סיכום פרויקט סיום קורס

1. הגדרת הבעיה/ עולם הבעיה/ מה רוצים לפתור/ מה מצפים שיהיה

Approximately two-thirds of speech is voiced which has important intelligibility property. Invoiced speech is caused by air passing through a narrow constriction of the vocal tract as when consonants are spoken, which is non-periodic, random-like sounds. Because of the periodic nature of voiced speech, it can be identified, and extracted more precisely than unvoiced speech [1]

Speech can be divided into numerous voiced and unvoiced regions. The classification of speech signal into voiced, unvoiced provides a preliminary acoustic segmentation for speech processing applications, such as speech synthesis, speech enhancement, and speech recognition. [3]

ניתן לראות שהבעיה של הבדלה בין חלקים בהקלטה של דיבור ל voiced/unvoiced זו בעיה ידועה, ואנחנו רוצים למצוא דרך לזהות אותה בצורה הכי מדוייקת.

ניתן להשתמש בזה בהרבה מקומות, אחת מהם למשל היא בזיהוי אותיות, ניתן לשלב את המידע שנקבל אם הקטע הוא קולי או לא יחד עם פרמטרים אחרים על מנת להבדיל בין אותיות.

אנחנו נרצה לאפיין קטע קולי ב DATA שלנו, ולהציג אותו בsignal.

2. קריאת מאמר או חומר נלווה להבנת התחום

יש כמה דרכים שאיתם יודעים להתמודד עם בעיה זו. פתרון שלנו מתבסס על:

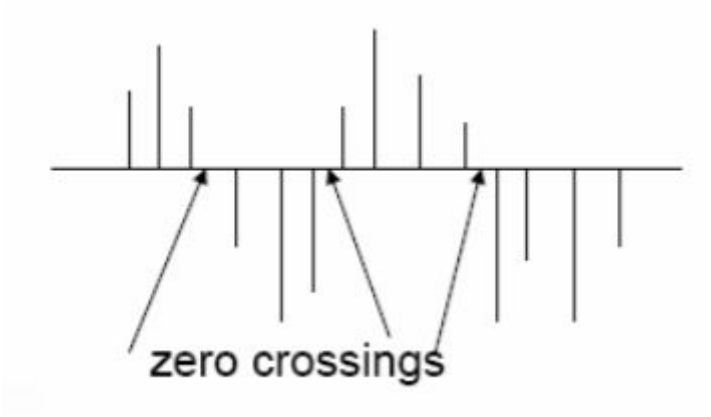
ZCR - Zero Crossing rate

ועל Short Term Energy (STE).

ניתן לסווג חלק בדיבור על ידי מדידה של אותם החלקים, ולהסיק האם זה קטע קולי או לא על ידי שילוב המידע.

ZCR - Zero Crossing Rate

הגדרה: מספר הפעמים שבה האמפליטודה של האות הנדגם מגיעה ל 0 באינטרבל של זמן או פריים.



ניתן לחשב על ידי:

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]\right| w(n-m) \qquad (1)$$

where

$$\operatorname{sgn}[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases} \qquad (2)$$

and

$$w(n) = \begin{cases} \frac{1}{2N} \, for, 0 \leq n \leq N-1 \\ 0 \, for, otherwise \end{cases} \qquad (3)$$

STE - Short Term Energy

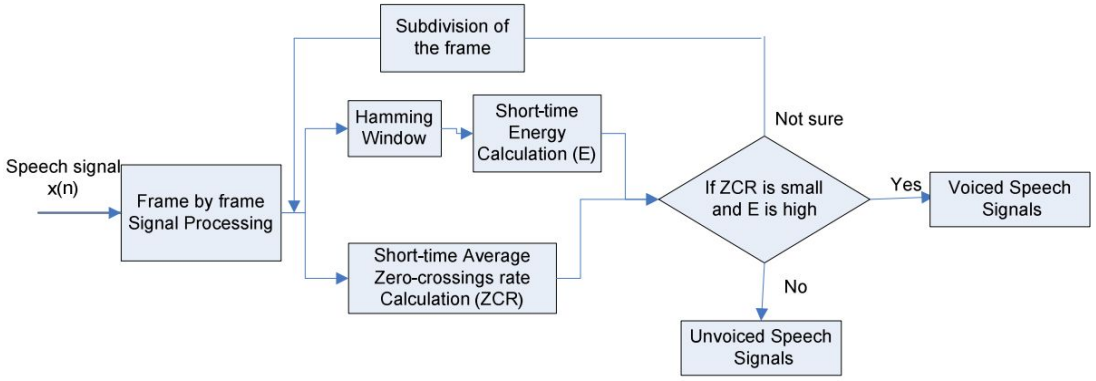
אמפליטודת ההקלטה הנדגמת משתנה בזמן, באופן כללי, אמפליטודה של קטע לא-קולי הרבה יותר נמוכה מאשר קטעים קוליים. אנרגיה של קטע דיבור נותנת לנו ייצוג שמשקף לנו את ההבדלים באמפליטודות.

