׳ רננה אופוצ׳ינסקי. מלבד התמיכה והייעוץ במהלך השנה, אנו מעריכים מאוד את המוכנות של פרופ׳ גנות אשר קיבל את הבקשה שלנו לבחור פרויקט התואם את המסלול המשולב של הנדסת חשמל ומוסיקה, ואף אישר לנו לעסוק בפרויקט שאנחנו הצענו. באופן דומה, גברת אופוצ׳ינסקי, שלמרות שלא הייתה בקיאה בתחום המוסיקלי ששילבנו במהלך הפרויקט, הייתה מוכנה לקבל אותנו כפרויקטנטים שלה והצליחה להביא את הידע שלה ברשתות ובלמידת המכונה לידי ביטוי בסיוע ובהנחיית הפרויקט.
נודה גם כן לד״ר שי כהן, ששמח לעזור לנו כבר בתחילת הדרך, ייעץ לנו רבות והכוון אותנו בדרכים שלא חשבנו עליהם מלכתחילה.
לבסוף נציין גם כן את פרופ׳ יואל גרינברג, ראש המחלקה למוסיקה, ד״ר לאה סילבר ומר פיני טדייטניק שגם כן סייעו לנו במהלך הפרויקט.
לא פחות חשוב, לעמית ולאהרון יש אחים ובני משפחה שמכירים קצת את התחום, יש מעט ניסיון או פשוט יצירתיים. להם, ובכלל לכל בני המשפחה אנחנו אסירי תודה שתמכו בנו במשך השנה ובכלל לאורך כל התואר. אין ספק שלא היינו מצליחים בלעדיהם.

3. **הצגת הבעיה:**

בתור שני הסטודנטים היחידים מהשנה שלנו בבר אילן, הלומדים במסלול המשולב של הנדסת חשמל ומוסיקה, ראינו בכך הזדמנות לעשות פרויקט המשלב את שני התארים האלה. כבר ראינו בקורסים שונים במהלך השנים הראשונות שלנו כיצד יש חפיפה יותר ממה שנראה לעין בין הנדסת חשמל ומוסיקה, וככל שהתקדמנו והתמקצענו בשני התחומים הללו גילינו עד כמה הקורולציה בין שני התחומים אכן חזקה.
לכן, כחלק מההכנה להצגת הפרויקט, התעמקנו במחקרים שונים המשלבים את שני תחומים אלה, חיפשנו בעיה שמעניינת אותנו, שמתאימה מבחינת רמת הידע והזמן שהיה לנו בתחילת השנה, וגם תואמת את הדרישות של הפקולטה עבור פרויקט גמר בהנדסת חשמל.

מתוך רשימת הרעיונות שהצענו, הרעיון לפרויקט שנבחר נקרא: **"Speech-to-Sing"**.
הרעיון הוא ליצור אפליקציה (מערכת כלשהי), אשר מקבלת קובץ של שיר כלשהו, ובנוסף גם קובץ של המשתמש בו הוא נשמע מדבר במהלך פרק זמן קצר, ותוכל לעבד את הקלטים ולהוציא כפלט קובץ בו נשמע המשתמש שר את השיר הנבחר. נציין שהקלט המקורי של המשתמש (הדובר החדש) אינו כולל בתוכו שירה אלא דיבור בלבד, ולכן דרושה היכולת ללמוד את השירה של המשתמש מתוך קובץ דיבור בלבד.
ישנן כיום אפליקציות דומות אשר מתעסקות בתחום זה, מה שלרוב נקרא ***Style Transfer***. ברשתות כאלה ישנם שני קלטים, אחד בתפקיד התוכן (content) והשני בתפקיד הסגנון (Style) והמטרה היא לייצר פלט בו מקבלים את התוכן בסגנון החדש ולא בסגנון המקורי.

נציג דוגמה המבהירה את מושג ה-Style Transfer שהזכרנו:

3.1. <u>דוגמה - החלפת סגנון לתמונה:</u>

בדוגמה זו יש שתי תמונות שנקלטות במערכת - אחת בתפקיד התוכן והשנייה בתפקיד הסגנון. המטרה היא ליצור תמונה חדשה שנראית כמו התמונה הראשונה בסגנון התמונה השנייה.

שתי התמונות המוצגות למטה הן הקלטים למערכת. התמונה מימין (התמונה של האבא עם הילדה) היא בתפקיד התוכן , התמונה משמאל (הצבעונית) בתפקיד הסגנון. הרעיון הוא ליצור רשת לומדת אשר יודעת להבין ולהבדיל בין תוכן לסגנון של תמונה - דבר שנפרט על כך בהמשך. לאחר שההפרדה נעשית בהצלחה, ניתן ליצור תמונה חדשה אשר משלבת בין 2 התמונות.

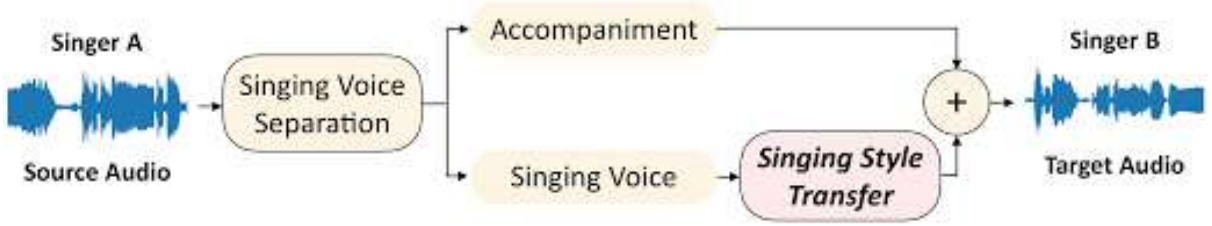
במקרה הזה, השתמשנו בשיטות קיימות המבצעות את החלפת הסגנון בין התמונות[1]:

ולאחר שהרצנו אצלנו את המכונה עם שני הקלטים לעיל, זו הפלט שקיבלנו:

[1] arXiv:1508.06576 [cs.CV]ת "A Neural Algorithm of Artistic Style",,
https://github.com/lengstrom/fast-style-transfer

דוגמא זו ממחישה את הרעיון של הפרדת התוכן מהסגנון, והחלת סגנון מסוים על תוכן אחר לקבלת תוצר חדש לגמרי. באופן דומה, לאחר שראינו דוגמאות בסגנון הזה עבור תמונות, חשבנו שנרצה למצוא דרך דומה לבצע style-transfer עבור אודיו. בהמשך ספר הפרויקט נסביר מה מצאנו בנושא.

מבלי לפרט את דרך הפתרון בצורה מעמיקה, ניתן בכל זאת לתאר את הקווים המנחים ואת השלבים שאנו רואים שהמערכת תצטרך לכלול. ניתן לראות בתרשים הבא:



לפי התרשים לעיל, ניתן לראות כי נכנס קטע שמע למערכת. ויתבצעו מספר פעולות:

1. הפרדת השירה מהליווי המוסיקלי. זה השלב הראשון כיוון שברור היה לנו שהעיבוד על החלק של השירה צריך להיעשות בתנאים סטריליים ככל שניתן (כלומר אות דיבור/שירה נקי ללא ליווי מוזיקה).
2. נבצע את חלק החלפת סגנון השירה (שישמע כמו הדובר החדש) - זה יהיה לב הפרויקט, היכולת להמיר שירה של אדם אחד לשירה של אדם אחר.
3. חיבור מחדש של השירה והליווי.

חשוב לציין כי בשלב זה טרם היה ברור לנו באיזה חלק של התרשים נכון להכניס את הקלט של הדובר החדש מדבר. בהמשך נפרט גם על החלופות הקיימות השונות והיכן כל אחד מהם ממקם את הכנסת סגנון היעד של הפלט נמצא.

כמו כן, מכיוון שמערכת כוללת כזו דורשת עבודה רבה, ידענו מראש ובתיאום עם המנחים שנשתמש גם בתוכנות וכלים קיימים (למשל בחלק הראשון - כלים להפרדה בין שירה ומוזיקת ליווי).

3.2. <u>האתגרים בפתרון בעיה:</u>

מהסתכלות ראשונית, ניתן להתחיל לראות מהם האתגרים בניסיון למצוא פתרון לבעיה:

<u>אתגרים טכניים:</u>

1. היכולת להפריד באופן חד וברור בין השירה לבין החלק הליווי של השיר. התמודדות עם אתגר זה מאוד משמעותית כיוון שבמידה וההפרדה לא תתבצע בצורה טובה מספיק, חלק הליווי של ההפרדה יכלול את השירה המקורית מה שיפגע באיכות הפלט לאחר שנחבר את השירה החדשה יחד עם הליווי שהפרדנו מהשיר המקורי. כמו כן, היינו צריכים לחשוב איך אנחנו מתמודדים עם הפרדה מזמרי רקע וקולות אנושיים אחרים שאינם בתפקיד הזמר הראשי (שאותם כנראה נרצה להשאיר כמו שהם במקור).
2. על אף שכן נעשה מחקר ופורסמו עבודות בנושא החלפת סגנון דיבור, לא הייתה כמעט התייחסות להחלפת סגנון עבור קטעי שירה. כפי שנציג בחלק התיאורטי, ישנם הבדלים ניכרים ומשמעותיים מאוד בין דיבור לשירה, הבדלים אלה גרמו לנו לחשוב ולהתאים את המערכת שלנו בהתאם.
3. במוסיקה, לעומת דיבור רגיל וטבעי של בני אדם, ישנה חשיבות יתרה עבור מקצב השירה והטמפו על מנת שיתאים לליווי המוזיקלי. אם בהחלפת סגנון עבור דיבור רגיל ניתן לאפשר גמישות בתזמון ובקצב המילים. בהחלפת סגנון שירה אין אפשרות להשאיר מקום לגמישות כזו - ויש צורך בהתאמה מדויקת של המילים לשיר המקורי.
4. אתגר נוסף שנאלץ להתמודד איתו הוא מציאת מאגרי מידע בתחום זה. בהנחה שחלק מהמערכת יכלול היבט של למידה אזי מאגרים של שירה יהיו הכרחיים להצלחת הפרויקט. גילינו שישנם מאגרים רבים עבור דיבור (ועוד יותר עבור תמונות), אבל מאגרים המיועדים ללמידה של שירה הם מאוד מוגבלים, הן מבחינת הכמות והן מבחינת מבנה המאגר (מבנה המאגר, קבצים נלווים וכו). בנושא זה נרחיב בהמשך.
5. <u>ההבדל בין עיבוד דיבור/שירה ועיבוד תמונה:</u>

אחד הכיוונים עליהם חשבנו הוא בהשראת רשתות הפועלות על תמונות, כמו שהצגנו את הדוגמה בה ביצענו החלפה של סגנון. אולם לפי דעתנו ישנם מספרים גורמים משמעותיים המבדילים בין עיבוד תמונה ודיבור, אשר גורמים לעיבוד דיבור להיות קשה יותר:

a. האוזן האנושית הרבה יותר רגישה לרעשים ולטעויות מאשר מה שהעין יכולה לסבול עבור תמונות. תמונה עם שגיאות רבות עדיין תתקבל באופן יותר טבעי אצל המסתכל מאשר קטע קול מעוות.
b. לרוב כמות המידע שמתעסקים איתו בעיבוד אודיו מכיל הרבה יותר משקל בקטע קול מאשר בתמונה.
c. הרבה יותר מחקרים ועבודות (בתחום ה- Deep) כבר נעשו בעיבוד תמונה ובפרט בהחלפת סגנונות אומנותי של תמונות מאשר של דיבור.
d. דיבור הוא אות המתפתח בזמן, זאת לעומת תמונה עליה ניתן להתבונן בבת אחת ולקלוט את כל המידע. אות הדיבור הוא כמו שוידאו הוא אות מתפתח בזמן לעומת תמונה שהיא סטטית.

<u>אתגרים נוספים ושאלות נוספות שעלו עם בחירת רעיון זה לדוגמא:</u>

1. האם נרצה לתת למשתמש אפשרות לבחור כל שיר שהוא ירצה עבור קובץ תוכן או שאנחנו מגבילים לבחירת שיר מתוך סט שירים קיימים?
2. מהי הדרישה שלנו מהדובר החדש על מנת שנוכל להחליף את הסגנון באופן מספיק טוב? האם כל קובץ דיבור מתאים או שאנחנו דורשים טקסט ספציפי שצריך להיאמר על ידי המשתמש? כמה קצר קטע זה יכול להיות?
3. בהנחה שאנחנו רוצים שאכן ירצו להשתמש במערכת זו, הבנו שכל התהליך צריך להתקיים כמה שפחות זמן שניתן, ולכן עולה השאלה מהי הארכיטקטורה על מנת להשיג תוצאה בזמן מינימלי?

ולבסוף עלו גם שאלות מהותיות לגבי הפרויקט ומהי החזון שלנו בעבודה זו:

1. כאשר אנו רוצים ללמוד את סגנון הדיבור/ השירה מתוך המשתמש, מה זה כולל? האם זה רק גוון הקול? האם זה כולל גם כן גמגומים, ליקויי הגייה המתבטאת בעיצורים שונים, קצב דיבור אופייני לדובר ועוד מאפיינים נוספים שניתן לייחס לסגנון הדיבור של הדובר?
2. האם במידה ובמצב אמת המשתמש היה בעל יכולות שירה גרועות (קול לא נעים, זיופים, בעיות בקצב ועוד) נרצה לייצר פלט התואם את היכולות הגרועות האלו או דווקא לאפשר למשתמש לשמוע גירסה מקצועית ומשופרת שלו שר?
3. לאלו שימושים אנו מייעדים את הטכנולוגיה הזו? האם רק למטרות הנאה משחק ולמידה או שישנם יישומים אחרים אשר יהנו מיכולות כאלה ובתחומים אחרים גם כן?

ככל שהתקדמנו בפרויקט, היה ברור לנו שנגלה עוד אתגרים בדרך.

4. רקע עיוני:

בחלק זה נסביר על מספר נושאים חשובים המהווים ידע בסיסי והכרחי על מנת להבין את הנושאים עליהם נסביר במהלך ספר הפרויקט.
מעבר לכך - הידע הזה היה נדרש עבורנו על מנת להבין כיצד לגשת לפתרון הבעיה, על מנת לחקור לעומק את המאפיינים של דיבור ושירה ואת ההבדלים ביניהם.

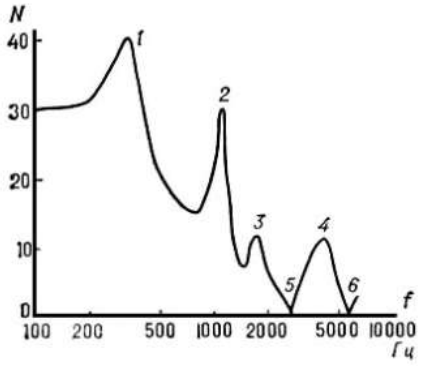
נציין שבמהלך העבודה על הפרויקט קראו והעמקנו בנושאים רבים, קראנו מאמרים רבים, אך לא את כולם נוכל להזכיר ולפרט עליהם. נעשתה הרבה עבודה בינינו ומול המנחה, ואת התיעוד שמרנו במסמך משותף שליווה אותנו לאורך השנה.

4.1. טון בסיס (F0) והרמוניות באותות אקוסטיים:

בקורסים השונים בתואר העוסקים בגלים ותדרים אנו מכירים שקול מורכב משילוב שונה של תדרים.
ישנו את תדר הבסיס, נסמנו F0, ואליו מצטרפות הרמוניות - כלומר כפולות שלמות של אותו התדר הנובעות מגוף התהודה. ההרמוניות של התדר הן משהו קבוע בטבע (מכפלות שלמות של תדירות הבסיס) ולכן כאשר שני הזמרים שרים אותו הצליל (=אותו התדר) אז גם ההרמוניות שלהם זהות. מה שמבדיל בין שני הצלילים הנשמעים (שהרי הם שונים) הוא העוצמה של כל אחת מהן הרמוניות, מה שנקרא גם המעטפת ההרמונית. עבור כל אדם, ובכלל כל צליל (זה יכול להיות כלי נגינה, צופר של רכב או כל רעש אחר) המעטפת ההרמונית מייצגת את גוון הצליל כפי שאנו שומעים אותו.