מגדירים אנרגיה של קטע קולי על ידי סכום של מכפלות בריבוע (אמפליטודה כפול החלק היחסי של אורך הקטע הנדגם - החלק של האמפליטודה, מאורך כלל הפריים הנדגם) כפי שמופיע במשוואה הבאה:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2$$

זיהוי קטעים קוליים - בדיבור בדרך כלל האנרגיה גבוהה יותר בקטעים קוליים בתדרים מתחת ל 3khz, לעומת זאת בקטעים לא קוליים, רב האנרגיה נמצאת בתדרים הגבוהים.

מכיוון שתדרים גבוהים מאופיינים על ידי ZCR גבוה, ותדרים נמוכים על ידי ZCR נמוך, ישנה קורלציה בין והאנרגיה בדיבור. בהערכה גסה ניתן לומר כי ZCR נמוך מצביע על קטע קולי, וZCR גבוה על קטע לא קולי.[3]



ניסינו לפתור את הבעיה בשתי דרכים. קיימת דרך המבוססת על מאמר [3], וניסינו גם להתבסס על מאמר [2]. בסוף ראינו שהאלגוריתם שהציעו במאמר [3] נתן לנו רמת דיוק הרבה יותר גבוה. השארנו בקוד את שתי הדרכים, אך את הדרך מתבסת על מאמר [2]. השארנו את הקוד בהערה רק בשביל שתוכלי לראות את הכיוון.

3. תיאור הנתונים שנאספו או תאור הנתונים עליהם עבדתם

ס"כ השתמשנו בתשע דגימות קוליות:

א) "shee_mono.wav"
ב) "bad.wav"
ג) "bed.wav"
ד) "come_here_ususal.wav"
ה) "ImTooOldForThis.wav"
ו) "MyNameIsBojan.wav"
ז) "SheHad.wav"
ח) "OhMySon.wav"

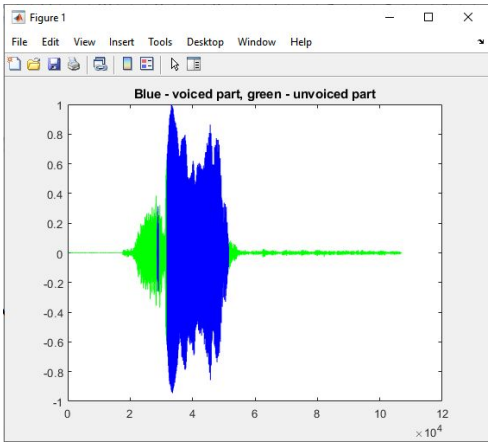
ט) "SheHas_me.wav"
הקלטות א', ב', ג', ז', ט' נלקחו ממטלות בית.
הקלטות ה', ח' לקוחות מתוך הקלטות של משחק.
הקלטה ו' נלקחה בשביל לראות תקינות עבודה של אלגוריתם עבור הקלטה ארוכה.

4. תיאור הflow של העבודה כולל מימוש קוד

ניתן לחלק את כל התהליך עבודה לשלבים הבאים
א) חיפוש וקריאה מאמרים בנושא של בעיה
ב) בחירה של מאמרים שעליהם העבודה שלנו תתבסס
ג) בחירה של הקלטות שעבורן נבדוק תקינות אלגוריתם
ג) חלוקת משימות תכנות בין חברי צוות
ד) חיבור קוד
ה) בדיקת תוצאות

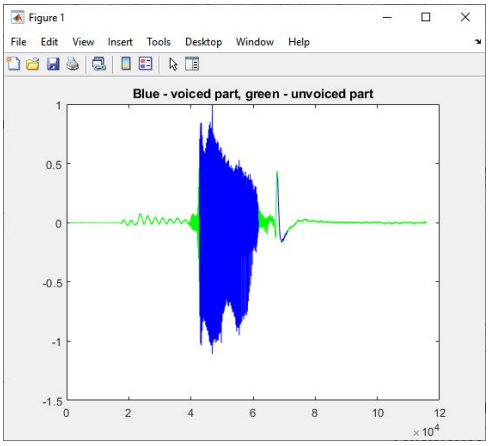
5. הנסיונות עצמם - מה נעשה, מה הצליח ומה לא, תוצאות וניתוח תוצאות

אלגוריתם שלנו בונה דיאגרמה שמסמנת voiced part בצבע כחול unvoiced part בצבע ירוק
תוצאה עבור 'shee_mono.wav':



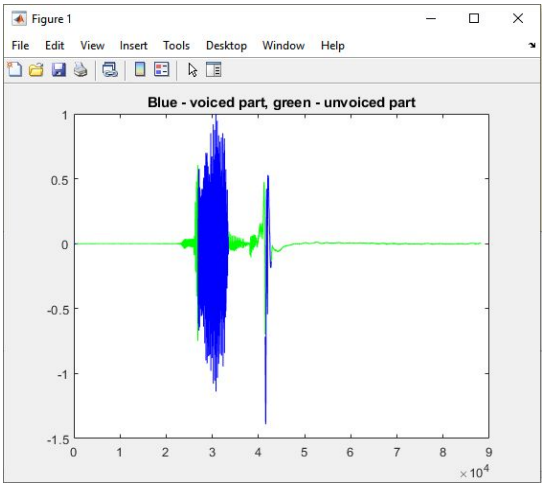
כאן אלגוריתם היה אמור לסמן את 'sh' ואת קטע שקט בסוף בירוק כי הוא unvoiced. ניתן לראות שזה אכן קרה חוץ מקטע ממש קצר לקראת סוף של "sh".

תוצאה עבור 'bad.wav':

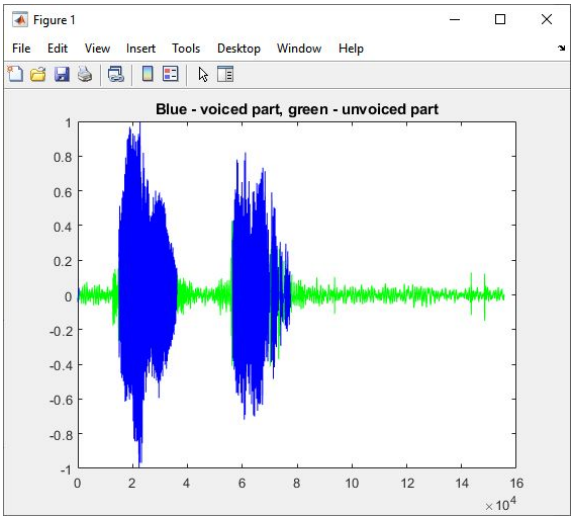


בעקבות זאת שגם 'b' גם 'a' וגם 'd' הם voiced, אלגוריתם היה צריך לצבוע הכל חוץ משקט בכחול וזה אכן קורה.

תוצאות עבור 'bed.wav':

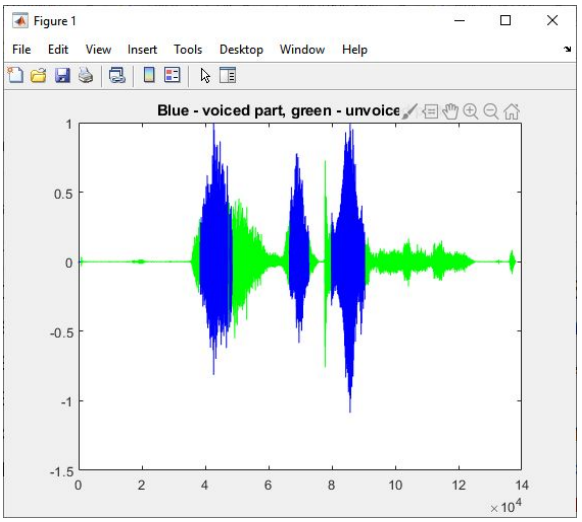


תוצאות עבור 'come_here_usual.wav':



ניתן לראות ש'h', 'k' ומקטעים עם silence מסומנים unvoiced

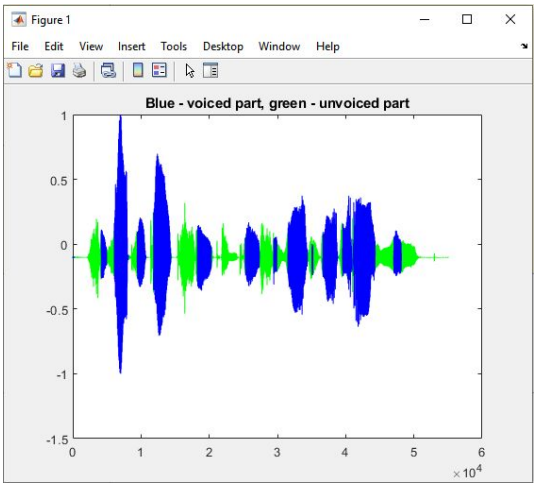
תוצאות של 'Ah, I'm too old for this':