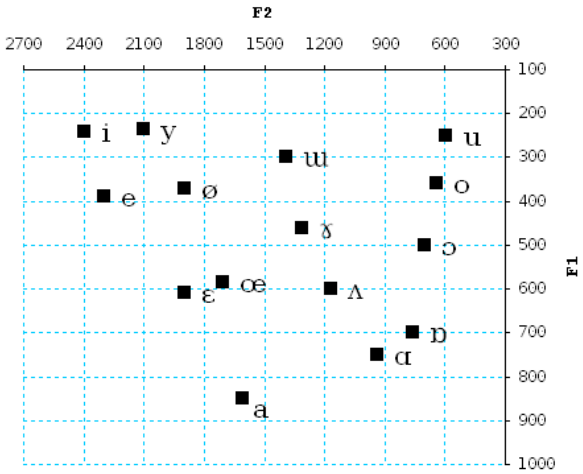
4.2. פורמנטים:

המושג "פרומנטים" (Formants) מתאר את מבנה המעטפת המעטפת ההרמונית.
פורמנט הוא שיא בספקטרום התדרים של צליל הנובע מתדרי התהודה של מערכת אקוסטית. המונח משמש בפונטיקה ובאקוסטיקה לתיאור תדרי התהודה של מיתרי הקול וכלי נגינה.
התמונה הבאה תדגים מהם פורמנטים:



ניתן לראות בתמונה לעיל כי הפורמנטים הם ריכוזים של תדרים חזקים יחסית בטווחים מסוימים.
בפועל, מבחינת הפורמנטים בדיבור ובשירה המשפיעים ביותר הם ארבעת הפורמנטים הראשונים, כאשר לפי רוב המחקרים הפורמנט הראשון והשני - הגובה/עוצמה שלהם והמיקום שלהם (סביב איזה תדר הם) מייצגים את ההבהרה הנשמעת: A E I O U.

ישנה מפה שעוזרת להגדיר על פי 2 הפורמנטים הראשונים מהי התנועה הנשמעת:

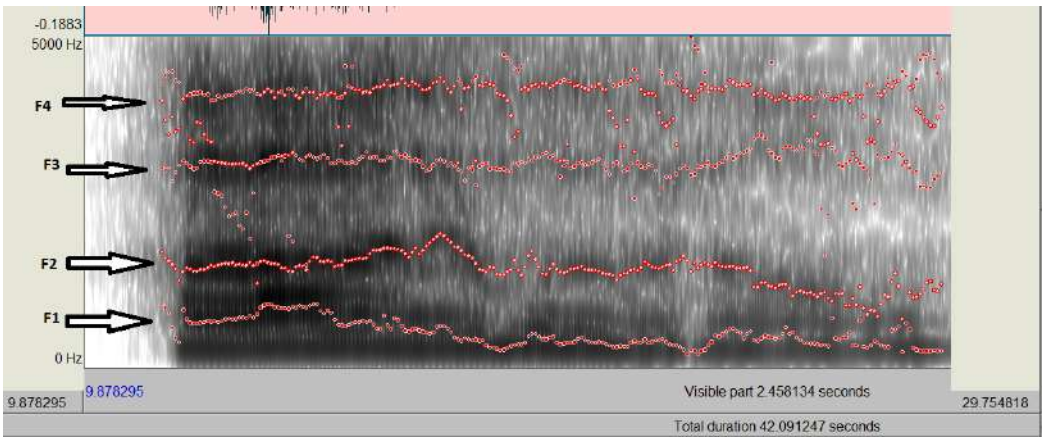


על פי שילוב התדרים, של הפורמנט הראשון F1 והפורמנט השני F2, ניתן לסווג פחות או יותר את התנועה הנשמעת. המפה הזו מהווה בעצם סוג של הורדת מימד של הדאטה העוזרת בסיווג של תנועות.

שני הפורמנטים הבאים, השלישי והרביעי, מייצגים את גוון הקול של הדובר.

נוח מאוד להסתכל ולזהות את הפורמנטים בספקטוגרמה.

נתבונן בספקטוגרמה הבאה:



באופן די ברור ניתן לראות את המקבצים השונים של ההרמוניות החזקות = מיקום הפורמנטים השונים. הסימונים האדומים הם שיערוך של מיקום הפורמנט (איור זה נעשה באמצעות תוכנת Praat, תוכנה מאוד שימושית ונפוצה בתחום עיבוד וחקירת צלילים והקלטות).

4.3. מקדמי LPC:

מקדמי LPC (קיצור עבור Linear Predictive Coding) הם שיטה לייצוג המעטפה הספקטראלית של אות בצורה דחוסה, תוך שימוש במידע של מודל חיזוי ליניארי. מקדמי ה-LPC מייצגים את המקדמים עבור ייצוג של האות כתהליך אוטורגרסיבי (תהליך AR). בקורס SSP1 למדנו רבות בנושא שערוך ספקטרום פרמטרי. למדנו על תהליכי AR, תהליכי MA ותהליכי ARMA.

לשם חישוב מקדמי ה-LPC יש לבחור את סדר המודל - כלומר את כמות המקדמים שאנו רוצים למצוא שבעזרתם נייצג את האות. לאחר מכן נוכל להיעזר בכלים אוטומטיים אשר מחשבים את מקדמי ה-LPC.

מבחינת משוואות אנו יכולים להיעזר במשוואות שלמדנו בקורס SSP1:

אנו מניחים שהאות הוא לפי המודל:

$$x[n] = -\sum_{k=1}^{p} a_k x[n-k] + w[n]$$

מתקיים קשר רקורסיבי בין ערכי הקורלציה:

$$R_{xx}[\ell] = -\sum_{k=1}^{p} a_k R_{xx}[\ell - k]; \quad \ell > 0 .$$

קבלנו שמקדמי המסנן צריכים לקיים את המשוואה המטריצית הבאה:

$$\underbrace{\begin{bmatrix} R_{xx}[0] & R_{xx}[-1] & \cdots & R_{xx}[-p+1] \\ R_{xx}[1] & R_{xx}[0] & \cdots & R_{xx}[-p+2] \\ \vdots & \ddots & \ddots & \vdots \\ R_{xx}[p-1] & \cdots & R_{xx}[1] & R_{xx}[0] \end{bmatrix}}_{\text{R-Toeplitz, symmetric matrix}} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}}_{\mathbf{a}} = -\underbrace{\begin{bmatrix} R_{xx}[1] \\ R_{xx}[2] \\ \vdots \\ R_{xx}[p] \end{bmatrix}}_{\mathbf{r}}$$

ויש לפתור ולחלץ את וקטור הפרמטרים a.

ניתן לאחד המשוואות לקבלת:

$$\begin{bmatrix} R_{xx}[0] & R_{xx}[-1] & \cdots & \cdots & R_{xx}[-p] \\ R_{xx}[1] & R_{xx}[0] & \ddots & & R_{xx}[-p+1] \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ R_{xx}[p] & \cdots & \cdots & R_{xx}[1] & R_{xx}[0] \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma_w^2 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

מערכת המשואות מכונה משואות Yule-Walker.

שערוך הספקטרום מתבצע עפ"י השלבים הבאים:

1. שערוך קורלציה עקיב $\hat{R}_{xx}[\ell] = \frac{1}{N-|\ell|} \sum_{n=0}^{N-1-|\ell|} x[n]x[n+\ell]; \quad 0 \le \ell \le p$
2. $\hat{R}_{xx}[-\ell] = \hat{R}_{xx}[\ell]$
3. פתרון משואות Yule-Walker עם ערכי הקורלציה המשוערכים לקבלת $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p, \hat{\sigma}_w^2$.
4. שערוך ספקטרום-Z $\hat{S}_{xx}(z) = \frac{\hat{\sigma}_w^2}{\hat{A}(z)\hat{A}^*\left(1/z^*\right)}$ כאשר $\hat{A}(z) = 1 + \sum_{k=1}^{p} \hat{a}_k z^{-k}$

   והספקטרום ע"י $\hat{S}_{xx}\left(e^{j\omega}\right) = \frac{\hat{\sigma}_w^2}{\left|\hat{A}\left(e^{j\omega}\right)\right|^2}$.

תהליך חישוב מקדמי ה-LPC המתואר מעלה הוא עבור מציאת מקדמים לספקטרום, אולם אנו רוצים לעשות באותות המשתנים בצורה משמעותית בזמן. לכן יש לבצע חלוקה של האות לחלונות, לחשב את מקדמי ה-LPC לכל חלון, ולבסוף אנו נקבל **מטריצת** מקדמים המתארת את האות.

בדומה לשיקולים הידועים כמו בחישוב stft, יש לקחת בחשבון ולשים לב לגודל החלון, החפיפה בין המסגרות וכו'.

<u>שחזור אות בעזרת LPC</u>

לאחר שחישבנו את מקדמי ה-LPC, על מנת לשחזר את האות, אני מייצרים:

- מסנן על פי מקדמי ה-LPC
- אות רעש לבן

לאחר שיש לנו את 2 האובייקטים הנ"ל, אנו מעבירים את אות הרעש דרך המסנן, הרעש נצבע ומתקבל אות משוחזר. ככל שמקדמי ה-LPC יתארו טוב יותר את האות, כך האות שנקבל יהיה איכותי יותר.

נציין שבהמשך העבודה השתמשנו בטכניקה דומה, איך ביצענו סינו מעט שונה, ולא עם רעש לבן.

בהקשר של <u>ניתוח אותות דיבור</u>, נעשה בעבר שימוש במקדמי LPC על מנת להעביר בצורה דחוסה מאוד את המידע.

אחת הדוגמאות היא LPC10 [2], אשר מקודדת אותות דיבור בתדר דגימה של 8kHz בעזרת 10 מקדמי LPC.

שיטה זו אפשרה להעביר אותות דיבור בצורה די טובה, ולא דרשה זיכרון רב עבור המימוש.

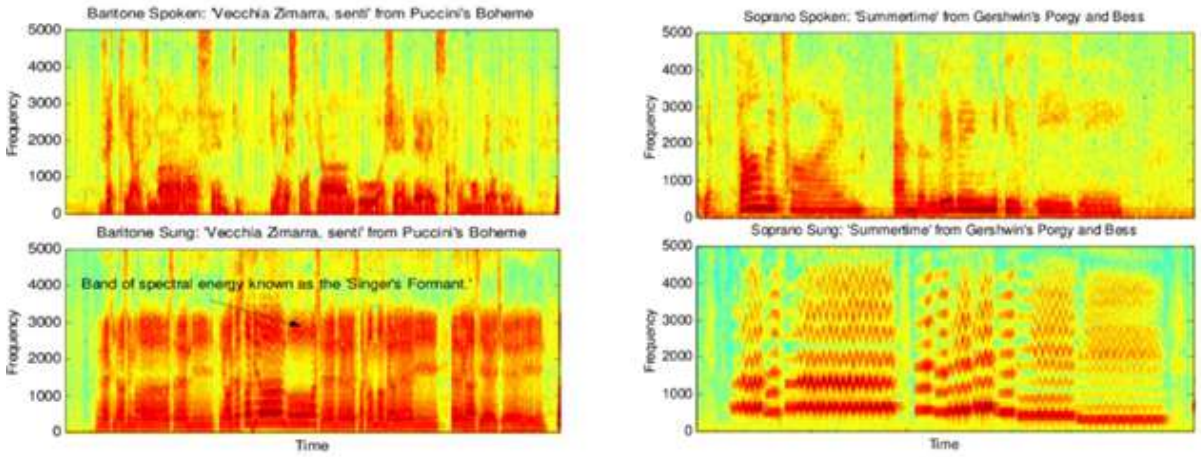
4.4. <u>הבדלים בין דיבור ושירה, וכיצד הם משתקפים בפורמנטים ובספקטוגרמות וכו'.</u>

כאשר אנו שומעים דיבור או שירה או כל דבר אחר, אנו שומעים <u>אות אקוסטי</u>, כלומר תנודות אוויר. התנודות נקלטות אצלנו באוזן ואנו מפענחים אותן לצליל כלשהו – דיבור, שירה, מוזיקה, רעש וכו'. נציג מספר מאפיינים אשר עוזרים בניתוח והבנה של אות כזה, וכמו כן נוכל לראות כיצד ההבדלים בין דיבור ושירה משתקפים דרכם וכיצד ניתן להבחין בהבדלים פיזיים ממש ביניהם.

ניתן לאפיין כמה פרמטרים עיקריים כמו גובה צליל (pitch), עוצמה (loudness), משך הצליל (duration) וגוון (timber). הם באים לידי ביטוי באופן קצת שונה, מוגדר יותר או פחות בשירה ובדיבור. נציג עוד כמה פרמטרים נוספים שעוזרים לתאר ולאפיין אותות שמע.

<u>ספקטוגרמה ומנעד הצליל - טון הבסיס והטונים העיליים בספקטוגרמה:</u>

נתבונן בספקטוגרמות הבאות:



ימין למעלה: דיבור זמרת סופרן, ימין למטה: שירת זמרת סופרן

שמאל למעלה: דיבור זמר בריטון, שמאל למטה: שירת זמר בריטון.

[2] "FIPS PUB 137, Analog to Digital Conversion of Voice by 2,400 Bit/Second Linear Predictive Coding" (PDF). National Institute of Standards and Technology. Retrieved 2018-08-17.

בתמונת הספקטוגרמה ניתן לראות את טון הבסיס, הקו התחתון ביותר בספקטוגרמה האדום ביותר יהיה טון הבסיס אותו אנו תופסים כגובה הצליל אותו אנו שומעים. ניתן לראות בבירור בתמונה מעלה בשירה של זמרת הסופרן שיש לפחות 7 פסים אדומים גליים. זאת אומרת שבנוסף לטון הבסיס, מצטרפים (לפחות) 7 טונים עיליים שהם בעוצמה חזקה מספיק על מנת שהכלי ליצירת הספקטוגרמה יציג אותם.

בנוסף, ניתן לראות שהעוצמה של הטונים (אשר מתבטאת בגוון) אינה זהה בכולם, ואינה משתווה לעוצמה של טון הבסיס – טון הבסיס כמעט תמיד יופיע בעוצמה גדולה הרבה יותר. יחד עם זאת, ההבדלים בין העוצמות בין הטונים העיליים, והדגשים על טונים מסוימים או אחרים – כל אלה הם שנותנים אופי שונה לכל אדם / כלי נגינה.

<u>ספקטוגרמה ומנעד הצליל - ההבדלים הנשקפים מתוך תמונת הספקטוגרמה:</u>

אף על פי שדיבור ושירה שניהם נוצרים בעזרת מיתרים הקול שלנו, בעזרת אותו "כלי", אנו בכל זאת מסוגלים להבחין בהבדלים ביניהם. ההבדלים העיקריים הם:

- **בדיבור**: בדיבור אנו יכולים לראות תמונה ברורה הרבה פחות, מעט "מרוחה". זאת אומר שגובה הצליל – התדר – מוגדר הרבה פחות. כלומר קשה לזהות צליל בסיס ולהפריד בצורה טובה אל מול הצלילים העיליים. תכונה זו נובעת מכך שבזמן דיבור אנו לא שמים דגש ומנסים לדבר עם טון ספציפי (להוציא מהכלל שפות טונאליות בהן יש חשיבות לגובה הצליל עבור משמעות המילה), דיבור הוא חופשי יותר מאשר שירה ומוקפד פחות.
- **בשירה**: בשירה אנו יכולים לראות את טון הבסיס ואת הצלילים העיליים בצורה ברורה מאוד (בדוגמה הנ"ל ספציפית, אך ניתן להכליל זאת גם לדוגמאות אחרות של שירה). בנוסף אנו רואים מאפיינים נוספים אשר בלטו פחות מאשר בדיבור:
  - <u>המנעד</u> – בשירה יש מנעד רחב הרבה יותר, כלומר בשירה אנו עושים שימוש במגוון רחב של צליל , ובספקטוגרמה אנו רואים זאת בטווח תדרים הרבה יותר רחב. בשירה ישנם תדרים פעילים בצורה משמעותית גם ב- 4000-5000Hz לעומת דיבור שהוא עד בקירוב 1,500-2,000Hz.
  - <u>ויברטו</u>– ניתן לראות בדוגמה של זמרת הסופרן שימוש נרחב בטכניקת ויברטו. טכניקה זו מתבטאת בקווים אופקיים גליים מאוד אך בצורה מדויקת, זאת אומרת שהזמרת מרעידה את הקול שלה, מעלה ומורידה את גובה הצליל, בצורה מחזורית ומדויקת.

<u>קצב דיבור ושירה:</u>

קצב בדיבור ובשירה בא לידי ביטוי בצורה שונה לגמרי.

בשירה הקצב מוכתב בצורה ברורה מאוד מתוך המנגינה והשיר. הזמר צריך לבצע את השירה בדיוק כפי הנדרש – הן מבחינת הדיוק בגובה הצליל, והן מבחינת <u>הקצב</u>.

בדיבור, לעומת זאת, הקצב לא מוגדר וברור, הוא תלוי בדובר, בהגייה שלו, בנושא עליו הוא מדבר ובמילים בהן משתמש. ישנם אנשים שמדברים מאוד מהר, ויש כאלה שלאט. כמו כן, קצב הדיבור ישתנה על פי הקהל אליו פונים – בשיחה לא פורמלית או מול קהל רב. השליטה של הדובר בתוכן
הטקסט משפיעה גם היא, אם אדם מדבר על נושא שהוא לא מכיר, ייתכן ויעצור וישתהה מעט.

אנו רואים הבדל משמעותי בתחום זה בין דיבור ושירה – בעוד שבדיבור הקצב עשוי להשתנות בצורה משמעותית. ייתכן ושני אנשים יאמרו את אותו משפט בדיוק אך יעשו זאת בצורה שונה מאוד מבחינת הקצב. לעומת זאת, בשירה על הזמר לעמוד בדרישה ברורה ומדויקת ובאופן מכליל ניתן לומר ששני זמרים שונים אמורים לשיר את אותו השיר באותה צורה מבחינת הקצב.

4.5. יצירת וקטור embedding לדובר:

כחלק מפתרון בשיטות של Deep Learning, הגענו להבנה שאנו צריכים לייצג "אופי" של בן אדם, כלומר את החתימה הקולית שלו, ולאחר מכן להזין אותה לרשת על מנת לבצע החלפה. ייצוג כזה נקרא embedding, זאת אומרת שאנו רוצים וקטור בגודל מסוים וקבוע שיקיים לפחות את 2 התכונות הבסיסיות הבאות:

- עבור אותו דובר, אך עבור קטעי אודיו שונים שלו - נרצה לקבל וקטורים דומים מאוד. אם יהיה כך, זה אומר שהייצוג הזה אכן מייצג את האדם ולא למשל את תוכן הדברים שהוא אומר.
- עבור דוברים שונים, נרצה לקבל ייצוגים שונים, גם אם הם אומרים את אותו תוכן. כך באופן דומה לדובר יחיד, אנחנו מבטיחים שהוקטור אכן מייצג את האדם ולא את תוכן הדברים.

מעבר לכך, ניתן לבדוק תכונות נוספות, למשל שאנשים שנשמעים דומה ייצוגו בצורה דומה (אך לא זהה כמובן). או למשל שניתן לבצע clustering של נשים לעומת גברים (כיוון שבדרך כלל קולות של גברים הם נמוכים יותר מנשים).

למטרה זו, מצאנו מספר כלים שמייצרים embedding. חלקם היו כחלק מובנה ברשתות ממאמרים שונים, אך אנו מצאנו כלי חיצוני, שמבצע רק את הפעולה הזו - Resemblyzer[3]:

5. הכלים הקיימים טרם העבודה על הפרויקט:

לאחר שהגדרנו את הבעיה ואת הרעיון של הפרויקט יש צורך לבדוק האם יש מאמרים או כלים קיימים כאשר עובדים על אודיו ומבצעים פעולות דומות שיכולות לעזור לנו בפתרון הבעיה.

ישנם מספר כלים קיימים כיום כאשר מבצעים פעולות דומות למה שאנו רוצים:

5.1. כליי text-to-speech :

עולם ה-text-to-speech קיים וישנן מערכות המסנתזות דיבור מתוך טקסט. על אף שיש מערכות שעושות את זה בצורה טובה ואיכותית, עלו מספר בעיות בהקשר לפרויקט זה, שעיקרן היו **תזמון המילים** - מערכות ה- TTS אינן בנויות עבור שירה. במערכות שהצלחנו למצוא כקוד שניתן להשתמש בו, לא היה ניתן לשלוט בתזמון המילים - שזה כלי מרכזי וקריטי בשירה. מעבר לכך, הכלים הקיימים אינם לרוב כקוד פתוח שניתן להשתמש בו בקלות.

5.2. כליי החלפות דוברים:

קיימים כלים המסוגלים להחליף בין דוברים שונים, אשר ברובם בנויים על Deep Learning וההחלפה מתבצעת בדרך כלל בין דוברים קיימים בתוך הדאטה עליו המערכת התאמנה. ישנם מספרים כלים שמצאנו תוך כדי העבודה על הפרויקט אשר מחליפים בין דובר חדש ובין דובר מהמאגר עליו הרשת התאמנה. דוגמה לאחד הכלים הוא של החברה "deepdub" אשר מייצרת דיבוב עבור סרטים - היא מנתחת את הקול של השחקן, ולאחר מכן ניתן לשמוע אותו בקולו בשפות שונות. אולם הכלי מסחרי ואינו נגיש לנו.

ניתן לראות המחשה של הכלי כאן: Coverage of deepdub.ai on N12 Israel News

5.3. כלים של DeepFake:

בשנים האחרונות ניתן לראות שימוש נרחב בשיטות של Deep Learning שנקראות DeepFake, מלשון זיוף. דוגמאות כאלה ניתן לראות למשל בסרטונים מוכרים באינטרנט של נאומים של פוליטיקאים.

[3] https://github.com/resemble-ai/Resemblyzer
https://pypi.org/project/Resemblyzer/

6. **תיאור של חלופות לפתרון**

במהלך העבודה על הפרויקט ניסיתי לפתור את הבעיה בכמה דרכים שונות, ובכמה סוגים שונים של גישות.
ניתן לחלק באופן גס את השיטות ל-2 קבוצות עיקריות:

- שיטות קלאסיות
- שיטות של Deep Learning

לכל אחת מהשיטות יתרונות וחסרונות משלה, קשיים בתחומים שונים וידע נדרש שונה.
נתאר מספר פתרונות אפשריים שניסינו לייצר בכל תחום, ובחלק הבא נציג את הפתרון שאנו חושבים שעשוי להיות המוצלח ביותר.

כאשר ניגשנו בתחילת העבודה על הפרויקט לחשוב על פתרון, עוד לא ידענו בדיוק באיזה כיוון הכי כדאי להתמקד, האם בשיטות קלאסיות או שיטות חדשות יותר (כמו Deep). לכן, חקרנו בשני הכיוונים והתנסנו עם אופציות שונות לפתור את הבעיה. נציג מספר שיטות עליהן עבדנו לאורך השנה. על אף שלא בכולן הצלחנו להגיע לפתרון, אנו סבורים שהן לימדו אותנו רבות, גם בתחומים של עיבוד אותות בשיטות קלאסיות וגם בתחום ה-Deep, ולכן ראוי שנציין את כולן ונסביר עליהן. נשב את תשומת הלב לכך שסדר הצגת השיטות בספר הפרויקט **אינו** הסדר הכרונולוגי בו הגענו לכל שיטה במהלך העבודה על הפרויקט.

6.1. **חלופות לפתרון - שיטות קלאסיות:**

בשיטות קלאסיות אנו נדרשים לנתח את המידע בכלים שלמדנו במהלך התואר כמו יצירת ספקטוגרמות (stft), סינון ועוד. על מנת לעשות כן אנו נדרשים להצליח לנתח את אותות הדיבור והשירה בעזרת פרמטרים פיזיקליים ולאחר מכן לחשוב על דרך להצליח ולבצע את השינוי וההחלפה הנדרשת בעזרת פעולות על אותם פרמטרים שניתחנו.

6.1.1. חלופות לפתרון - שיטות קלאסיות - חלופה א - סינתזה בעזרת מקדמי LPC

בחלק התיאורטי הסברנו מהם מקדמי LPC, כעת נציג שיטה בה ניסיתי להגיע לפתרון הבעיה.

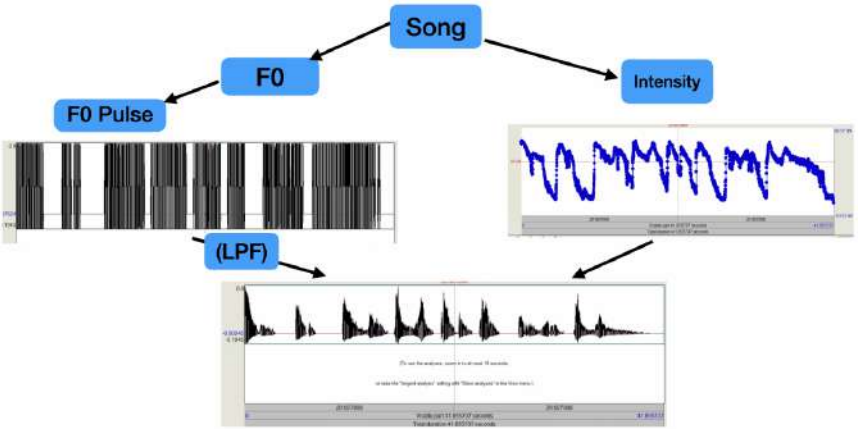
כפי שהסברנו בחלק התיאורטי, בעזרת מקדמי ה-LPC ניתן ליצור מסנן ודרכו להעביר אות. את הפעולה הזו ניתן לעשות בקלות בעזרת תוכנת Praat וכמו כן בעזרת חבילה עבור Praat בשפת פייתון.

נסביר תחילה כיצד אנו מבצעים את הסינון, שכן הוא מעט שונה מהחלק התיאורטי. כמו כן, על מנת לבדוק שהדבר בכלל אפשרי, ניסינו קודם כל לבצע את האלגוריתם ללא החלפה של דוברים. כלומר לקחת קטע שירה (שירה בלבד, ללא ליווי) ולבדוק אם אנו מצליחים להעביר אותו במערכת ולקבל חזרה את השירה.

לאחר מכן נתאר כיצד להרחיב את השיטה על מנת לבצע החלפה, וכמו כן נסביר היכן נתקענו בשיטה זו ומה היה החלק שלא הצלחנו להתגבר עליו על מנת שבסופו של דבר היא תעבוד ונבחר בה לפתור הבעיה.

נתאר את שלבי האלגוריתם - **ללא** החלפת דובר:

1. מתוך השירה, בעזרת תוכנת Praat נחלץ את ה-pitch. נקבל אובייקט שמתאר את תדר הבסיס בלבד בכל רגע, נסמן אותו כ- 'F0'.
2. נייצר בעזרת פקודה ב praat קובץ אודיו הבנוי מפולסים לפי ה- F0 שחילצנו בשלב הקודם. השלב הזה מייצר קובץ אודיו שבו מלבד לתדר הבסיס כפי שחולץ, מתווספות כל ההרמוניות (עד תדר מקסימלי שניתן לקבוע). נכנה את האות הזה F0_pulses. האות הזה אמור לדמות את האות שנוצר במיתרי הקול - אות עם תדר בסיס והרמוניות.
3. נעביר את F0_pulses במסנן LPF, למשל בתוכנת audacity, על מנת שההרמוניות הגבוהות יהיו פחות חזקות ודומיננטיות, שכן בזמן יצירת F0_pulses תוכנת Praat מייצרת את כל ההרמוניות באמפליטודה זהה. מעבר של האות במסנן LPF מנחית את ההרמוניות הגבוהות והאות המתקבל דומה יותר לצורה הטבעית שבה נוצר הקול במיתר הקול וחלל הפה.
4. מתוך קובץ השירה, נחלץ מטריצת LPC, כאשר בכל חלון זמן ניצר P מקדמים. את הפרמטר P יש לקבוע לפי תדר הדגימה של האות. למשל עבור אות שנדגם בתדר 48kHz וחלון זמן של 0.025 שניות, ניקח 32 או 64 מקדמים.
5. נייצר אות רעש באורך של השירה, ולאחר מכן נאפס אותו בכל המקומות שבשירה אין תנועות, למשל אותיות שורקות (s,sh וכו'). התהליך הזה בשלב הראשון יהיה ידני, אך יש לזכור שיש לעשות אותו רק פעם אחת עבור כל שיר שרוצים להכניס שהמערכת תתמוך בו, ולא בכל פעם שרוצים לבצע החלפה.
6. נאחד את אות הרעש ואת F0_pulses שעבר סינון. שילוב שני האותות מאפשר לקבל תוצאה טובה יותר, כוון שאות F0_pulses אינו מייצג היטב עיצורים, ועל כך יפצה אות הרעש.
7. נעביר את האות המאוחד של F0_pulses והרעש בסינון דרך ה-LPC שחילצנו.



התרשים הנ"ל מתאר את שלבי האלגוריתם.

נזכיר השלבים הנ"ל מאפשרים לבדוק אם האלגוריתם מסוגל "לסגור מעגל" - כלומר להתחיל באות שירה וחזור ולקבל את אותו האות בחזרה. מצורפים דוגמאות עבור התהליך הזה, קובץ שירה מקורי, וקובץ בו ביצענו את התהליך הנ"ל.

לאחר שביצענו מספר בדיקות[4] של השלבים הנ"ל, ראינו שניתן לקבל אות שנשמע בצורה די טובה ונאמנה למקור.

[4] ראה בנספחים קישור לתיקיית דרייב בה נמצאים תוצאות מהפרויקט

כעת נותר להבין כיצד לבצע תהליך דומה על מנת **לבצע החלפה** בין דוברים שונים

על מנת לבדוק היתכנות לפתרון הבעיה, ניסינו לבצע את המהלך הבא:
הקלטנו את אהרון **אומר** (בדגש ללא pitch) את מילות השיר בצורה מתוזמנת לגמרי כמו השירה. זאת אומרת שהוא צריך לומר את המילים בדיוק על בתזמון של השירה, בין אם מאריכים או מקצרים מילים או תנועות מסוימות. זו דרישה לא טבעית ולא דרישה שנרצה לדרוש מהמשתמש, אך היא אפשרה לנו לבדוק אם השיטה עשויה לעבוד.
כעת לאחר שיש את קובץ הדיבור המתוזמן של הדובר החדש, נחלץ ממנו מטריקת LPC כמו בשלב 4 באלגוריתם שתיארנו לעיל, ולאחר מכן נמשיך בשאר השלבים כרגיל.

כאשר ביצענו את הניסוי הזה, קיבלנו קובץ אודיו[5] אשר לפי דעתנו ניתן לשמוע בו את אהרון שר השיר.

אם כן, הניסוי הזה הצליח להראות לנו שהשיטה הזו אפשרית על מנת לבצע החלפה. יחד עם זאת, אנו זקוקים עדיין לגשר על בעיה אחת - דרשנו מהדובר החדש לומר את מילות השיר בצורה מתוזמנת ומדויקת מאוד. זו דרישה קשה מדי ולא אפשרית בפועל. לכן, הבנו שאנו צריכים להתגבר עליה בכלים אחרים - וכעת בחלק הזה לשלב כלים מעולם ה-Deep.

נגדיר את הבעיה שניסינו לפתור בצורה ברורה יותר:
אנו דורשים מהמכונה לקבל כקלט קטעי דיבור של דובר חדש, מהם היא תצטרך לייצר embedding.
נכניס למכונה גם קובץ שירה שאותו נרצה שהדובר ישיר (ייתכן ובשלב ראשון נבצע החלפה רק על שירים מתוך רשימה מוגדרת). יחד עם ה-embedding המכונה צריכה להחזיר מטריצת LPC של הדובר החדש שמייצגת את הדיבור/שירה אילו הוא היה אומר/שר את השיר בצורה מתוזמנת בדיוק. לאחר שתהיה לנו המטריצה הזו, נוכל להמשיך בשלבי האלגוריתם כרגיל.

בשלב הזה חיפשנו מאמרים בנושא, וחיפשנו רשתות אשר יודעות לשלב 2 סוגים קלטים שונים - קלט embedding וקלט נוסף שייצג תוכן מסוים. אולם לאחר חיפושים רבים, וגם נסיונות לכתוב רשתות בעצמו, לא עלה בידנו הדבר, ושלב זה בפתרון נשאר עדיין כשאלה פתוחה.

יחד עם זאת, חשוב לציין כי דווקא החיפושים הללו, וההבנה שיש לשלב בין 2 סוגי קלטים, הובילו אותנו להגיע למאמר שנציג בהמשך בנושא של style-transfer. ומעבר לכך, הפתרון בסגנון הזה לימד אותנו רבות על ניתוח של אותות דיבור ושירה.

[5] וראה בחלק נספחים, קישור לתיקיית דרייב בה נמצאים החומרים והתוצאות

6.1.2. חלופות לפתרון - שיטות קלאסיות - חלופה ב - Formant Domain Adaptation:

כאשר התחלנו לחשוב על פתרון לבעיה של החלפת סגנון שירה ניסינו לחשוב בעצמנו מהן הגורמים מבחינה פיזיקלית, פיזיולוגית ומוסיקלית המייחדים את סגנון השירה של כל יחיד. כמובן שיש מספר רב של מאפיינים שונים בדיבור ובשירה המבדילים בין כל אחד ואחת, אז ניסינו לצמצם את הגורמים השונים לדוגמא מבטא, קצב דיבור, ליקויי דיבור ועוד, והתרכזנו תחילה בגוון הקול בלבד. במילים אחרות, רצינו לדעת למה כאשר שני זמרים שונים שרים את אותו הצליל בדיוק (למשל הצליל 'לה' - 440 הרץ) ושניהם זמרים באותה רמה ושרים עם אותה האפקט, עדיין ניתן להבדיל ביניהם.

בחלק העיוני הסברנו על הרמוניות באותות אקוסטיים ועל פורמנטים, ניסינו לחשוב על פתרון לבעיית החלפת הדוברים בעזרת שינוי הפורמנטים. ראינו שחלק ממה שמגדיר את אופי הקול של אדם הם הפורמנטים בדיבור שלו.

המטרה היא ליצור מערכת אשר תקבל כקלט את השירה של הדובר המקורי, תחלק את כל קטע השירה לחלונות של מקטעים, "תנקה" את הדגשים על ההרמוניות כמו שהם (הכוונה שתעלים את הדגשים הקיימים שמייצרים את המעטפת ההרמונית) כך שיישאר רק תדר הבסיס. לאחר מכן, המערכת תקבל דגימה של הדובר יעד, תלמד את המעטפת ההרמונית שלו, קרי מהן הדגשים בהרמוניות שלו, מהם הפורמנטים והמעטפת ההרמונית של זמר היעד. לאחר מכן, המערכת תשנה עבור כל חלון עבור כל הרמוניה את עוצמתו ותרכיב את האות בחזרה.

בשלב זה בתהליך הפתרון, הבנו שאנחנו צריכים לענות על מספר שאלות חשובות הקשורות למה שלמדנו עד כה, על מנת להבין האם הפתרון הזה אפשרי ויישים:

1. האם הדגשים על כל אחת מהן הרמוניות נשארים קבועים ביחס לתדר הבסיס? זאת אומרת, האם למשל תמיד ההרמוניה השנייה היא בעוצמה 90% מתדר הבסיס, ההרמוניה השלישית היא בעוצמה 65% מתדר הבסיס, ההרמוניה השלישית היא בעוצמה 75% מתדר הבסיס וכו.. או שהמעטפת ההרמונית משתנה כתלות בתדר הבסיס?
2. האם המעטפת ההרמונית תלויה בזמן ובמיקום שלו בשיר? האם ניתן להסתכל על כל חלון בנפרד או שצריך להתייחס גם כן לחלונות המקדימים והמאחרים של החלון הנוכחי?
3. האם החלפת הדגשים של כל אחת מהן הרמוניות, גם אם תתבצע באופן מושלם תהווה שינוי משמעותי מספיק כך שהפלט אכן ישמע כמו זמר היעד? כלומר נצליח לקחת את קטע השירה של זמר המקור ולהתאים את ההרמוניות בדיוק באופן שהן אמורות להיות אם זמר היעד היה שר - האם נשמע שאכן התבצעה החלפה בין הזמרים?

כפי שהסברנו בחלק התיאורטי לגבי פורנמנטים (Formants), הם מתארים את ההרמוניות המודגשות בקול, והיחסים ביניהם מגדירים את התנועה שאנו תופסים. בנוסף הפורמנטים הם חלק ממה שמגדיר את ההבדלים בין אנשים שונים ומהווים חלק מהתכונות שמגדירות את גוון הקול של הדובר.

ניסינו לבצע מספר בדיקות על מנת להצליח לענות על השאלות שהעלנו.

<u>בדיקה ראשונה - דובר בודד, תנועה יחידה:</u>

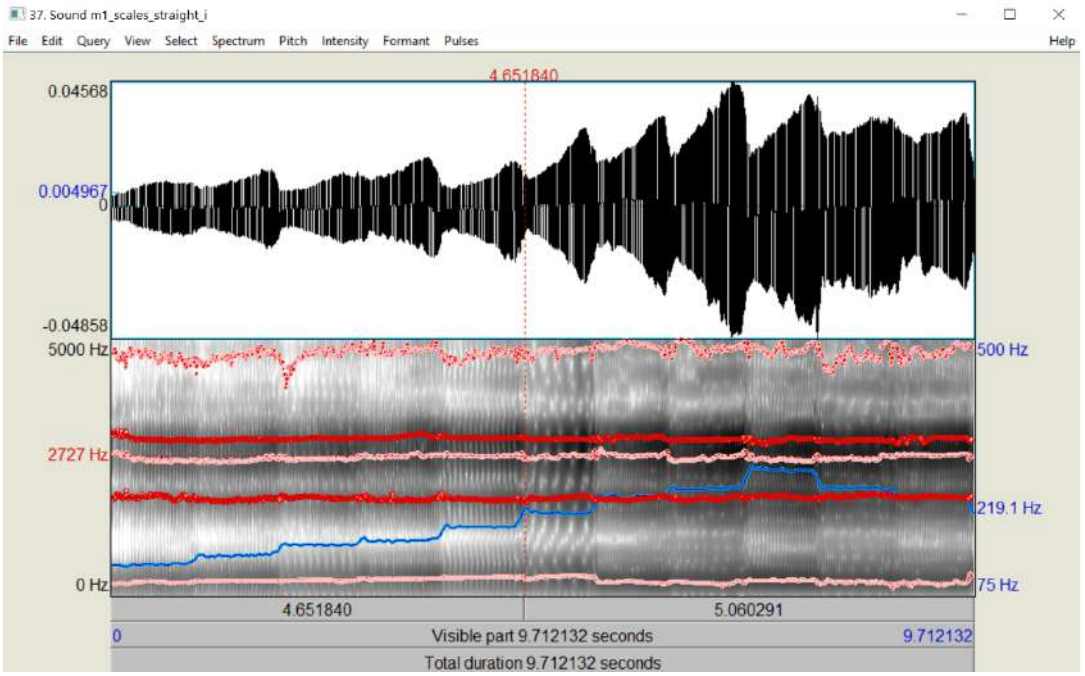
מה שניסינו לבדוק הוא אם מיקום הפורמנטים והמרווחים ביניהם נשארים קבועים כתלות בגובה הצליל של הזמר (הפורמנט הראשון והשני, וגם הפורמנט השלישי והרביעי).

על מנת לבדוק זאת נעזרנו ב-dataset מאוד שימושי שנעזרנו בו רבות - VocalSet.[6] [7]

בדאטה סט הזה ישנם זמרים מקצועיים אשר מבצעים טכניקות שירה שונות. השתמשנו בו רבות על מנת לנתח מאפיינים שונים בשירה.

על מנת לאשש את מה שקראנו במחקרים, לקחנו קובץ של זמר מקצועי מתוך הדאטה הנ"ל, אשר שר הברה קבועה 'i' (על מנת לבדוק את השינוי או חוסר שינוי בפורמנט) אך הוא משנה את גובה הצליל כך שהוא עולה עד השיא ויורד.

נתבונן בתמונה הבאה:



בתמונה הנ"ל ניתן לראות בחלקו העליון את השירה של הדובר בתחום הזמן. בכל התרחבות אמפליטודה של האות יש עלייה בדרגה בסולם (טון/ חצי טון בהתאמה). בחצי התחתון של התמונה יש את הייצוג של האות בתחום התדר. השתמשנו ביכולות של הכלי praat על מנת לשערך את מיקום הפורמנטים [אדום] ואת תדר הבסיס [כחול] (הצליל שהדובר שר).

*(**הערה**: יש לשים לב שסקלת התדרים לפורמנטים בצד שמאל, ולתדר בסיס בצד ימין!)*

---

[6] Wilkins, Julia et al. "VocalSet: A Singing Voice Dataset." *ISMIR* (2018). , VOCALSET: A SINGING VOICE DATASET

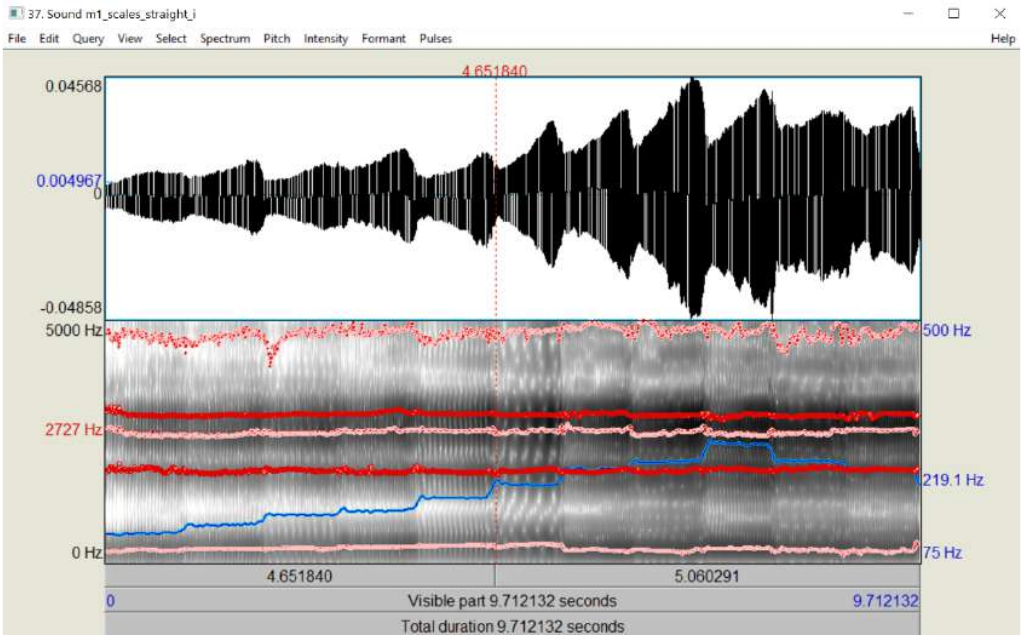
[7] Wilkins, Julia, Prem Seetharaman, Alison Wahl, & Bryan Pardo. (2018). VocalSet: A Singing Voice Dataset (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1193957

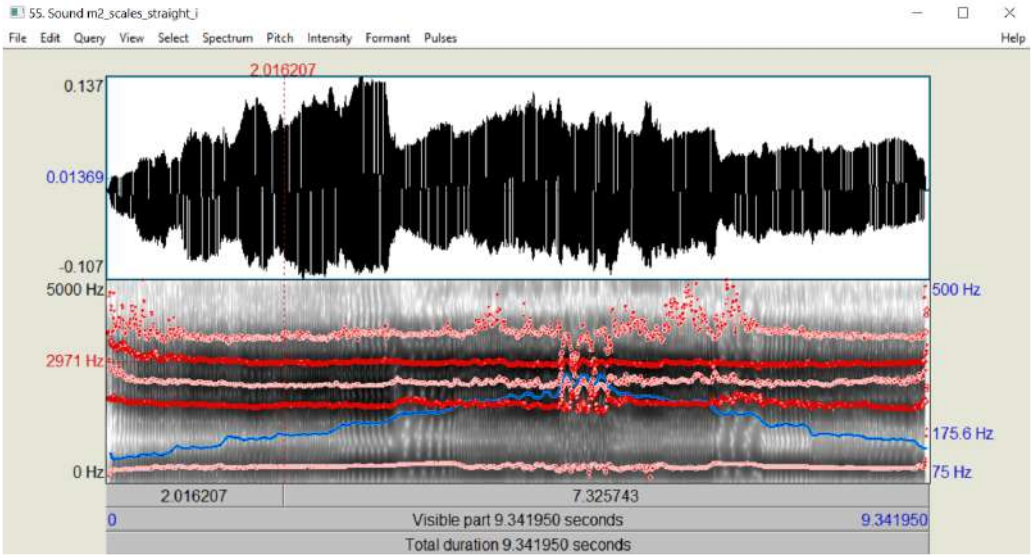
ניתן לראות באופן די ברור שלמרות שהצליל משתנה (הקו הכחול עולה בהתאמה עם הדרגה בסולם), הפורמנטים נשארים קבועים לכל אורך הקטע. הסיבה היא שלעומת גובה הצליל שמשתנה, ההברה נשארה זהה (i) ולכן גם הפורמנטים נשארו במקומם. ניסוי קצר זה אכן אימת את העובדה שמיקום הפורמנטים לא תלוי בתדר צליל השירה.

<u>בדיקה שניה - דוברים שונים, תנועה יחידה:</u>

על מנת לבדוק האם אכן 2 הפורמנטים הראשונים מתארים רק את התנועה הנאמרת, ולא כוללים (כמעט) את אופי הדובר. לקחת 2 דוברים שונים מתוך הדאטה ששרים את אותה התנועה.

נתבונן ב-2 הספקטוגרמות הבאות, כאשר מסומנים עליהן הפורמנטים:





שני הדוברים שרים את התנועה (i) , ניתן לראות ששני הפורמנטים הראשונים הם כמעט זהים, אולם השניים הבאים שונים בצורה די בולטת לעין. כמו כן, יצרנו סקריפט המחלץ את זה את שערוך הפורמנטים ואת רוחב פס התדר שלהם.

ניתן לראות למשל בטבלה הבאה:

time(s)	nformants	F1(Hz)	B1(Hz)	F2(Hz)	B2(Hz)	F3(Hz)	B3(Hz)	F4(Hz)	B4(Hz)	F5(Hz)	B5(Hz)
2.002225	5	399.274	293.878	1924.847	101.776	2387.261	99.105	2953.022	137.530	3569.389	371.592
2.002225	5	399.274	293.878	1924.847	101.776	2387.261	99.105	2953.022	137.530	3569.389	371.592
0.745612	5	378.915	220.480	1956.189	38.611	2738.811	711.911	3289.376	88.457	4627.653	710.835
time(s)	nformants	F1(Hz)	B1(Hz)	F2(Hz)	B2(Hz)	F3(Hz)	B3(Hz)	F4(Hz)	B4(Hz)	F5(Hz)	B5(Hz)
1.470965	4	368.305	300.855	1992.921	54.447	2323.421	313.232	3197.955	146.477	--undefined--	--undefined-
3.377579	5	251.832	92.386	2253.417	62.621	2892.490	231.574	3479.413	64.379	4272.719	559.596
1.914660	5	351.519	21.516	1923.358	93.796	2546.744	148.682	3614.641	59.516	4454.899	373.107
1.901308	5	352.640	34.145	1930.269	23.566	2468.131	64.179	2867.727	65.962	4355.983	558.319
1.613343	4	343.487	40.918	1980.980	188.038	2289.681	449.937	3474.139	195.848	--undefined--	--undefined
0.753031	5	259.885	236.331	1720.869	139.581	2416.079	82.056	2766.895	165.595	4598.376	111.287
2.015553	5	367.057	313.952	1935.483	180.773	2774.050	148.823	3146.372	47.489	3868.173	373.263

<u>בדיקה שלישית - תנועות בהקשרים שונים:</u>

נרצה לבדוק האם הפורמנטים של דובר מסוים, תחילה עבור התנועות הבסיסיות (A I E O U), נראים באופן זהה בהקשרים שונים. כלומר האם הפורמנטים עבור התנועה i במילים שונות יהיו זהים. למשל עבור המילים tip, sip. לאחר בדיקה על מספר קטעי דיבור, נוכחנו לגלות כי יש חשיבות גם להקשר ולמיקום ההברות במילה. זאת אומרת שהפורמנטים אינם קבועים לגמרי! יש לקחת בחשבון את העיצור/תנועה לפני ואחרי. יש לקחת בחשבון אם התנועה נאמר בזמן קצר או נמשכת לאורך זמן. כל הגורמים הללו משפיעים על צורת הפורמנטים.

<u>האם ניתן להגיע לפתרון בצורה הזו?</u>

פתרון אפשרי, מעט נאיבי, יהיה לבנות טבלת צירופים גדולה של תנועות ועיצורים, אשר ממנה ניתן להרכיב מילים רבות כל התנועות והעיצורים מופיעים בה בהקשרים שונים (כפי שראינו שיש חשיבות והשפעה על הפורמנטים לפי המיקום בתוך המילה). לאחר מכן, לקחת דאטה גדול של דיבור/שירה ולבצע מעין סיווג של הדוברים.
בזמן שנרצה לבצע החלפה, יש לדרוש מהדובר החדש להגיד את כל הצירופים מתוך הטבלה. כך בתקווה יהיו שמורים עבורו מספיק צירופים שונים של וידע על הפורמנטים שלו על מנת לבצע החלפה.
לאחר מכן, נשתמש בשיטות של סיווג, כמו "שכן קרוב", על מנת לבדוק למי מתוך הדאטה הדובר הכי דומה.
בשלב הסופי, המערכת תצטרך להצליח לבנות מתוך הדאטה שהיא למדה עליו שחזור של השירה לפי הדובר הקרוב שהיא סיווגה לפיו את הדובר החדש, למשל לשנות את המרווחים בין הפורמנטים שלו, לשנות את התדרים של הפורמנטים ואת הדגשים בהם.

על אף שייתכן והפתרון הנ"ל עשוי לעבוד בצורה מסוימת, חשבנו שהוא לא מספיק אופטימלי ובאופן פרקטי הוא מסובך מדי, על מנת לנסות ולכתוב עבורו קוד ומערכת שתעשה את כל הדרוש.

נציין, שגם בסוג הפתרון הזה, ניתן לבצע החלפה באלגוריתם דומה לפתרון שתיארנו קודם לכן בעזרת מטריצת LPC בעזרת תוכנת Praat. אולם התוצאות שהתקבלו היו באיכות טובה פחות מאשר בעזרת LPC, ולכן לא המשכנו בכיוון ההוא.

יחד עם זאת שלא הגענו לפתרון בדרך זו, העבודה בנושא הזה, למדה אותנו המון, ובחינה וניתוח של קטעי שירה בכלים שתיארנו עזרו לנו להבין טוב יותר כיצד באים לידי ביטוי ההבדלים בין שירה ודיבור. וכיצד באים לידי ביטוי ההבדלים בין אנשים שונים בכלי קלאסיים.

6.2 **חלופות לפתרון - שיטות של Deep Learning:**

טרם נציג את הפתרונות השונים שניסינו בשיטות של למידה עמוקה, נציג מספר נושאים נוספים אשר יש להתמודד איתם בפרויקט זה כאשר ניגשים לפתרון את הבעיה בעזרת שיטות של למידה עמוקה.

6.2.1 חיפוש Dataset מתאים:

שיטות של למידה עמוקה, Deep Learning, דורשות מאגרי מידע, Datasets, גדולים אשר צריכים להתאים לבעיה אותה מנסים לפתרון. למשל כאשר רוצים לבצע סיווג בשיטות כאלה, יש לאסוף מספיק דוגמאות (בין אם מתוייגות או לא) בתחום. לדוגמה - בזיהוי ספרות בכתב יד, יש למצוא מאגר גדול (ואכן קיים כזה - MNIST) של דוגמאות שונות של ספרות שנכתבו בכתב יד, ואת התיוג שלהם - איזו ספרה באמת כתובה שם.

כך בדומה, אם נרצה לפתור את בעית הפרויקט בשיטות של למידה עמוקה, נצטרך למצוא דאטה מספיק גדול אשר מיועד ולתחום, ויכול לעזור לנו לפתור את הבעיה. פתרון בעיות בעזרת למידה עמוקה אינו חדש עבור דיבור, ישנם Datasets רבים עבור דיבור, אך כאשר התחלנו לחפש מאגרים עבור שירה, גילינו שהם מעטים הרבה יותר.

נחלק את סוגי הדאטה שמצאנו ל-2 סוגים:

- דאטה מקבילי
- דאטה לא מקבילי

דאטה מקבילי - הוא דאטה בו עבור כל דובר, וכל דוגמה בדאטה, יש גם את הדיבור וגם את השירה.
דאטה לא מקבילי - כשמו כן הוא, דאטה שעבור כל דובר אין לנו גם את הדיבור וגם את השירה.

הצלחנו למצוא מאגרים גם מקביליים וגם לא מקביליים, אך כמובן שדאטה מקבילי היה קשה למצוא הרבה יותר, והוא גם מועט יותר.

השיקולים בכל אחד מסוגי הדאטה, מקבילי ולא מקבילי:

- דאטה מקבילי: בתחילת העבודה בעזרת שיטות של למידה עמוקה, חשבנו שמכיוון שאנו רוצים לדעת רק מתוך דיבור כיצד תשמע השירה, אז יש צורך בדאטה שיתאים לצורה הזו. כך הרשת תוכל לקבל את הדיבור כקלט, להעביר אותו ברשת, לקבל תוצאה כלשהי - ואותה תוכל להשוות אל מול השירה (כמו סוג של label). אולם מציאת דאטה כזה הוא קשה יותר, וגם מודל הרשת והאימון שלו צריכים להיות בנויים בצורה מסוימת שתתאים לסוג הדאטה הזה.
- דאטה לא מקבילי: קל יותר למצוא דאטה לא מקבילי, אנו יכולים למצוא דאטה רק של דיבור (קיימים מאגרים גדולים) ולמצוא בנפרד דאטה של שירה בלבד, ולאחר מכן להבין כיצד להשתמש בצורה חכמה בכל אחד מהם על מנת לאמן את הרשת כנדרש. בהמשך העבודה על הפרויקט, ראינו אף מספר רשתות ומאמרים שמדגישים את העובדה שהם **אינם** דורשים דאטה מקבילי והציגו זאת כיתרון גדול. מבחינת אימון הרשת, כאשר המודל לא דורש דאטה מקבילי, נוכל להתאמן תחילה על הרבה דאטה של דיבור, להגיע למודל מאומן באופן מסוים, ולאחר מכן לבצע fine-tuning עבור החלפה של דיבור-לשירה בעזרת דאטה של שירה.

6.2.2. חלופות לפתרון - שיטות של Deep Learning - חלופה ג - Phoneme Learning:

עוד טרם הגעת יכולות הלמידה והבינה המלאכותית לתפקיד עיקרי ברוב הטכנולוגיות הקיימות היום, עוד הייתה דרישה למערכות של TTS ו STS (טקסט לדיבור ודיבור לדיבור). למשל, במכשירי הניווט GPS הייתה את היכולת להשמיע את הוראות כיוון הנסיעה. עבור הוראות נפוצות ("פנה ימינה", "הגעת ליעד" וכו) הייתה הקלטה שהכינו מראש. לעומת זאת, הGPS יכל גם להשמיע שמות של רחובות ומקומות, ולא היה פרקטי להקליט את כל מאגר שמות הרחובות ולכן שמרו מילון של כל הפונמות של הדובר ובכך יכלו לשזר יחד את הפונמות לפי סדרם בשם. נזכיר - פונמה היא התת חלק הקטן ביותר של מילה.

על אף שהשיטה הזו עבדה, התוצאות מאוד רחוקות ממה שניתן להשיג היום באמצעות שיטות של למידה. בשיטה זו, דורשים מדובר היעד לומר את כל הפונמות, וכל הצירופים האפשריים. לאחר מכן לבצע הרכבה מחדש לפי הצירופים שלמדנו. השיטה מסורבלת, והפלט גם לא נשמע הומוגני מספיק ואמין, הוא בדרך כלל נשמע מלאכותי ושומעים שהורכב מחלקים בלתי קשורים.

אחד המאמרים בהם עסקנו במהלך העבודה על הפרויקט עשה שימוש באלמנט דומים. המאמר *"A NEURAL PARAMETRIC SINGING SYNTHESIZER"* [8]
ובקיצור נסמנו NPSS, מבוסס על דטאה מתוייג של שירה לרמת חלוקה של הפונמות הנאמרות בו.
למשל, ישנם קבצי lab אשר מתארים את הפונמות שנאמרות:

```
0.000000 2.910000 sil
2.910000 3.070000 ey
3.070000 3.100000 d
3.100000 3.150000 ah
3.150000 3.240000 l
3.240000 3.280000 v
3.280000 3.630000 ay
3.630000 3.740000 s
3.740000 3.760000 sil
3.760000 3.920000 ey
3.920000 3.950000 d
3.950000 4.020000 ah
4.020000 4.070000 l
4.070000 4.120000 v
4.120000 4.390000 ay
4.390000 4.640000 s
4.640000 5.130000 sil
5.130000 5.270000 eh
5.270000 5.300000 v
5.300000 5.330000 r
5.330000 5.360000 iy
5.360000 5.380000 sil
5.380000 5.470000 m
5.470000 5.500000 ao
5.500000 5.530000 ah
```

כאן ניתן לראות חלק מאחד קבצי ה.lab כאשר בעמודה הראשונה מופיע זמן תחילת הפונמה, העמודה השנייה היא זמן סוף הפונמה והעמודה השלישית היא הפונמה עצמה. (<SIL> מתייחס לשקט).

המאמר מציג רשת אשר אמורה ללמוד כיצד אדם שר מתוך דוגמאות של **שירה** שלו (מתויגות באופן שתארנו).

[8] Blaauw M, Bonada J. A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs. *Applied Sciences*. 2017; 7(12):1313. https://doi.org/10.3390/app7121313

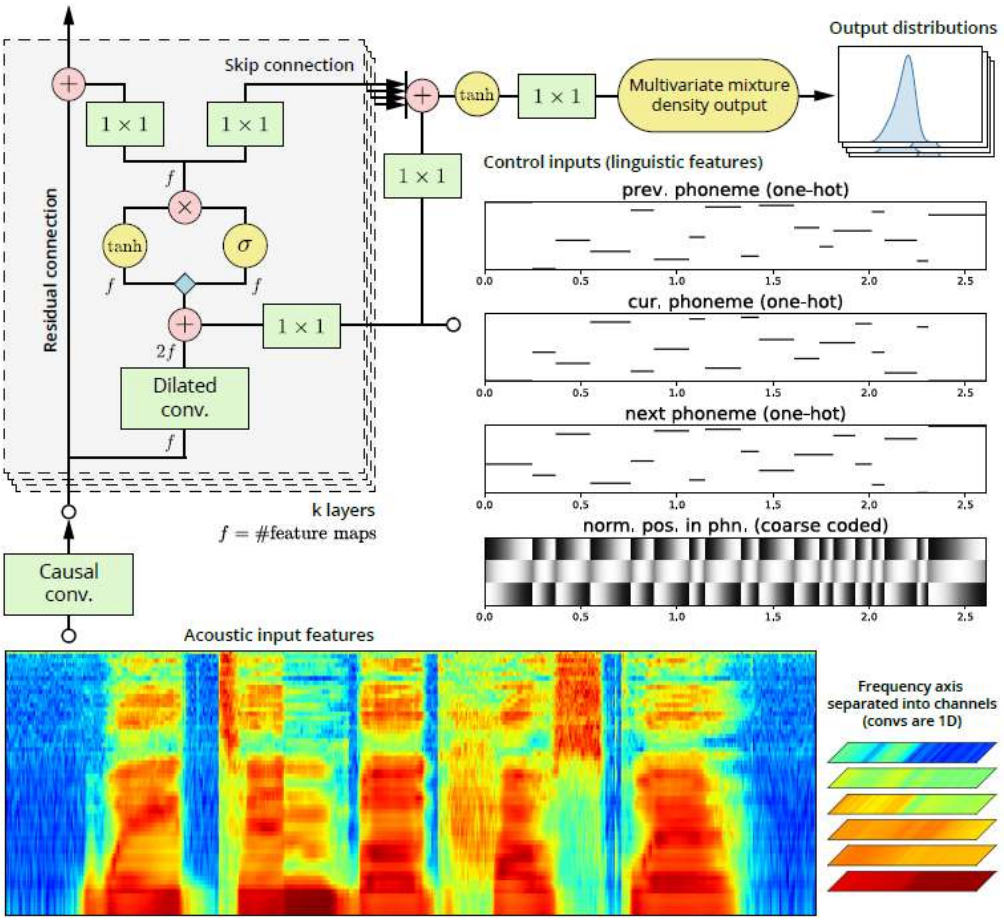
הרשת בנויה מ-3 תתי מודלים:

- מודל הרמוני (Harmonic spectral envelope): המודל אמור ללמוד את החלק ההרמוני עבור השירה.
- מודל אפריודי (Aperiodicity envelope): מודל שצריך ללמוד את החלק הא-פריאודי עבור השירה.
- מודל Voiced/Unvoiced. נסמנו V/UV

3 המודלים הללו יחדיו מרכיבים את הרשת, והם מאומנים בנפרד, על פי המאמר, זאת על מנת שלא לפגוע באיכות ובביצועים של כל אחד מהם.

המודלים מוזנים האחד לשני - המודל Harm הוא הראשון, הוא מוזן לתוך מודל ה- V/UV, ולבסוף התוצאה מוזנת לתוך מודל Aper. ארכיטקטורת 3 המודלים זהה, אך יש הבדלים בהיפרפרמטרים, והאימון כפי שאמרנו נעשה בצורה בלתי תלויה.

תיאור הרשת מתוך המאמר: (עמ 3/23)



מהם הקלטים לרשת?

הרשת אינה מקבלת ישירות את קבצי האודיו, אלא מקבלת פיצ'רים מקודדים עבור השירה.

וקטור הפיצ'רים, במאמר מכונה ה- Control,עבור יש ניתוח של קבצי האודיו בחלונות זמן, ועבור כל חלון זמן שומרים את הפרטים הבאים:

- מחלצים את תדר הבסיס, F0 עבור מסגרת הזמן.
- שמירת פונמות, מתוך הדאטה, שומרים את המידע על הפונמות:
  - הקודמת
  - הנוכחית
  - הבאה
- מיקום מסגרת הזמן בתוך הפונמה:

מכיוון שמסגרת הזמן קטנה יותר מאורך ממוצע של פונמה, אזי נקבל שישנן מסגרות זמן רבות עבור אותה פונמה. לכן מחשבים את מיקום המסגרת בתוך הפונמה, כאשר החלוקה היא לאחת מתוך 3 קטגוריות: התחלת הפונמה (עד 150 מיל" שניות מתחילתה), אמצע הפונמה, סוף הפונמה (עד 150 מיל" שניות מסופה).

אנו יכולים לראות שבעזרת וקטור הפי'צרים הנ"ל אנו שומרים הרבה מאוד מידע בכל מסגרת זמן. כפי שראינו בניסיון אחר לפתור את הבעיה (ראה חלק (Formant Domain Adaptation) הפורמנטים ובכלל הספקטרום של שירה (ודיבור) אינו מוגדר באופן מוחלט רק בעזרת הפונמה הרגעית, אלא יש תלות גדולה מאוד בהקשר. לכן, אשר ברשת המתוארת כאן, שומרים את במסגרות זמן קטנות, את הפונמות הקודמות והבאות, אזי לרשת ייתכן ויהיה מידע מספיק על מנת להצליח ללמוד את הצורות המותנות עבור 3 תתי-המודלים.

הערה חשובה: במהלך אימון הרשת, בכל פעם שמגיעה פונמה חדשה, היא מתווספת לרשימה ששומרים על הזמר. תנאי הכרחי לביצוע החלפה בהמשך יהיה שהפונמות הנדרשות הופיע במהלך אימון הרשת!

כאשר הרצנו את הרשת בעזרת המודל המאומן על הדאטה לדוגמה שניתן יחד איתו, מקבלים תוצאות די טובות אולם כאשר ניסינו להשתמש במודל על דאטה שאינו היה שייך לדאטה האימון של הרשת, התוצאות כבר לא היו טובות. לא הצלחנו לקבל תוצאות הנשמעות כמו שירה, או אפילו דיבור.

יחד עם זאת שדוגמאות ה- demo היו מוצלחות, יש כמה בעיות מהותיות עם המודל הנ"ל:

1. מכיוון שתנאי הכרחי לביצוע החלפה הוא שכל הפונמות בשיר שנרצה להחליף הופיע כבר במהלך האימון. כאשר ניסינו להחליף שיר עם פונמה שלא נראתה קודם לכן - הקוד קרס. יש להבין כיצד להוסיף עוד אלמנט ברשת שיצליח להתמודד עם מצב של פונמה לא מוכרת - למשל, להגדיר אילו פונמות הן הכי דומות, ואז להשתמש בהן במצב כזה.
2. הרשת הזו בנויה ללמוד שירה מתוך שירה! על מנת ללמוד שירה מתוך דיבור, נדרש להבין כיצד לשנות אותה בהתאם. אין מספיק רק לשנות את דאטה האימון, אלא לשנות בצורה מהותית את הרשת.
3. הרשת בנויה ללמוד שירה מתוך שירה של אותו הדובר. אין כאן התייחסות להחלפה בין שני אנשים שונים - זו בעיה מהותית, שניסינו למצוא דרכים כיצד להתמודד איתה, אך עזבנו את הדרך הפתרון הזו טרם הגענו לפתרון.
4. צמצום האפשרויות לבחירת שיר עבור התוכן - כיוון שכל שיר מצריך קובץ טקסט נלווה אשר מתאר איזו פונמה נאמרה באיזה זמן לאורך השיר (כמו שהצגנו לעיל) אזי נצרכת הכנה מראש עבור השירים על מנת שנוכל להשתמש בהם במהלך ההחלפה. לאור עובדה זו, לא נוכל לתת למשתמש לבחור מכל שיר שיחפוץ בו בדיוק בגלל ההכנה מראש שדרושה.

בצענו מספר נסיון ומחקרים על מנת לנסות לפתור את הבעיות הנ"ל, אך לא הצלחנו להגיע לפתרון בכיוון הזה שיענה על הבעיות הללו ויענה גם על הדרישות שלנו של הפרויקט.

7. **תיאור מפורט של השיטה הנבחרת והסיבות לבחירתה.**

7.1. הקדמה:

האפשרות לגרום לדיבור של אדם אחד, ובכלל שירה של אדם אחד להישמע כמו אדם אחר היא יכולת מרגשת המתקרבת לכדי מימוש בזמן הקרוב. כיום, הרעיון של החלפת קול בדיבור או בשירה מוצג כמשימה של style transfer (החלפת סגנון). על אף מספר הרב של ניסיונות לפתור את האתגר של החלפת קול הדיבור כמו שהצגנו לעיל בפרקים הקודמים, טרם הצליחו להגיע לתוצאות מספקות.

מספר בעיות עלו עם השיטות שהוצגו עד כה:

1. רוב השיטות להחלפת סגנון דיבור או שירה מכילות בארכיטקטורה שלהם את הצורך בדאסט מקבילי - מקבץ של זוגות או קבוצות של דוברים מדברים ושרים את אותם המשפטים/קטעים, כך שבזמן האימון הרשת מקבלת את קטע הדיבור כמקור ואת הקטע התואם של השירה כיעד ובכך לומדת את ההבדלים בין הדיבור לשירה. רק מעטים מכל השיטות הקיימות יכולות להתאמן על דאטה שאינו מקבילי כזה.
2. מבין השיטות הבודדות היכולות להתמודד עם דאטה סט כלשהו שאינו מקבילי, כמעט ואין אף שיטה שבנויה בצורה של many-to-many קרי רשת שיכולה לקבל כקלט מספר רב של דוברים ולהוציא כפלט מספר רב של זמרים (ולא מצב לדוגמא שבו הרשת התאמנה על זמר ספציפי ועבור דובר/שיר שנכנס כקלט נוכל לשמוע את אותו הדובר הקבוע שר בפלט).
3. מעטות מאוד הרשתות שמאפשרות יכולת של zero-shot-conversion, הכוונה שהרשת לא נדרשה ללמוד את דובר היעד שלנו במהלך שלב הלמידה אלא שניתן ממש לקבל את דובר היעד בשלב ה-test ולבצע את ההחלפה בהצלחה.
4. הצבנו לעצמנו כמטרה בתחילת הפרויקט שאורך הקטעים של דובר היעד ממנו נלמדת סגנון השירה הרצויה לא יכול להיות ארוך במיוחד (לא קבענו זמן ספציפי אך במחשבה רצינו פרק זמן סביר עבור משתמש שיוכל להקליט את עצמו ללא הסחות דעת והפרעות. משהו בסדר גודל של מספר שניות או דקות בודדות). בנוסף לכך, דרשנו שכל תהליך ההחלפה יהיה מהיר מאוד (עניין של כמה דקות במקסימום) ולא פרקי זמן ארוכים במיוחד, בדומה למספר שיטות אחרות שכבר קיימות.

בהסתכלות על משימה זו כמשימה של style transfer, רצינו להשתמש בכלים מתקדמים לפתירת בעיות דומות (כפי שהצגנו בתחילת ספר הפרויקט כדוגמה עבור החלפת סגנון תמונות וכו).

2 מהכלים המובילים לעמידה באתגרים אלה הם **GAN** ו-**CVAE**.

GAN (=Generative adversarial network) , **CVAE** (=Convolutional Variational Autoencoder)

על אף התוצאות המרשימות של כלים אלה, הן אינן מושלמות:

ה-GAN באופן תיאורטי פועל בשיטה בה ישנה הנחה כי ההתלפגות של התוצר מתאים להתפלגות של המקור. שיטה זו מצליחה להניב תוצאות מאוד טובות, במיוחד בתחום של computer vision. לעומת זאת, תכונה נפוצה עבור ה-GAN היא שמאוד קשה לאמן את הרשת הזו, ההתכנסות שלה מאוד שברירית והרשת לא עיקבית בתוצאות. יתר על כך, למרות העלייה בשימוש של GAN בבעיות של החלפת סגנון קול, אין מספיק מידע מהימן

המוכיח כי GAN יכול לייצר קבצי דיבור המשכנעים את האוזן האנושית בהתאם לאיכות התוצר. העובדה שהרשת הצליחה "לעבוד" על ה-discriminator לא בהכרח אומר שהתוצר ישמע טוב לאוזן האנושית.

ה-CVAE לעומת זאת היא רשת שיחסית קל לאמן אותה. יתר על כך, הרשת מתאמנת אך ורק עם פונקציות מחיר על השחזור של הקלט. כיוון שהרשת מצמצמת את המידע הנקלט ומשחזרת אותו, יש דרישה ותנאים מקלים יותר ביחס לכמות וסוג הדאטה לטובת אימון הרשת. אך כאן ישנה בעיה של חוסר התאמה בין התפלגות המקור לעומת התוצר, ולכן נוצרות בעיות כמו החלקת יתר ועיבוד יתר של הדאטה.

השיטה שאנו בחרנו לפתח במהלך הפרויקט שלנו בנויה על טכניקה שפורסמה במאמר
[9]AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss

המוטיבציה של חוקרים היא ליצור רשת קלה לאימון בתכונותיה בדומה ל-CVAE אך יחד עם זאת שמצליחה לייצר תוצאות התואמות את ההתפלגות של הקלט.

7.2. הכנת הדאטה

במאמר, שנכנה אותו Autovc, צוין שאימון הרשת בוצע על הדאטהסט של VCTK, מאגר חינמי של דוברים רבים אומרים משפטים קצרים באנגלית (פירוט נוסף בהמשך על הדאטהסט).

בסוף שלב ההכנה של הדאטה, נוצר אובייקט מסוג .pkl אשר מכיל את כל שמות הדוברים השונים.

עבור כל דובר יש וקטור embedding (כלומר וקטור בגודל קבוע המתאר את אופי סגנון הדיבור של אותו הדובר), וכן רשימה של ספקטרוגרמות עבור כל קטע בדאטהסט של אותו דובר יש ספקטרוגמרה (במאמר המקורי שימוש ב- mel-spectrogram) של אותו הקטע.

וקטור ה-embedding בקוד המתאר את המאמר נעשה על ידי רשת שאומנה על ידם. בהמשך העבודה על הפרויקט אנו החלפנו את הרשת הזו בכלי אחר, הנקרא resemblyzer, ראה חלק זה בהמשך ספר הפרויקט.

7.3. תיאור השיטה:

ראשית, חשוב לציין כי השיטה המוצעת במאמר זה מיועדת להחלפת סגנון של דיבור, קרי הרשת מקבלת קובץ של דובר א׳ מדבר אשר ישמש עבור התוכן של הפלט, וקובץ (או מספר קבצים) של דיבור על ידי דובר ב׳, שממנה הרשת תלמד את הסגנון של דובר ב׳. הרשת תבצע החלפה של הסגנון ותייצר את תוכן דובר א׳ בסגנון דובר ב׳.

כיוון שאנו מעוניינים דווקא לבצע החלפת סגנון קול של שירה מתוך דיבור (ולא דיבור מתוך דיבור) משימתנו הייתה מאתגרת אף יותר ממספר בחינות:

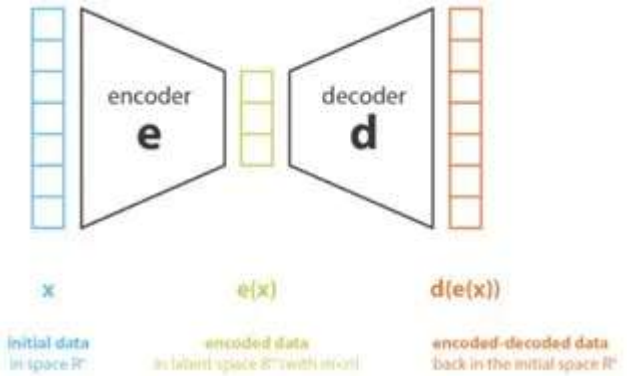
1. שני הרשתות מיועדות לייצר פלט של משהו שלא היה קיים לפני (=רשת גנרטיבית), אך במקרה של יצירת דיבור, זו בעצם מה שנקרא more of the same - קיבלנו דיבור של אדם והרשת תייצר פלט של אותו דובר גם כן מדבר אך בהבדל תוכן המילים. כאשר מדובר במשימה שלנו - יצירת פלט של שירה

[9] Qian, K., Zhang, Y., Chang, S., Yang, X. & Hasegawa-Johnson, M.. (2019). AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. *Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 97:5210-5219 Available from https://proceedings.mlr.press/v97/qian19c.html

מתוך דיבור - למעשה הרשת (וגם אנחנו) בכלל לא יודעת איך הדובר שר, ולכן קפיצת המדרגה מדיבור לשירה נעשית ללא ניסיון או הכרה של אופי הפלט.

2. מלבד השוני בין החלפת סגנון דיבור להחלפת סגנון שירה, היו עוד מספר מרכיבים בשיטה שלהם שהיה עלינו לשפר על מנת להגיע לתוצאות שישיגו את המטרות שהגדרנו בתחילת הפרויקט - אורך הקבצים הנלמדים עבור סגנון השירה, איכות הפלט, קצב למידה ועוד.

למרות הבדלים מהותיים אלה, ראינו יתרונות רבים בשיטה המוצעת במאמר ובביסוס הפתרון שלנו על הממצאים שפרוסמו, ולכן תחילה נפרט בנושא המאמר.



באופן כללי, השיטה של autovc, היא אלגוריתם של רבים-לרבים (מספר קבצי תוכן שונים למספר דוברים חדשים) המבצעת החלפת סגנון ללא שימוש בדאטה מקבילי. באמצעות auto encoder ותכנון מאוד מדויק של צוואר הבקבוק שלו, בנוסף להורדת מימד מדודה וסינון דצימציה של המידע, האלגוריתם מצליח לייצר פלט של דובר שלא נראה מעולם ברמה גבוהה מאוד ביחס לשיטות האחרות הקיימות.

<u>נגדיר בצורה מתמטית:</u>

תהי U_i אופי השירה של דובר i , ותהי Z_j תהליך אקראי שהיא הדגימה של פונמה j כלשהי $Z = Z(1:T)$ מתוך התפלגות P_Z אזי X_{ij} הינו תהליך אקראי ארגודי סטציונארי של דובר i שר את פונמה j. ולכן גם $X = X(1:T)$. X יכול לייצג דגימה של קובץ אודיו או פורמט אחר, במקרה שלנו הייצוג היא ספקטרוגרמה של השירה בקטע ספציפי.

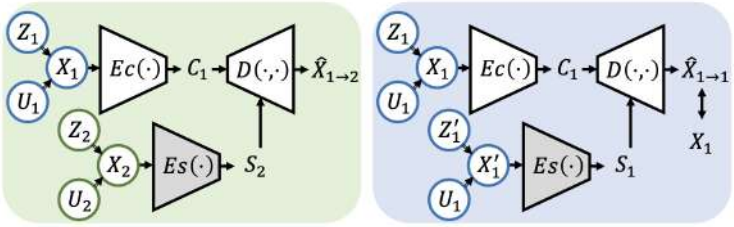
התוצאה הרצויה עבור תהליך החלפה צריכה להתאפיין באופן דומה לנוסחה הזו:

pXˆ1→2(·|U2 = u2,Z1 = z1) = pX(·|U = u2,Z = z1)

כאשר U1 ו- U2 שניהם מהווים חלק מדאטהסט האימון של הרשת, אזי המשימה היא משימה נפוצה עבור החלפת רבים-רבים בסגנון הדיבור ו/או השירה. אך במידה ו-U1 או U2 הם לא חלק מדאטהסט האימון, המשימה הרבה יותר מורכבת , בסגנון zero-shot-conversion, שזו גם אחת המטרות לפרויקט זה.

7.4. <u>מבנה הencoder-decoder:</u>

כפי שניתן לראות בתרשים המצורף, ה-encoder-decoder מורכב מ-3 חלקים:



(a) Conversion (b) Training

1. **חלק Ec - 1** , זה encoder המקבל דיבור של דובר א׳ (דובר התוכן) ומוציאה קידוד של **תוכן הדיבור**. לפי המאמר, מכניסים קלט בגודל 80 (מספר ה-mel) לרשת של 3 שכבות קונבולוציה בגודל 512, כשלאחר כל שכבת קונבולוציה יש שכבת נרמול וReLU. לאחר מכן, המוצא עובד דרך LSTM דו כיוונית בגודל 32+32=64 (forward+backwards) . יש דצימציה בסדר של 32 לשני הכיוונים אך יש משמעות שונה כאשר מדובר בforward לעומת הbackward (אם לוקחים את הדגימות הראשונות של כל חלון או האחרונות).
2. **חלק Es - 2** , זה encoder המקבל דיבור של דובר ב׳ (דובר הסגנון) ומוציאה וקטור אמבדינג של **הסגנון הדיבור** של דובר ב׳. לפי המאמר, ה-encoder של הסגנון בנוי משתי שכבות של LSTM בגודל 768 וגודל הפלט הוא וקטור אמבדינג בגודל 256x1. את ה-encoder הם אימנו על הדאטהסטים : Librispeech 'VoxCeleb1. ההגיון של חלק זה הוא שעבור שני קטעי דיבור שונים של אותו הדובר, יתקבל אותו הפלט (וקטור ייצוג של סגנון הדיבור, בלתי תלוי בתוכן המדובר).
3. **חלק D - 3** , זה ה-decoder אשר מקבל את הקידוד של התוכן ואת וקטור האמבדינג של הסגנון מהencoder הראשון והשני בהתאמה, ומוציאה פלט של דיבור על פי התוכן והסטייל שהתקבלו.

הארכיטקטורה של ה-decoder בנויה באופן של קונקטינציה (שרשור) של וקטור הסגנון על כל קידוד של תוכן, והעברת הוקטור המאוחד המכיל את הסגנון והתוכן דרך 5 שכבות קונבולוציה בגודל 512, שלאחר כל שכבה יש נרמול ו-ReLU. לבסוף הרשת מסיימת עם מעבר דרך 3 שכבות LSTM בגודל 1024 אשר מוציאות פלט בגודל 80 שזה מספר ה-mels.

בשלב אימון הרשת ובשלב ה-test יש שינויים בקלטים המתקבלים בכל אחד מהרכיבים.

בשלב האימון, הרשת מקבלת את אותו הדובר עבור שני החלקים: גם התוכן וגם הסטייל. לאחר שהיא מצמצמת את הקלטים לוקטורי האמבדינג, היא משחזרת את קובץ הדיבור ומשווה לקובץ המקורי. כיוון שההחלפה נעשתה בין דובר לעצמו, שחזור התהליך צריך להיות שווה לקובץ המקורי.

בשלב ה-test, מכיוון שרוצים להחליף בין דוברים, אזי מקבלים קלטים של דוברים שונים עבור התוכן ועבור הסטייל, ובתקווה, המודל יצליח לייצר פלט שהוא החלפה בין הסטייל של הדוברים.

ישנם עוד 2 שלבים להשלמת תהליך ההחלפה:

1. רשת נוספת - postnet: על מנת לשחזר גם את הפרטים הקטנים ביותר של הספקטרוגמרה, הציגו במאמר רשת נוספת הבנויה מ-5 שכבות של קונבולוציה, עם שכבת נרמול ו-tanh לאחר כל שכבת קונבולוציה. מוצא הרשת הנוספת הוא השארית שאמורה להשלים בין ההחלפה האידיאלית לבין ההחלפה שבוצעה בפועל. הפרש זה גם מחושב ברשת כפונקציית מחיר שלישית המרכיבה את כלל פונקציית המחיר.
2. מעבר מספקטרוגמה חזרה לאודיו: מעבר זה אינו טריוויאלי כלל שהרי במעבר ההפוך - אודיו לספטרוגמרה אנו שומרים רק את הערך המוחלט, מה שגורם לאיבוד הפאזה של האות המקורי. ומעבר לכך, אנו מחליפים בין דוברים, אז גם שמירת הפאזה לא הייתה יכולה בהכרח לעזור כי היא לא מתאימה לדובר החדש. השחזור שמוצע במאמר משתמש ב- *Wavenet* - רשת שפורסמה ב-2016 ומאז נהייתה מאוד נפוצה ברשתות העוסקות בגנרציית דיבור. היתרונות שלה בין היתר הן איכות השחזור, המהירות בה היא מתבצעת ביחס לשיטות נפוצות שקדמו לה, וגם הכמות המינימלית של דיבור שהרשת מצריכה על מנת ללמוד היטב את אופי הסגנון של הדיבור.

7.5 **אופן הלמידה של הרשת:**

במאמר זה צוין כי יוצאים מנקודת הנחה שאת וקטור האמבדינג של הסגנון הרשת יודעת לעשות מספיק טוב, וכי אין צורך להמשיך לאמן אותה.

** הערת אגב, כאשר נפרט בהמשך את השיטות שלנו להתאים את הרשת לשירה, ואת הצעדים שעשינו על מנת לשפר את התוצאות, החלפת ה-encoder של סגנון הדיבור (כלומר האמבדינג) היה חלק מאוד משמעותי.

את פונקציית המחיר של הרשת מרכיבים שני פונקציות מחיר שונות: $L = Lrecon + \lambda Lcontent$

החלק הראשון - פונקציית מחיר שמחושבת על ידי ההפרש בין הקובץ המשוחזר של התוכן בסגנון הסטייל, כאשר שניהם לקוחים מאותו הקובץ המקורי, ולכן הפלט אמור להיות זהה לחלוטין לקלט:

$$,[Lrecon = E[||X\hat{}1\rightarrow 1 - X1||^2$$

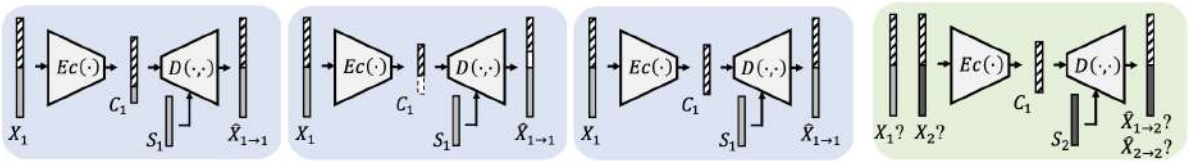
החלק השני - פונקציית מחיר מחושבת על ידי ההפרשים בין קידוד של התוכן של המקור ושל הקובץ המשוחזר. בעצם, בתהליך הצמצום והשחזור, מתבצעת הורדת מימד, ואז העלאת מימד - שחזור. את וקטור קידוד שומרים ומשווים לוקטור של הקובץ המשוחזר:

$$/Lcontent = E[||Ec(X\hat{}1\rightarrow 1) - C1||]1$$

במאמר החליטו שפונקציית המחיר הראשונה (השוואת השחזור עצמו) תהיה נורמה 2 (MSE), פונקציית המחיר על הקידודים של התכנים תהיה נורמה 1.

למה זה עובד?

אחד המרכיבים הרגישים והמתוכננים היטב במבנה של הרשת הוא צוואר הבקבוק שבין ה-encoder ל-decoder. הרעיון מאחורי הגודל האופטימלי הוא שמצד אחד אנו מעוניינים להעביר מספיק מידע לגבי התוכן ולכן צוואר הבקבוק צריך להיות מספיק גדול, אך מנגד לא נרצה שיותר מדי מידע יעבור, מה שיכיל גם מידע על סגנון דובר המקור. לאחר מציאת הגודל האופטימלי, מאמנים את ה-encoder לקחת רק את החלק המייצג את תוכן הדיבור. דבר זה נעשה משום שה-decoder מקבל שוב וקטור אמבדינג של סגנון הדיבור ובכך הוא מוותר על המידע שנמצא בקובץ הקורי המתאר את סגנון הדיבור.



(a) Bottleneck too wide (b) Bottleneck too narrow (c) Bottleneck just right (d) Conversion

את המצבים השלכות של גודל צוואר הבקבוק ושיטת הלמידה של הרשת ניתן לראות באיור לעיל.

** מלבן חלק מייצג את הסגנון ומלבן פסים מייצג את תוכן הדיבור **

בתרשים a, צוואר הבקבוק גדול מדי ולכן יתקיים שחזור מושלם אך עם מידע נוסף ומיותר. הסיבה שתתקיים שחזור מושלם למרות שצוואר הבקבוק לא מושלם הוא כיון שברמה עקרונית הייתה אמורה להיות "התנגשות" בין סגנון הדיבור של המקור לבין סגנון הדיבור של היעד. בשלב הtrain, דובר היעד ודובר המקור הם אותו הדובר (עושים זאת בשביל חיושב פונקציית המחיר כי אנחנו יודעים איך התוצאה צריכה להישמע) ולכן אין באמת התנגשות בין הסגנונות כי זה אותו האדם (אותו וקטור אמבדינג של הדובר).

בתרשים b, ניתן לראות את ההשלכות של צוואר בקבוק קטן מדי (לא לומדים את התוכן).

בתרשים c ובתרשים d ניתן לראות שמתבצעת החלפה טובה עם אותו הקובץ עצמו וגם עם דובר חדש.

בסיכום המאמר כמובן מפרטים שהשיטה המוצעת משיגה את השיטות הקיימות כיום בהחלפת סגנון דיבור. אכן ראינו בעצמנו כאשר הרצנו את הרשת אצלנו שניתן לייצר תוצאות טובות מרשת זו, אך לא מספיק טובות בשבילנו.

7.6. **הסיבות שבגללן בחרנו לבצע שינויים ברשת:**

- התאמה להחלפת שירה - כמו שפירטנו במספר מקומות לעיל, ההחלפה של דיבור לדיבור מול דיבור לשירה אינה אותו הדבר ומצריכה התאמות במבנה הרשת ובחיבור החלקים השונים של המערכת.
- למרות שהפלטים שהוצגו באינטרנט נשמעו יחסית טוב, כאשר הרצנו בעצמנו את המודל מבלי לשנות כלום, לא הצלחנו לשחזר תוצאות באותה רמה כמו שפורסמו. ומעבר לכך, גם את התוצאות של המאמר רצינו לשפר עוד יותר בשביל תוצאות איכותיות יותר.
- גם כאשר הצלחנו לייצר תוצאות שנשמעו סבירות, משך הזמן שלקח לאימון ולביצוע ההחלפה היה ארוך מאוד. מטרה ברורה שהצגנו לעצמנו בתחילת הפרויקט בדיוק הייתה לשפר גם כן את מהירות ההחלפה.

7.7. **השינויים שלנו ברשת המתוארת במאמר:**

7.7.1. **החלפת ה-encoder של הסגנון - ה-embedding:**

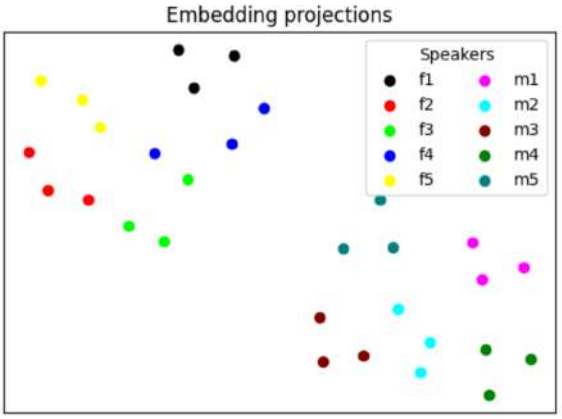
מטרת העל של ה-encoder היא לקבל קלט של אדם א׳, ולייצר פלט של וקטור בגודל קבוע המייצג את סגנון הדיבור של אדם א׳ - embedding. ההצלחה של ה-encoder תימדד בכך שעבור שני קלטים שונים, וקטור ה-embedding יהיה זהה במידה ושני הקלטים היו דיבור של אותם אדם, ושונים כאשר מדובר בשני בני אדם שונים. בקוד המתבסס על המאמר ישנו גם המימוש של ה- encoder עבור הסטייל/אמבדינג. בנוסף ישנם גם מספר פלטים לדוגמא כחלק מקובץ metadata.pkl - קובץ המכיל גם את הספקטרוגרמות של השירים כתפקיד התוכן וגם את וקטורי הembedding של הדוברים - מה שרלוונטי לפסקה זו.

אנחנו החלטנו להחליף את שיטת ייצור ה-embedding לכלי אחר - **Resemblyzer**. כלי זה הוא גם כן encoder המיועד לייצר embedding המייצג סגנון דיבור מסיום.

היו מספר סיבות שגרמו לנו להחליף ביניהם:

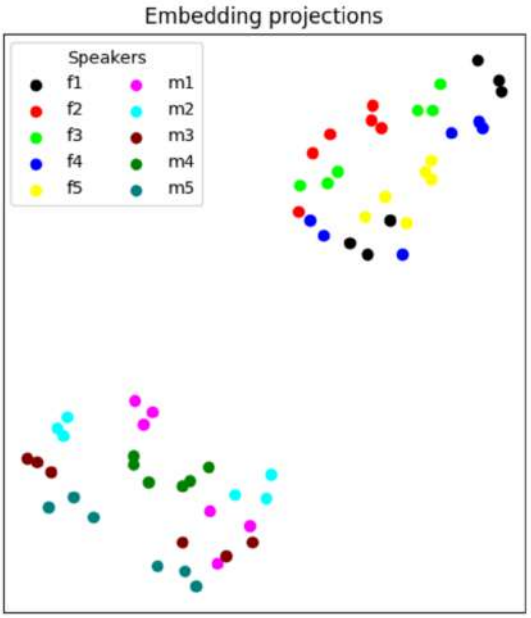
- אמינות השיטה המקורית: ראינו שכאשר אנחנו מנסים לייצר embedding באמצעות הכלים שסופקו בקוד, לא קיבלנו בהכרח אותם תוצאות שצורפו גם כן עם הקוד כדוגמה לפלטים שיוצרו על ידי הכותבים של הקוד. אך בשורה התחתונה רצינו להשתמש ברשת embedding שאנחנו יכולים להיות יותר בטוחים בטיב התוצאות שלה.
- הצורך לשפר את הרשת: כיוון שה-Resemblyzer הוא כלי יחסית נפוץ שנמצא בשימוש ועדכון כמעט יומיומי, הרשת משתפרת ומתייעלת כל הזמן. לעומת זאת, הרשת שסופקה לנו כחלק מהקוד אומנה ע"י כותבי הקוד, אנו לא בטוחים אם היא המשקולות שניתנו לנו עבורה הם אכן במצב מספיק טוב. בנוסף במידה והיינו רוצים להמשיך לאמן או לשפר אותה היינו צריכים לעשות זאת בעצמנו, דבר שלא רצינו להעמיק בו. שימוש בכלי Resemblyzer מאפשר לנו להשתמש בכלי עם מתוך ספרייה מובנית, שעדיין ממשיכים לשפר אותה ולעדכן אותה.

את ה-encoder החלופי Resembyzer בדקנו על מספר זמרים וזמרות שונים על מנת לאמת את טיב הרשת. נציג כעת תוצאות של איכות ה-embedding. את התוצאות נציג בשיטת clustering של KMeans והורדת מימד ל-2 מימדים.



כפי שניתן לראות בגרף הראשון, ישנה הבדלה ברורה בין גברים לנשים: יש 2 קלאסטרים מובהקים המכילים כל אחד חצי מסך הדאטהסט, מופרד באופן ברור בין הגברים לשנים. יתר על כך, ניתן לראות שישנה הפרדה בין כל אחד מהזמרים והזמרות. כמובן שזו הפרדה במימד נמוך ממימד האמבדינג המקורי, וזו לא הפרדה מובהקת כמו בהפרדת המינים, אך פחות או יותר ניתן לראות שהנקודות בצבעים הדומים קרובים יותר אחד לשני.
נציין גם שכאשר שמענו שוב את הזמרים השונים, הייתה התאמה בין קרבת הנקודות בעלי הצבעים השונים לזמרים וזמרות שנשמעו לנו יחסית דומים.

גרף נוסף, המראה את ההבדלים בין דיבור ושירה:



בגרף זה ניתן לראות את התוצאה כאשר ניסינו שוב להשתמש ב-encoder לייצר ולייצג בגרף. השוני כאן הוא שלקחנו עבור כל דובר 3 קטעים של דיבור ו-3 קטעים של שירה. ניתן לראות שיש הפרדה בין הקטעים של הדיבור לבין הקטעים של השירה. הבנו מכאן שה-encoder לא מייצג את הדיבור והשירה של אותו האדם על ידי אותו הוקטור. עניין זה עלול להוות בעיה עבורנו שהרי אנו מעוניינים לייצר שירה של אדם מתוך הדיבור שלו בלבד ואם אנו מזהים שיש הבדל ב-embedding של הדיבור לעומת השירה אזי הפלט יהיה בסגנון הדיבור של האדם ולא בסגנון השירה שלו. טרם הצלחנו להתגבר על בעיה זו, אך אנו לא בטוחים עד כמה בעיה זו קריטית לאיכות התוצאות.

7.7.2 **החלפת ה-vocoder:**

אחד המרכיבים המאתגרים בשלב זה הוא דווקא השלב האחרון ביצירת קובץ השמע: שלב ה-vocoder. תפקידו הוא קבלה כקלט מערך דו מימדי : הספטקרוגרמה של הקובץ (ספקטרוגרמה של הדובר יעד שר את תוכן שיר המקור, או במילים אחרות - ספקטרוגרמה של השלב לאחר ההחלפה) ולהמירו חזרה לקובץ אודיו.
אנו ראינו שלמרות שרשת wavenet צוינה במקומות רבים כרשת אמינה וטובה, הזמן שלקח לרשת היה ארוך מאוד מבחיתינו. משך הזמן הארוך לא היה בעיה רק בגלל הבדיקות וההשוואות של הרשתות המאומנות השונות שלנו, אלא שוב במחשבה שהטכנולוגיה שלנו תספק תוצאות בזמן המהיר ביותר ידענו שבסוף נצטרך שיטה הרבה יותר מהיר ולא פחות איכותית.
לאחר מחקר של החלופות השונות גילינו שהספריה הנפוצה לעיבוד ותכנות של אודיו בפייתון, "Librosa" מכילה פונקציה המבצעת את אלגוריתם Griffin-lim.

אלגוריתם Griffin-Lim או GLA הינו אלגוריתם לשחזור פאזה מתוך יתירות של ה-STFT. אלגוריתם זה אינו מצריך מידע מקדים על הספקטרוגמרה או על הסיגנל המקורי. האלגוריתם מצפה לשחזר ספקטרוגרמה בעלת ערכים מרוכבים, כך שהשחזור עקבי עם האמפליטודה הנתונה.

$$\mathbf{X}^{[m+1]} = P_{\mathcal{C}}\left(P_{\mathcal{A}}\left(\mathbf{X}^{[m]}\right)\right)$$

עיקר האלגוריתם הוא - כאשר מבצעים המרה ל-stft מקבלים אמפליטודה ופאזה, ולאחר המעבר לספקטרוגרמה נשאר רק אמפליטודה (בגלל הערך מוחלט). עכשיו מנסים לשחזר את הפאזה, אז מנחשים פאזה כלשהי ועובדים באיטרציות. "מנחשים" פאזה ולאחר מכן בודקים אם לקיחת ערך מוחלט על האמפליטודה והפאזה נותן את אותה אמפליטודה, כפי שקיבלנו במקור. כך ממשיכים באיטרציה ומעדכנים בכל פעם את הפאזה.

עבור X, ערך של הספקטרוגמרה המשוחזרת (X בעל ערך מרוכב), הערך שלו מתעדכן על ידי איטרציות שונות של האלגוריתם. P הוא ההטלה של השחזור על סט מסוים ו-m הוא מספר האיטרציה. ישנם חישובים מעמיקים יותר של ההטלות שכוללות הכפלת ערך התמרת ה-STFT עם ה-iSTFT אך כיוון שזה לא הנושא העיקרי של הפרויקט, לא רצינ להתעמק בזה כאן.

המטרה העיקרית של ההחלפה ל-vocoder הזה הייתה קיצור זמן ההמרה מספקטוגרמה חזרה לאודיו. ראינו שבאמצעות האלגוריתם של Griffin-Lim אנו מצליחים לקצר באופן מאוד משמעותי את הזמן של כל התהליך (מסדר גודל של ~10 דקות לפרק זמן של פחות מ-~10 שניות! עבור פלט שאורך בין 7-10 שניות) .
ראינו גם כן שאיכות ההמרה לא נפגמת באופן שניכר לנו כמאזינים ולכן יישמנו את אלגוריתם זה בחלק מובנה ממערכת השחזור.

7.7.3. **שימוש ב-stft ולא ב-mel-spectrogram:**

אחד השינויים המשמעותיים ביותר שביצענו בקוד המקורי היה סוג הקלט עבור הרשת.

הרשת המקורית תוכננה לקבל ספקטרוגמות לא לינאריות אלא מותאמות לסקאלה של mel, מה שנקרא mel-spectrogram. למרות שלא הסבירו בפירוט מה התכלית של ההמרה מספקטרוגארמה למל- ספקטרוגמרה, אנו משערים שהשימוש ב- mel הוא מכיוון שנהוג בעיבוד קול ובעיקר דיבור לעבוד בספטרוגמרת מל . הסקאלה של ב-mel אינה לינארית והיא מרווחת יותר את התדרים הנמוכים על חשבון התדרים הגבוהים, כך באמצעות אותו מספר של ערכים ניתן לקבל רזולוציה גבוהה יותר עבור תדרים נמוכים, שהם התדרים הרלוונטיים יותר בדיבור מבלי לאבד את הערכים של התדרים הגבוהים (התדרים הגבוהים יהיו ברזולוציה נמוכה יותר אך העדיפות במקרה של עיבוד קול הוא בתדרים שבהם הקול האנושי יותר נוכח בהם).

החלטה זו - להשתמש בספקטרוגרמת-מל ולא ספקטרוגרמה סטנדרטית השפיע על כל שאר הארכיטקטורה של הרשת. אחת הסיבות לכך היא כיון שהוחלטת על 80 ערכים בציר התדר לכל חלון זמן בספקטרוגרמת-המל. ערך זה היה נמוך מדי לדעתינו ורצינו לשנות את זה. הבעיה השנייה הייתה שלמרות שהצלחנו להקטין משמעותית את זמן הריצה של ה- vocoder, באמצעות אלגוריתם Griffin-Lim עדיין היה עיכוב מאוד משמעותי בהמרה מספקטרוגרמת-מל לספקטרוגרמת stft, כי אלגוריתם Griffin-Lim פועל רק על ספקטרוגמרה stft. כאשר החלפנו את כל המערכת שתהיה מותאמת לספקטרוגרמה סטנדרטית הצלחנו לשפר את שתי הבעיות הללו: גם קיבלנו רזולוציה גבוהה יותר עבור כל התדרים (לא היינו מוגבלים ל-80 ערכים בתדר), וגם חסכנו את ההמרה אל וחזרה מספקטרוגרמת-מל ל-stft.

7.7.4. **אימון על Datasets נוספים:**

7.7.4.1. אימון על מאגרים קיימים

הקוד המקורי מציין כי במהלך האימון של הרשת, הוא נעזר בדאטה-סט מאוד נפוץ בשם VCTK - Voice Cloning ToolKit[10]. הדאטה-סט הזה פורסם על ידי אוניברסיטת אדינבורו שבסקוטלנד ומכיל 109 דוברי אנגלית בעלי מבטאים שונים. כל דובר מקריא 400 משפטים שונים בנוסף לקטע זהה על מנת לזהות את המבטא של הדובר.

למרות שהדאטה-סט הזה היה מאוד מרשים, רצינו לנסות ולשפר את התוצאות שהוצגו במאמר ולכן ניסינו גם להיעזר בדאטהסטים שונים. אחת הנפוצות ואיכותיות ביותר שמצאנו, דווקא כאשר התמקדנו בשיטה אחרת לפרויקט לפני שהגענו שליטה הזו, נקראת NHSS.

הדאטה NHSS הוא מאגר שפורסם על ידי אוניברסיטת סינגפור. מלבד האיכות של ההקלטות (הוקלטו באולפן מקצועי) וכמות ההקלטות (סך ההקלטות אורך כ-7 שעות), הייחודיות של דאטהסט זה על פני האחרים הנפוצים היא שזה דאטהסט מקבילי: עבור 10 זמרים יש 20 שירים אותם הדוברים גם מדברים (הקפדה רק על תוכן המילים ולא התזמון בהתאם לשיר או לטון המילים) בנוסף לקטעים בהם אותם הדוברים ממש שרים את השירים כמו בביצוע המקורי שלהם.

[10] Yamagishi, Junichi; Veaux, Christophe; MacDonald, Kirsten. (2019). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2645.

החלטנו לא להתבסס רק על דאטהסט אחד אלא לנסות לאמן את הרשת על שני הדאטהסטים האלה - VCTK, NHSS ובנוסף לאמן חלק על ההצלבה ביניהם - זאת אומרת להתחיל לאמן את הרשת על אחד הדאטהסטים ולהמשיך לאמן על השני, ובכך אנו יכולים לאמן את הרשת על החלפה של דיבור (בעזרת VCTK שהוא דאטה גדול), ולאחר מכן לבצע סוג של fine-tuning למשימה של שירה בעזרת הדאטה NHSS. מעבר לכך, אימון על דאטה נוסף עשוי למנוע מצב של overfitting עבור אחד מהם.

7.7.4.2. <u>יצירת מאגר שירים חדש ואימון מכונה</u>

על מנת ליצור מכונה שמצליחה להחליף דוברים חדשים על שירים קיימים ומוכרים, יצרנו מאגר חדש. אספנו שירים מוכרים (באנגלית וגם קצת בעברית) ובעזרת כלים קיימים[11] הפרדנו בין השירה ובין הליווי. לאחר שאספנו מספיק שירים (בזמן מצטבר של כמה שעות) אימנו מכונה חדשה על השירים הללו.

את האימון על הדאטה החדש עדיין לא סיימנו בזמן כתיבת ספר הפרויקט, ואנו מחכים לתוצאות בקרוב.

7.8. **<u>שינוי ובדיקת היפר-פרמטרים, ותצורות שונות לאימון הרשת:</u>**

על מנת לנסות להשיג תוצאות אופטימליות, ניסינו לאמן ולהריץ את הרשת בתצורות שונות, ועם היפר-פרמטרים שונים.

7.8.1. <u>בדיקת מימדים שונים לצוואר הבקבוק:</u>

במהלך אימון הרשת ניסינו לבדוק גדלים שונים של צוואר הבקבוק עבור ה- encoder:
ניסינו את הגדלים הבאים: 16,32,40,64,128. עבור כל גודל היו יתרונות וחסרונות.

16 - כיוון שפרמטר צוואר הבקבוק הוא בעצם פרמטר הדילול, כלומר ביחס הפוך לגודל האפקטיבי של צוואר הבקבוק, אזי 16 הוא צוואר בקבוק הכי גדול. כתוצאה מכך, בדוגמאות שבדקנו עם גודל זה, ה-encoder מעביר יותר מדי מתוך האופי של הזמר המקורי.

32 - ערך של 32 היה ערך ברירת המחדל של המודל מתוך הקוד שפורסם עבור המאמר, ואכן כאשר בדקנו מספר תוצאות של מודל מסוג זה התוצאות המתקבלות היו דיי טובות. כאשר מבצעים החלפה בין דוברים מאותו המין, התקבלו תוצאות טובות, ללא קפיצות אוקטבה וגם שינוי הסטייל של הדובר בוצע באופן די טוב. לעומת זאת, כאשר ניסינו להחליף בין דוברים ממין שונה התוצאות היו פחות טובות מאשר המודל עם פרמטר צוואר בקבוק 64.

40 - המודל שאומן עם הערך הזה היה הכי פחות מוצלח, בין היתר כי מסיבה שעדיין לא ברורה כל כך הוא לא ייצר פלטים באורך המלא של השיר, וגם ההחלפה בין הדוברים לא הייתה טובה, ומלאה בקפיצות אוקטבה וזיופים.

[11] Spleeter: A Fast And State-of-the Art Music Source Separation Tool With Pre-trained Models, https://paperswithcode.com/paper/spleeter-a-fast-and-state-of-the-art-music

64 - ערך זה של צוואר הבקבוק היה מבן המוצלחים, יחד עם ערך של 32. נתן תוצאות די דומות לערך של 32, אולם כאשר ניסינו להחליף בין דוברים ממינים שונים, התוצאות היו טובות יותר מאשר במודל עם ערך של 32.

7.8.2. <u>גדלים שונים של ספקטוגרמות</u>

כפי שגם תיארנו בחלק של הכנת הדאטה למודל, ניסינו לשנות את סוגי הספקטוגרמות (mel-spectrogram, stft) אך גם את גודל הספקטוגרמות: 80,128,256,512,1024.

בקוד המקורי הממש את המאמר, היה שימוש ב-mel-spectrogram בגודל 80. אולם כאשר ניסינו לאמן מכונות עם סוג הספקטוגרמה הזו, התוצאות לא היו טובות מספיק לדעתנו (וראו חלקים מתאים בספר). לכן בדקנו אפשרות לאמן את הרשת גם על גדלים שונים של ספקטוגרמות מסוג stft.
בדקנו את איכות המעבר מספקטוגרמה חזרה לאודיו עבור מספר גדלים שונים - 128,256,512,1024. ראינו שעבור ספקטוגרמות הקטנות מ-256 ניתן לשמוע באוזן ירידה באיכות. מעל 256 האיכות הייתה טובה, ללא הפרש ניכר בין הגדלים. לכן בחרנו להשתמש בספקטוגרמות של 256.

7.8.3. <u>סוגי דאטה שונים, והצלבות ביניהם</u>

כפי שהסברנו בחלקים קודמים בספר, על מנת לאמן את הרשת מצאנו מספר מאגרים, השניים בהם השתמשנו לאימון הרשת היו VCTK ו-NHSS, ובנוסף אימנו את הרשת גם על מאגר שירים שאנו יצרנו.
את הרשת אימנו מספר פעמים בתצורות שונות על כל מאגר בנפרד, ובנוסף גם אימנו מכונות על הצבה של המאגרים. כלומר התחלנו אימון על מאגר מסוים והמשך לאחר מכן על המאגר השני.

7.9. <u>ניתוח התוצאות שלנו:</u>

כפי שכתבנו, את הרשת אימנו על datasets קיימים וידועים (VCTK,NHSS) ועל dataset חדש שיצרנו.
על הדאטה המוכרים, התוצאות היו מגוונות, היו החלפות שהתבצעו בצורה טובה והיו כאלה שפחות.

על **VCTK, NHSS** כאשר מנסים להחליף דוברים ממינים שונים, ההחלפה פחות טובה.
בעיקר בהחלפה אישה לגבר זה לא נשמע כמו גבר. כאשר מחליפים גבר לאישה, נשמע כמו אישה אך יש קפיצות אוקטבה. כאשר מחליפים בין דוברים מאותו המין, ההחלפה מתבצעת בצורה טובה יותר, עם פחות קפיצות אוקטבה.

בכל ההחלפות, יש עדיין מעט בעיה עם תוכן השירה, הזמרים לא תמיד ברורים והמילים שהם שרים לא תמיד ברורות.

ב-dataset **החדש** שיצרנו כאמור השתמשנו רק לאחרונה, וניתוח התוצאות נעשה בזמן כתיבת הספר, ולכן עדיין לא נוכל להכניס אותן[12].

[12] ראה בחלק נספחים קישור לתיקיית דרייב בה נמצאים התוצאות של הפרויקט.

8. **מסקנות וסיכום**

הרעיון הראשוני עליו חשבנו לפני שנה, סינתזת שירה מתוך דיבור של אדם, הוביל אותנו למחקר מעמיק בנושא. חקרנו בנושא דיבור ושירה באופן כללי, כל אחד בנפרד, ועל הקשר ביניהם. למדנו על צורות קלאסיות בהן ניתן לנתח אותם, ועל ההבדלים שניתן לראות ביניהם.

התמודדנו עם בעיה שבתחילת הדרך נראתה לא פתורה עדיין, ושלא רבים חקרו לפני כן. לאורך הדרך ראינו מאמרים ומחקרים בנושא, אך נראה שאין עדיין כלים המבצעים את המשימה הזו בדיוק בצורה טובה ומהירה כפי שניסינו לשים לנו כמטרה בפרויקט.

לפי דעתנו, הרעיון לפרויקט היה מאתגר מאוד, אך מצד שני הוא היה מאוד מלמד וניתן לומר שהרעיון מוצלח. אנו סבורים כי המשך עבודה בנושא, הן בכלים קלאסיים והן בכלים של deep learning, יביאו לתוצאות טובות יותר, ושהבעיה ניתן לפתרון ברמה טובה ביותר.

בהשוואה לפתרונות הקיימים האחרים שמצאנו נכון להיום, לפי דעתנו הפתרון שמימשנו הוא המהיר ביותר וזה שגם מבצעת החלפה בצורה טובה. יתר על כן, אנו מבין הבודדים שנותנים פתרון לבעיה של סינתזת שירה מתוך **דיבור.**

9. **רעיונות להמשך**

רעיונות לטווח הקרוב

- להרחיב את מאגר השירים שיצרנו עליו הרשת לומדת, ואף לנסות ליצור כלי אוטומטי אשר מרחיב את מאגר המידע (זאת כיוון שאנו יצרנו את המאגר בצורה חצי-ידנית).
- להמשיך לחקור מהו משך הזמן המינימלי הדרוש על מנת לייצר וקטור אמבדינג אמין מספיק.
- להמשיך את מימוש הקוד שהמחליף בין אנשים, כך שיעבוד על כל דובר חדש ועל כל שיר חדש בהרצת קוד בודד פשוט.

רעיונות לטווח הרחוק

- לנסות לשנות ארכיטקטורה של הרשת, למשל ב-encoder עצמו, למשל להוסיף עוד שכבות קונבולוציה, או אפילו סוג של רשתות למידה עמוקה.

10. **ביבליוגרפיה**

A Neural Algorithm of Artistic Style", arXiv:1508.06576 [cs.CV] ",
https://github.com/lengstrom/fast-style-transfer

Blaauw M, Bonada J. A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs. *Applied Sciences*. 2017; 7(12):1313.
https://doi.org/10.3390/app7121313

Coverage of deepdub.ai on N12 Israel News,
https://www.youtube.com/watch?v=ma-IS96GUWs

"FIPS PUB 137, Analog to Digital Conversion of Voice by 2,400 Bit/Second Linear Predictive Coding" (PDF). National Institute of Standards and Technology. Retrieved 2018-08-17.

Qian, K., Zhang, Y., Chang, S., Yang, X. & Hasegawa-Johnson, M.. (2019). AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss.*Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 97:5210-5219 Available from https://proceedings.mlr.press/v97/qian19c.html

Resemblyzer: https://github.com/resemble-ai/Resemblyzer
https://pypi.org/project/Resemblyzer/

Spleeter: A Fast And State-of-the Art Music Source Separation Tool With Pre-trained Models,
https://paperswithcode.com/paper/spleeter-a-fast-and-state-of-the-art-music

Wilkins, Julia et al. "VocalSet: A Singing Voice Dataset." *ISMIR* (2018). , VOCALSET: A SINGING VOICE DATASET

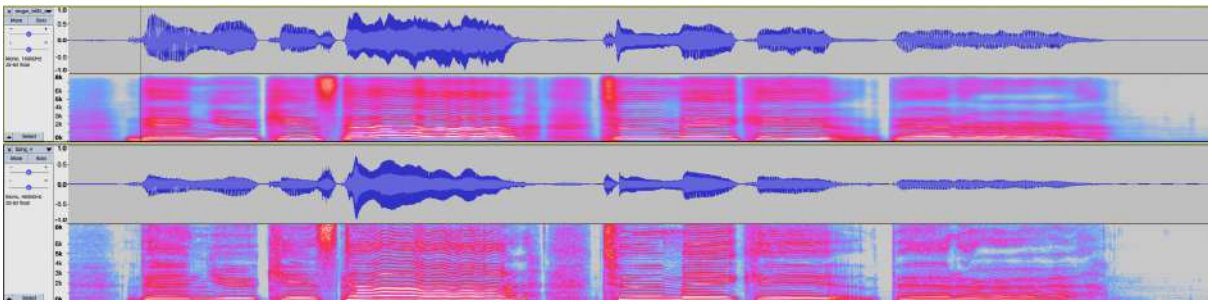
Wilkins, Julia, Prem Seetharaman, Alison Wahl, & Bryan Pardo. (2018). VocalSet: A Singing Voice Dataset (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1193957

Yamagishi, Junichi; Veaux, Christophe; MacDonald, Kirsten. (2019). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR).
https://doi.org/10.7488/ds/2645.

11. **נספחים**

נספח 1

2 ספקטוגרמות - "סגירת מעגל". כלומר זמר 02 מתוך הדאטה NHSS שר את שיר 05 , החלפה עם אותו זמר על מנת לייצר שוב את הקטע המקורי (עליון מסונתז, תחתון מקורי). החלפה בעזרת הרשת autovc שאימנו בעצמנו.



נספח 2:

בקישור הבא תוכלו למצוא תוצאות מתוך הפרויקט:

https://drive.google.com/drive/u/0/folders/1mW2Hzyx6AoilZ3YnZMyAQdRqgr8pkfni