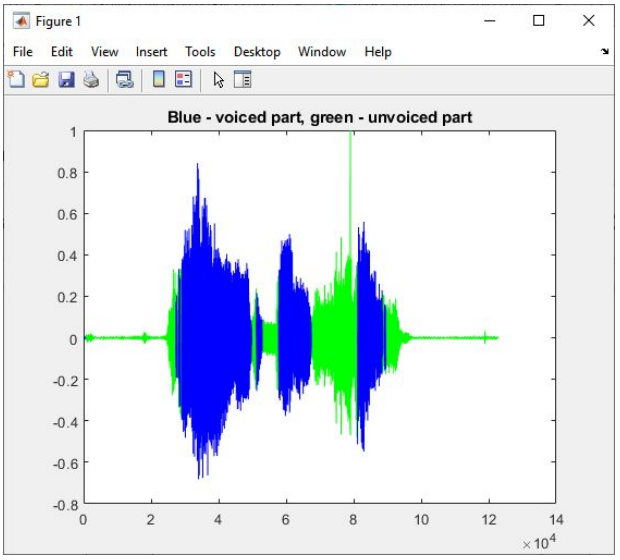
מ 3 עד 6 Ah

מסומן חצי כ VOICED וחצי שני UNVOICED
אחר כך I'm מסומן כ VOICED
אחר כך 't' סומן כ unvoiced ו'u' סומן כ voiced. אחרי זה 'old' סומן כ voiced.
אחר כך 'for this' סומן כ unvoiced כולו לעומת זאת ש'l', 'd', 'z' הן voiced. יכול להיות שזה קשור לדיבור שקט (אנרגיה נמוכה) במקטע הזה.

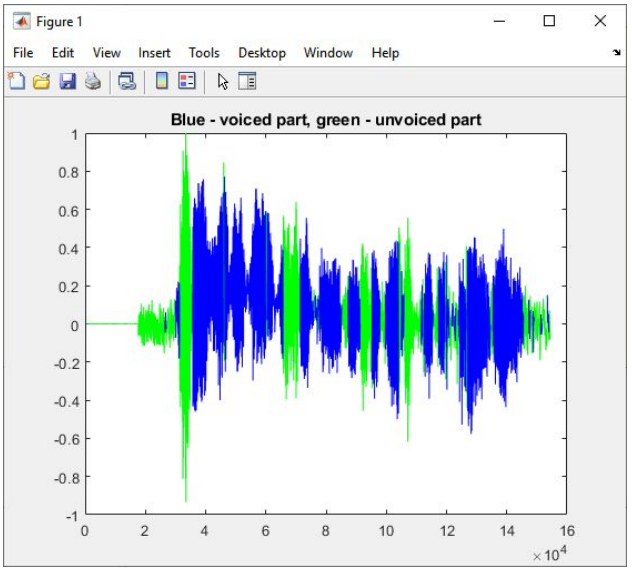
תוצאות של "SheHad.wav:



תוצאות עבור 'OhMySon.wav':



תוצאות עבור 'SheHas_me.wav':



מסקנה עבור התוצאות:
רואים רמת הדיוק אכן די גבוה אך לפעמים יש מקטעים קצרים שאלגוריתם מסמן voiced/unvoiced בצורה לא נכונה. אם בהקלטה בן אדם מדבר פחות ברור (מדלג על אותיות, משנה פיצ' בצורה משמעותית וכו') זה משפיע באופן שלילי על התוצאות.

6. מה אפשר עוד לבדוק ולא בדקתם
לא הצלחנו לבנות אלגוריתם מתבסס על [3] כך שזה יתן לנו דיוק טוב. יכול להיות שאם היינו מצליחים אז היה אפשר לשלב [3] ו[2] וזה היה נותן לנו רמת דיוק יותר גבוה.
יכול להיות שthresholds שלקחנו עבור ZCR וSTE אינם אופטימליים וthresholds יותר טובים היו נותנים תוצאות יותר מדויקות.

7. מה למדת מהעבודה
כל העבודה הזאת היא תזכור טוב של כמה העולם של speech recognition הינו גדול ואנחנו רק למדנו קטע מאוד קטן ממנו.
למדנו עצמאית על STE ו ZCR, ועל הקורלציה שביניהם שזה נתן לנו כלי חדש לניתוח ה DATA, ולהבנה שלו.
שמחנו מאוד לראות, או ליתר דיוק, לשמוע את התוצאה עובדת כאשר השמענו רק את הקטע הקולי או הלא קולי.

ספרות:

[1] Jong Kwan Lee, Chang D. Yoo, “Wavelet speech enhancement based on voiced/unvoiced decision”, Korea Advanced Institute of Science and Technology The 32nd International Congress and Exposition on Noise Control Engineering, Jeju International Convention Center, Seogwipo, Korea, August 25-28, 2003.
[2] A Method for Voiced/Unvoiced Classification of Noisy Speech by Analyzing Time-Domain Features of Spectrogram Image Kazi Mahmudul Hassan, Ekramul Hamid, Khademul Islam Molla
[3] Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal Bachu R.G., Kopparthi S., Adapa B., Barkana B.D. Electrical Engineering Department School of Engineering, University of Bridgeport

דברים נוספים שהשתמשנו בהם: