המכללה האקדמית להנדסה סמי שמעון

זיהוי דובר לא תלוי טקסט מקבוצת דוברים בסביבה רועשת

Text-Independent Speaker Identification in Noisy Environment

פרויקט הנדסי

דו"ח מסכם

הוכן לשם השלמת הדרישות לקבלת

תואר ראשון בהנדסה B. Sc

מאת

ניק אלסקובסקי

גבריאל בנגייב

בהנחיית מר ולדימיר וולפין

הוגש למחלקה להנדסת חשמל ואלקטרוניקה

המכללה האקדמית להנדסה סמי שמעון

אשדוד

תשע"ד **יוני 2014**

המכללה האקדמית להנדסה סמי שמעון

זיהוי דובר לא תלוי טקסט מקבוצת דוברים בסביבה רועשת

Text-Independent Speaker Identification in Noisy Environment

פרויקט הנדסי

דו"ח מסכם

הוכן לשם השלמת הדרישות לקבלת

תואר ראשון בהנדסה B. Sc

מאת

ניק אלסקובסקי

גבריאל בנגייב

בהנחיית מר ולדימיר וולפין

הוגש למחלקה להנדסת חשמל ואלקטרוניקה

המכללה האקדמית להנדסה סמי שמעון

אשדוד

חתימת הסטודנט: גבריאל בנגייב	**תאריך: 30.05.2014**
חתימת הסטודנט: ניק אלסקובסקי	**תאריך: 30.05.2014**
חתימת המנחה: _____	**תאריך: 30.05.2014**
אישור ועדת הפרויקטים: _____	**תאריך: _____**

תחום הפרויקט: עיבוד אותות (Signal Processing)

סוג הפרויקט: תכנון

Key words: GMM ,EM, Spectral Subtraction, MFCC

תקציר

במסגרת הפרויקט נעסוק באלגוריתם Gausssian Mixture Models (GMM) למטרת זיהוי הדובר מכלל האוכלוסייה, ללא תלות בטקסט הנאמר על ידי הדובר. בשימוש מודל זה מאפיינים את זהותו של הדובר ע״י סדרה של מרכיבים גאוסיאניים במשקלים שונים המתאימים למאפיינים האקוסטיים של אות הדיבור ולפיכך נוצר מודל ייחודי עבור כל דובר ודובר במאגר הדוברים.

רעשי רקע אקוסטיים המתווספים לאות הדיבור גורמים לירידה ברמת הביצועים של מערכות עיבוד קול דיגיטאליות המיועדים ליישומי דחיסה, זיהוי ואימות של אות הדיבור.

במסגרת הפרויקט נממש ונחקור את אלגוריתם ה- GMM באמצעות תוכנת MATLAB בתנאי סביבה שונים כגון: רעשים מתווספים, רעשים בהתפלגויות שונות, ערכי יחס אות לרעש משתנים והשפעתם על אחוזי ההצלחה של גילוי הדובר מבין מאגר הדוברים הידוע למערכת. כמו כן נעסוק בשיטת השבחה של אות דיבור על ידי הפחתה ספקטראלית של הרעש האקוסטי המתווסף אליו.

לבסוף נערוך אומדן עבור רמת הביצועים והיעילות של אלגוריתם ה-GMM לאחר ההשבחה ולפניה, כלומר עד כמה, אם בכלל משתפרים אחוזי ההצלחה של המודל.

Abstract

In the course of this project the Gaussian Mixture Models (GMM) will be used for robust text independent speaker identification. By using this model each speaker will be represented by various set of weighted Gaussian components that corresponds to the acoustic characteristics of the speech that will be unique for each speaker from the known speakers data base.

Background noise acoustically added to the speech can degrade the performance of digital voice processors used for applications such as speech compression, recognition and authentication.

by using MATLAB software we will simulate and study the effects of a noisy environment on the success rate of the speaker identification using the GMM algorithm as well as the success rate of the speaker identification after suppressing the acoustic noise using the spectral subtraction method.

Finally, we will evaluate the performance and efficiency of the GMM algorithm before and after using the acoustic noise suppression.

המכללה האקדמית להנדסה סמי שמעון

תוכן עניינים

רשימת איורים

רשימת טבלאות

1 מבוא

מערכות לזיהוי דובר, המכונות לעיתים מערכות ביומטריה של הדובר, כוללות זיהוי, אימות, סיווג, ואף עקיבה אחר הדובר. מונח כללי מעין זה מגדיר כל תהליך בו מעורבת ידיעת זהות הדובר על סמך קולו. ראוי לציין שלמונח זיהוי דובר מספר מונחים נוספים, דבר שיצר בלבול רב. המחקר העוסק במערכות לזיהוי דובר אוטומטי מזה כבר שנים יצר בלבול בקרב הציבור בין מערכות לזיהוי דיבור למערכות לזיהוי דובר, מונח נוסף שתרם לבלבול הינו זיהוי קול, מונח אשר משתמשים בו במספר חוגים בתור תחליף לזיהוי דובר. ביישומים לזיהוי דיבור, זה לא קולו של הדובר אשר מזוהה, אלא התוכן של הדיבור, לעומת זאת במערכות לזיהוי דובר, זהו קולו של הדובר אשר מזוהה.

השלב הראשון בתהליך זיהוי הדובר הינו חילוץ מאפייני מערכת הקול של הדובר. מאפיינים אלו יכולים להיות מיוצגים על ידי מודל מתמטי המתאר את המערכת הפיזיולוגית אשר מפיקה את אות הדיבור, או באמצעות מודל סטטיסטי אשר מתאר בקירוב את אותם המאפיינים. לאחר קביעת המודל, אשר יהיה ייחודי לדובר, בהינתן אות דיבור נוסף, יהיה ניתן ליחסו על ידי ערך סבירות מסוים לדובר ביחס ליתר המודלים הנצפים של יתר הדוברים. זוהי השיטה היסודית בבסיסה של כל מערכת לזיהוי דובר.

בכל הנוגע לחשיבות מערכות אלו, ראוי לציין שזהות הדובר הינו הביומטריה שקיימת בשימוש נרחב בתשתיות מגוונות, מבין הבולטות שבהן היא הרשת הטלפונית. דבר זה מקנה חשיבות רבה למערכות אלו במגוון רחב של יישומים בכל העולם. ראוי לציין שבעקבות מגמת עליה גבוהה של השימוש בטלפונים ניידים, ומורכבותם ההולכת וגדלה, מסיבה זו מערכות לזיהוי דובר ימשיכו להיות פופולאריות גם בעתיד.

למערכות זיהוי דובר ישנם מספר רבדים אשר קשורים ביניהם באופן ישיר או עקיף. באופן כללי, קיימים שני קבוצות עיקריות של רבדים, הראשונה הינה הקבוצה "הפשוטה" והשניה הינה "המורכבת". הקבוצה הפשוטה כוללת בתוכה את הרבדים הבאים: זיהוי, אימות וסיווג הדובר. הקבוצה המורכבת כוללת בתוכה את הרבדים של חלוקת הדוברים לקבוצות, דהיינו סגמנטציה, וגילוי ועקיבה של הדובר. בהווה, מערכות לאימות הדובר הינו הרובד הפופולארי ביותר מבין יתר הרבדים כיוון שזו נחוצה לצורכי אבטחה, כמו כן מערכות אלו פשוטות יותר ליישום בניגוד למערכות לזיהוי דובר.

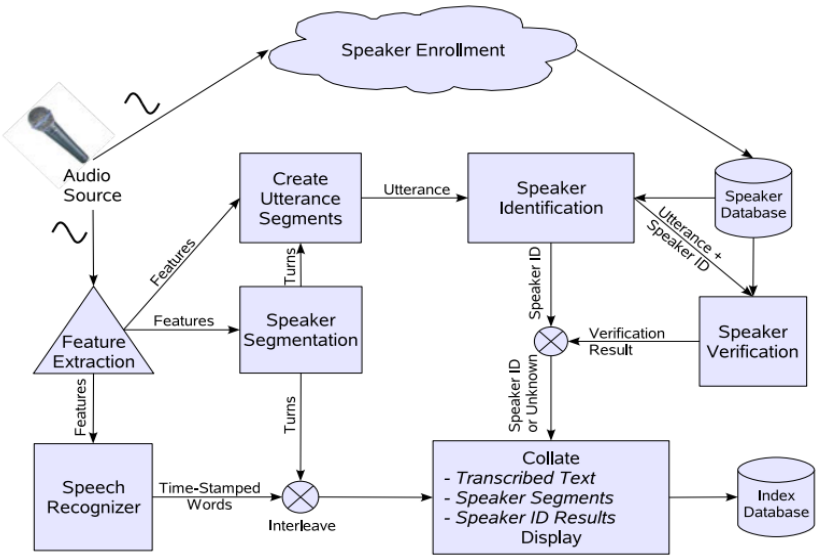
ביישומים של אימות הדובר, תחילה, הדובר יעבור תהליך של אימות באמצעים אשר אינם קוליים, דהיינו הקלדת שם משתמש, מספר מזהה וכו', בשלב זה מחלצים מהמאגר המידע את המודל הסטטיסטי הייחודי עבור אותו הדובר על סמך הפרטים המזהים שהקליד קודם לכן. בשלב הבא נערכת השוואה בין קולו של הדובר לבין הקול שנשמר במאגר.

ישנם שני גישות למערכות לזיהוי דובר, מאגר "סגור" ומאגר "פתוח" של דוברים:

- מאגר "סגור" –בשיטה זו אשר פחות מורכבת מבין השניים, מאפייני האות הקולי של הדובר הנבחן משווה ביחס ליתר הדוברים שקיימים במאגר כאשר בסוף התהליך, מוחזר שמו של הדובר המשוערך בסבירות המרבית. ראוי לציין שבגישה זו תמיד תתקבל הערכה לאחד הדוברים הקיימים במאגר, דהיינו המערכת לא מסוגלת לזהות מתחזים, או דוברים שכלל לא קיימים במערכת.
- מאגר "פתוח" – שיטה זו כוללת בתוכה את רובד הזיהוי והאימות, בשיטה זו, להבדיל מהשיטה הקודמת, הדובר הנבחן, במידה ולא יזוהה עם יתר הדוברים הקיימים במאגר, לא יזוהה על ידי המערכת. דבר נוסף שראוי לציין הינה המורכבות וסיבוכיות החישוב שתגדל באופן לינארי ככל שירבו מספר הדוברים במאגר, כיוון שהמערכת תיאלץ לערוך השוואה והבחנה בין כלל הדוברים הידועים למערכת.

נציין מספר שיטות לזיהוי דובר:

- מערכת לזיהוי דובר תלוי טקסט – שיטה זו כשמה כן היא, מתבססת על היכולת לזהות דובר באמירתו טקסט מוגדר מראש. ראוי לציין ששיטה זו שייכת לרובד האימות. שיטה זו מעניקה דיוק גבוה.
- מערכת לזיהוי דובר לא תלוי טקסט – שיטה זו הינה המגוונת מבין כל השיטות הקיימות. בשיטה זו ניתן להשתמש במגוון הרבדים שצוינו לעיל. שיטה זו, אשר איננה תלויה טקסט, או בשפה המדוברת, מסתמכת רק על המאפיינים של מערכת הקול עבור הדובר המיוחס וכלל לא מעניקה משמעות על תוכן הדיבור.
- מערכת לזיהוי דובר באמצעות טקסט מתוזמן – בשיטה זו, על הדובר לומר טקסט ספציפי במסגרת זמן מוגדרת מראש. שיטה זו פותחה בעיקר על מנת לזהות מתחזים, מכיוון שאמירת הטקסט הינה מתוזמנת, קשה יהיה מאוד להכשיל מערכת זו.



איור 1.1: דיאגרמת מערכת זיהוי דיבור הכוללת מספר רבדים [9].

1.1 יישומים

באופן מעשי, אין מגבלה על כמות היישומים האפשריים בתחום מערכות זיהוי הדובר [8],[7],[1]. בהינתן אות דיבור, הרי שניתן יהיה לנצל את אחד הרבדים של מערכות אלו. אף על פי כך, במונחים של יישום ושימוש, מערכות אלו נמצאות בפיתוח מתמיד. להלן רשימה מקוצרת של מספר יישומים נפוצים:

- יישומים פיננסיים – בהווה, חלק ניכר מהבנקים, מעניקים גישה מלאה ואוטמטית לחשבונות הפרטיים של הלקוחות באמצעות הטלפון.
- יישומי אבטחה וגישה – יישום זה נמצא שימוש במגזר הצבאי והאזרחי/פרטי, באמצעותו ניתנת גישה לסביבת מורשים מוגדרת מראש.
- יישומים משפטיים ומשטרתיים – הדיבור, הינו ייחודי מבחינת האופי הלא פולשני שלו, דהיינו הדיבור יכול להיאסף ללא מודעות הדובר, דבר הגורם לו להיות מועמד ראוי ליישומים משפטיים בהתמודדות עם משתמשים או עבריינים שאינם משתפים פעלולה.

1.2 רעש ועיוותים

רעשי רקע אקוסטיים המתווספים לאות הדיבור גורמים לירידה ברמת הביצועים של מערכות עיבוד קול דיגיטאליות המיועדים ליישומי דחיסה, זיהוי ואימות של אות הדיבור, לפיכך נדרשות מערכות או אלגוריתמים שונים של עיבוד אות לצורך הפחתה או סינון אותם הרעשים. רעש יכול להיות מוגדר כאות אקראי לא רצוי אשר עלול להפריע בתהליך התקשורת או בתהליך מדידת האות. רעש כשלעצמו נושא מידע אודות המקור הגורם להיווצרותו ביחס לסביבה בה הוא מתפשט. בתור דוגמא הרעש שנפלט ממנוע המכונית מכיל מידע אודות מצבו של המנוע ועד כמה טוב המנוע עובד. דוגמא נוספת הינה הקרינה הקוסמית אשר מכילה מידע אודות היווצרות ומבנה היקום. כמו כן אילו נתייחס למקרה בו מתנהלת שיחה בין דוברים בסביבה בה מתנהלות שיחות ברקע, הרי ששיחות הרקע עלולות להקשות על האזנה לשיחתם של זוג הדוברים.

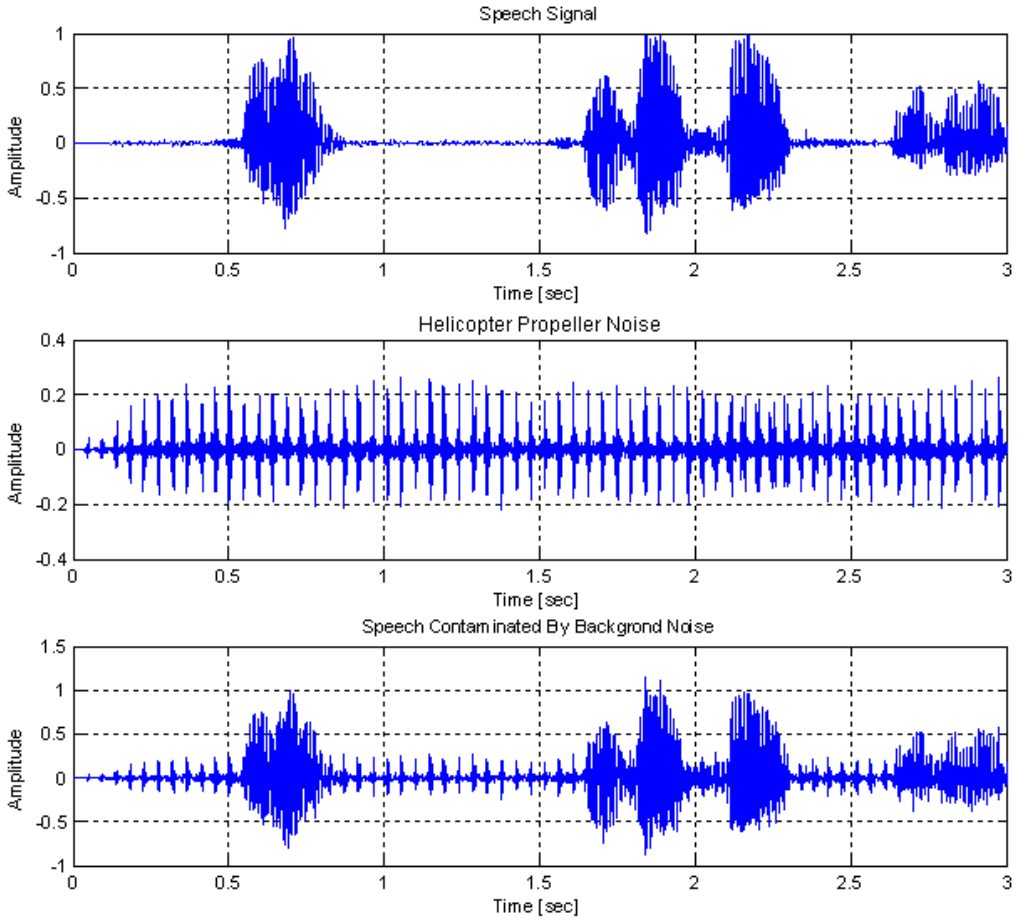
נוכחות הרעש, על מידותיו השונות, קיימת בכל סביבה שהיא, בתחום הסלולר למשל, יכול להיות מספר מגוון של רעשים אשר עלולים להנחית את איכות התקשורת, רעשים אלו כוללים רעשי רקע אקוסטיים, רעשי המכשיר האלקטרוני, דהיינו רעשים טרמיים ורעשי שוט, רעשים אלקטרומגנטיים בתחום תדרי הרדיו ורעשים הנוצרים כתוצאה מתהליך עיבוד האות.

הרעש עלול לגרום לשגיאות בזמן שידור המידע ואף יכול לשבש את תהליך התקשורת כולו, לפיכך עיבוד הרעשים הללו הינו חלק אינטגרלי וחשוב במשמעותו עבור תהליכי השידור והקליטה כמו גם בתהליך עיבוד האותות בתקשורת המודרנית. רמת ההצלחה של שיטת עיבוד הרעש תלויה למעשה ביכולת לאפיין ולמדל את תהליך היווצרותו, מאפיינים אלו ניתן יהיה לנצל על מנת להבחין ולהבדיל בין הרעש לבין מקור המידע הרצוי.

רעש ניתן לתאר בזכות המקור אשר יוצר אותו ואודות התהליך הפיזיקלי שגורם להיווצרותו. סוגי הרעשים כוללים: רעשים אקוסטיים, אלקטרוניים, אלקטרומגנטיים ואלקטרוסטטיים. זאת ועוד, בתקשורת הדיגיטאלית קיימים עיוותים הנגרמים על ידי הערוץ, כמו למשל רעשי כימות.

הפרעות אקוסטיות כוללות:

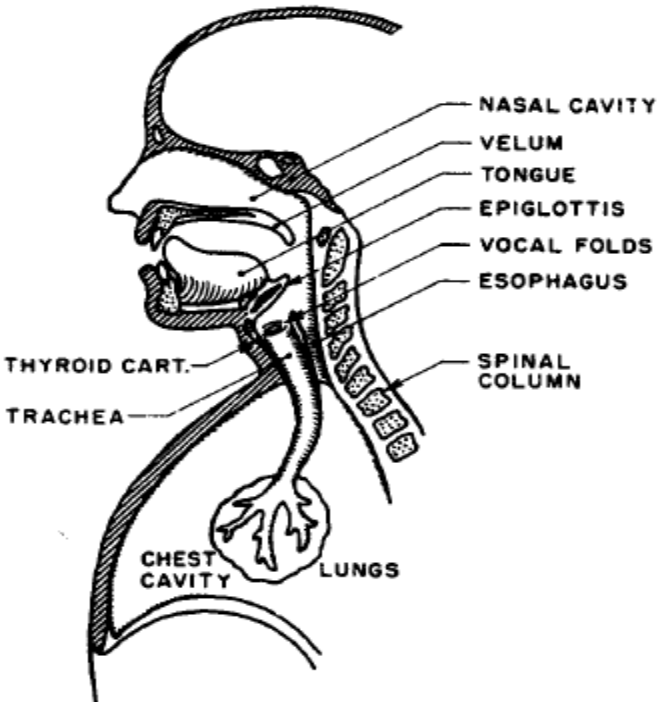
- רעשים אקוסטיים הנובעים מעצמים נעים, רוטטים או מתנגשים, אלו למעשה מקורות הרעש הידועים ביותר בסביבה היום יומית. דוגמאות למקורות רעש אלו הם מכוניות נעות, הקלדה על מקלדת, מערכת מיזוג אוויר, גשם, רוח וכדומה.
- משוב אקוסטי והד: אלו נובעים כתוצאה מהחזרים של הצלילים מקירות החדר למשל, או כתוצאה מצימוד בין המיקרופון לבין הדובר. להרחבה ראה נספח A.



איור 1.2: גרפים המתארים אות דיבור, רעש המופק ממנוע של מסוק והתוצאה שמתקבלת מסכומם.

1.3 מבנה אות דיבור ותכונותיו

על מנת לעסוק באנליזה של אות הדיבור, תחילה עלינו להבין כיצד הוא נוצר ומהם תכונותיו. לצורך המחשת תהליך יצירת הגלים האקוסטיים הנושאים את אות הדיבור נתבונן באיור הבא:



איור 1.3 מערכת הקול האנושית [1].

תחילת דרכו של יצירת אות הדיבור מתחיל כתוצאה ממעבר האוויר אשר נדחף מהריאות (Lungs) שניתן לתאר אותה כחלק ממערכת לחץ האוויר (Air pressure system), אשר עובר דרך קנה הנשימה (Trachea), ומשם ממשיך למיתרי הקול (Vocal folds) הנמצאים בבית הקול (Epiglottis) שאלו רקמות המוצמדות זו לזו, אשר נפתחים או נסגרים לסירוגין וכך גורמות להיווצרות פולסי אוויר מחזוריים, כאשר המרווח המחזורי בין פולסי האוויר הנוצרים נקרא בשם מחזור ה-"Pitch" שהיחס ההפוך אליו הוא התדר היסודי. מחזור זה מושפע משינויים בלחץ האוויר המתקבל מהריאות ובמתיחות של מיתרי הקול. קנה הנשימה ומיתרי הקול משמשים בתור מערכת האחראית על הרטיטה (Vibratory system), משם, זרימת האוויר עוברת דרך החך הרך (Soft palate) שתפקידו לשלוט על מעברו ממיתרי הקול אל תוך חלל הפה (Oral cavity) על פני הלשון (Tongue) השיניים והשפתיים (Lips) ואל חלל האף (Nasal cavity).

העירור עבור מערכת הקול נוצר על ידי זרימת האוויר כפי שנאמר, המגיע אל חלל הפה והאף, החך הרך, השפתיים והלשון יוצרים בתורם מעצורים שונים ואלו גורמים למערבולות בזרמי האוויר עבור העירור המקורי ועל ידי כך נוצר חיתוך הדיבור (Articulation), הייחודי עבור כל אדם. ניתן לייחס לחלל הפה והאף אנלוגיה של פונקציית תמסורת אשר משתנה בזמן הנגרמת כתוצאה משינוי צורת חלל הפה והאף בזמן הדיבור. פונקציית תמסורת זו יוצרת אפנון עבור פולסי האוויר וכך מתקבל אות הדיבור [8].

את אות הדיבור ניתן לסווג לשני סוגי צלילים עיקריים, הראשון מביניהם הינו צליל קולי (Voice) כאשר הגורם ליצירת הצליל הקולי בעל אופי מחזורי, כלומר צלילים אלו ניתן לאפיין על ידי תדר Pitch, אשר משתנה מאדם לאדם, לפי מינו וגילו, ונע בתחום התדרים של 50 Hz -400 Hz הינו מידת הרטיטה של מיתרי הקול.

סוג הצליל השני הינו צליל א-קולי (Unvoiced) בו, להבדיל מהצלילים הקוליים, מיתרי הקול אינם רוטטים ולפיכך האות המתקבל עבור צלילים אלו הוא בעל אופי רועש, ראוי לציין שאנרגיית הצלילים הקוליים מכילים כמות גדולה יותר של אנרגיה ביחס לצלילים א-קוליים, נוסף על כך קיימים צלילים המשלבים את המאפיינים של הצליל הקולי והא-קולי.

על פי האנלוגיה שהצגנו, ובה חלל הפה והאף מייצגים פונקציית תמסורת אשר משתנה בזמן, מתקבלת במישור התדר מעטפת ספקטראלית של אות הדיבור המאופיין במספר שיאים הנקראים בשם "פורמנטים" (Formants). שיאים אלו הם תדרי התהודה של הצלילים הקוליים, במערכת קולית ממוצעת יתקיימו לכל היותר ארבע פורמנטים בתחום תדרי הדיבור, ואלו נעים בין ערכים של 50 Hz ועד 4 KHz, נוסיף ונאמר שחלק ניכר מהאנרגיה של אות הדיבור מתרכזת בתחום שבין 300 Hz ועד 3 KHz, כמו כן נציין שלמטרות זיהוי ומובנות של אותות דיבור, נמצא כי התחום היעיל ביותר הוא תחום שבין 300 Hz עד 3.4 KHz.

ראוי לציין אחת הבעיות הטמונות בניתוח אותות דיבור היא היעדר תכונות סטטיסטיות סטציונאריות המאפיינות אות בו לא קיים שינוי של התדר עם הזמן, דהיינו פרמטר התוחלת והשונות של האות לא משתנים עם הזמן, ולכן לא ניתן לערוך לאות המשך עיבוד באנליזה של פורייה, אות הדיבור כשלעצמו נראה כאות אקראי. בעיה נוספת שמתעוררת הינה מציאת כלי יעיל לאפיון תכונות אות הדיבור, אשר תהיה ייחודית עבור כל דובר לצורך זיהוי הדובר והגדרת מודל הסתברותי אשר יהיה מסוגל להתאים עבור כל דובר ממאגר הדוברים מודל ייחודי על סמך מאפייניו האקוסטיים ועל פי מודל זה לזהות בסבירות המרבית מי מהדוברים מהמאגר דיבר.

2 סקר ספרות

2.1 עיבוד מקדים והשבחת אות הדיבור

2.1.1 עיבוד מקדים

תהליך העיבוד המקדים של אות הדיבור נחשב בעל חשיבות חיונית ביותר לצורך קבלת מערכת זיהוי דובר חסונה ויעילה. תהליך זה מכיל את השלבים הבאים:

- השלב הראשון בתהליך העיבוד המקדים הינו דגימת האות האנלוגי הרציף, דהיינו אות הדיבור המוקלט במיקרופון מסביבת הדובר והמרת אות זה לאות דיגיטאלי להמשך עיבוד ספרתי. ראוי לציין שאמנם קיימות מספר דרגות חופש לקביעת התדר בו יידגם האות המוקלט אך המטרה תמיד תהיה להימנע מתופעת ה-"Aliasing" [2] שעלולה לגרום לחפיפה בין המרכיבים הספקטראליים הדגומים במישור התדר ועל ידי כך ליצור עיוותים לאות המקורי, כך שאות הדיבור יתקבל בתדרים שונים מהתדרים המקוריים. כדי להימנע מתופעה זו יש להקפיד לדגום בתדר הגבוה בלפחות פי 2 מהתדר המקסימאלי של האות הנדגם, דהיינו תדר ניקוויסט, התדר מינימאלי לדגימת אות דיבור הינו 8 KHz.
- השלב השני בתהליך זה הינו הורדת מרכיב ה-DC. הסיבה לכך היא שמערכות ההקלטה השונות, לרבות מיקרופונים על תצורותיהם השונות, כוללים מרכיב DC הגורם לשיא גבוה במיוחד במרכיב הספקטראלי הראשון שכלל לא מכיל מידע אודות אות הדיבור, הורדת רכיב זה לפיכך מעניקה הבלטה משמעותית יותר עבור המרכיבים הספקטראליים הנחוצים לצורך המשך עיבוד האות, לצורך כך קיימות מספר שיטות, הראשונה מבניהם תהיה חיסור הממוצע מהאות, שיטה נוספת לצורך הורדת רכיב ה-DC הינה התמרת פורייה, איפוס המרכיב הספקטראלי הראשון ולבסוף התמרת פורייה הפוכה.
- השלב השלישי בתהליך העיבוד המקדים הוא מעבר האות דרך "Pre-Emphasis". בעת מעבר אות הדיבור מהמהוד דרך השפתיים אל התווך ניתנת הדגשה במונחי אנרגיה למרכיבי התדרים הנמוכים יותר ביחס לתדרים הגבוהים המרכיבים את האות, דבר הגורם למעשה לאיבוד מידע אודות האות דיבור. מערכת השמיעה האנושית מסוגלת להבחין ולזהות מרכיבים אלו די בקלות, לפיכך לצורך תכנון מערכת אוטומטית לזיהוי דוברים יש לתכנן מערכת שתהיה מסוגלת להבחין במרכיבים אלו על מנת שיהיה ניתן לנצל את המאפיינים הספקטראליים של האות בתדרים הללו. שיטה נפוצה לצורך מתן הדגשה למרכיבים הספקטראליים הללו היא העברת אות הדיבור הדגום דרך מסנן מעביר גבוהים מסדר ראשון מסוג "Finite Impulse Response (FIR)" בעל פונקציית התמסורת הבאה:

$$H_p(z) = 1 - az^{-1} \qquad (2.1)$$

כאשר הערכים הנפוצים ביותר עבור פרמטר a נעים בתחום 0.95 ועד 0.97 [9],[1] על מנת להדגיש ולאזן את התפלגות האנרגיות של המרכיבים הספקטראליים בתדרים הגבוהים והנמוכים המוכלים באות הדיבור.

- השלב הרביעי בתהליך העיבוד המקדים כולל את שלב חלוקת אות הדיבור למקטעים חופפים. על מנת לנתח את אות הדיבור באנליזת פורייה תחילה יש לחלק את אות הדיבור למסגרות זמן של 20 עד 40 אלפיות השנייה, כאשר ההנחה היא שבעבור מסגרות זמן תחומות אלו אות הדיבור איננו משתנה משמעותית והינו בעל מאפיינים סטטיסטיים סטציונאריים, פרמטר התוחלת והשונות של מסגרות אלו נשארות קבועות. נהוג לחלק מסגרות אלו בחפיפה של 50%, דהיינו 10 עד 20 אלפיות השנייה עבור מסגרות הזמן לעיל, דבר זה מאפשר כושר הפרדה גבוה יותר כמו גם מעקב טוב יותר אחר השינויים המתרחשים באות הדיבור במישור הזמן.
- השלב החמישי והאחרון בתהליך העיבוד המקדים הוא הכפלת מקטעי הדיבור החופפים בפונקציית חלון לצורך חישוב האנרגיה של כל אחת ממקטעי אות הדיבור בשלב מאוחר יותר. מבין מגוון פונקציות החלונות הקיימות ניתן למצוא את חלון Hamming, Hanning, Welch, Triangular, Gauss, Blackman ו-Bartlett , אולם פונקציית חלון מסוג Hamming ללא ספק הנפוצה והפופולארית מבין כל יתר הפונקציות בתחום עיבוד אות הדיבור, הסיבה העיקרית לפופולאריות לה היא זוכה נובעת מעצם העובדה שזו מעניקה אונות צד נמוכות וקבועות עבר כל חלון ביחס לפונקציות חלון מתחרות [5] הרחבה בנספח B.

2.1.2 <u>קצב חציית האפס "Zero Crossing Rate (ZCR)"</u>

לצורך זיהוי והפרדה בין התחומים הקוליים והא-קוליים של אות הדיבור, קיימות מספר שיטות וגישות, וביניהן זיהוי דפוס באות הדיבור וטכניקות שבהם קיים שימוש בתכונות הסטטיסטיות או האקוסטיות של אות דיבור. שיטה אחת לסיווג אות דיבור לתחומים קוליים או אלה שאינם הינה שיטת קצב חציית האפס "Zero-crossing rate" ולרבות נמצאת בשימוש נרחב בקדמת מערכות לזיהוי קולי. ניטור כמותי של חציית האפסים מהווה אינדיקציה עבור התדר בו מרוכזת האנרגיה בספקטרום אות הדיבור, כאשר בתחום קולי באות הדיבור מתקבלת חציית אפס נמוכה אשר מאפיינת עירור הנוצר על ידי מערכת הקול וזרימה מחזורית של האוויר ממיתרי הקול.

באשר לתחומים שאינם קוליים מתקבלת חציית אפס גבוה הנגרמת כתוצאה מזרמי מערבולת של האוויר הנדחף מהראות אל מערכת הקול. בהקשר האות בזמן הבדיד, חציית האפס מתרחשות במידה ודגימות האות מקבלות סימנים אלגבריים מנוגדים. קצב חציית האפס הינו מדד למספר הפעמים, באינטרוול זמן נתון, בה אמפליטודת אות הדיבור חוצה את ערך האפס. אות הדיבור הינו אות רחב פס ולכן פענוח של ממוצע קצב חציית האפס לא בהכרח מדויקת, למרות זאת שיערוך גס של מאפיינים ספקטראליים יכולים להתקבל על ידי ייצוג המתבסס על אינטרוול זמן קצר עבור ממוצע קצב חציית האפס.

הגדרה מתמטית עבור קצב חציית האפס נתונה באופן הבא:

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m) \qquad (2.2)$$

כאשר:

$$sgn[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases} \qquad w(n) = \begin{cases} \frac{1}{2N}, 0 \leq n \leq N-1 \\ 0, \text{אחרת} \end{cases}$$

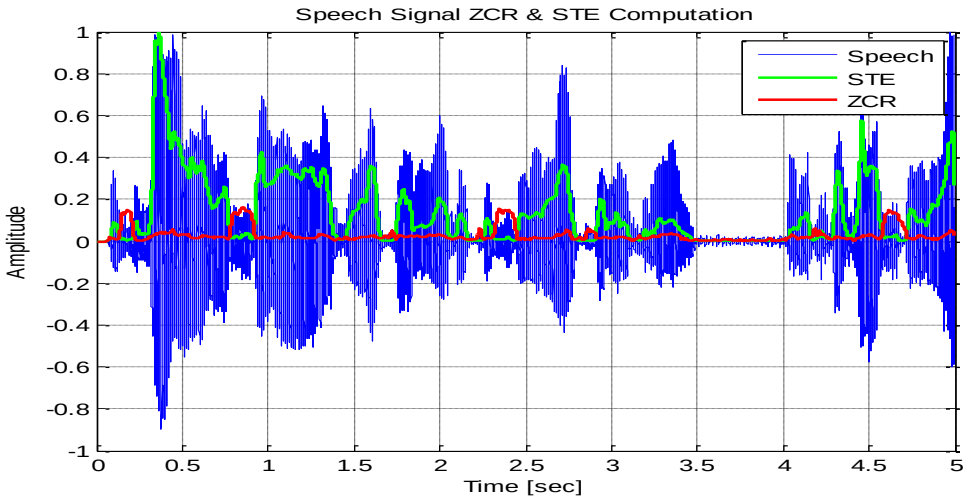
היות ותדרים גבוהים מרמזים על קצב חציית אפס גבוה יותר ואילו תדרים נמוכים מרמזים על קצב חציית אפס נמוך יותר, הרי שמתקיימת קורלציה בין קצב חציית האפס לבין התפלגות גודל האנרגיה לפרק זמן קצר. ההנחה היא שבעבור תחום אות קולי יתקבל קצב חציית אפס נמוך ובעבור תחום אותות שאינם קוליים יתקבל קצב חציית אפס נמוך יותר.

2.1.3 אנרגיית האות לפרק זמן קצר "Short Time Energy (STE)"

שיטה נוספת לסיווג אות הדיבור מתבססת על פרמטר האנרגיה הטמונה באות הדיבור, כאשר התחום באות דיבור יהיה בעל ערך אנרגיה גבוה בשל מחזוריותו ביחס לתחום אות שאיננו קולי בו קיימת אנרגיה נמוכה. אמפליטודת אות הדיבור משתנה עם הזמן, באופן כללי, אמפליטודת האות הדגום בתחום שאיננו קולי נמוכה יותר מאמפליטודת האות הדגום אשר בתחום הקולי. אנרגיה של אות דיבור מספקת ייצוג אשר משקף את שינויי האמפליטודה. ההגדרה המתמטית של אנרגיה לזמן קצר מוגדרת באופן הבא:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \qquad (2.3)$$

בשיטה זו האות הדגום מוכפל בפונקציית חלון אשר תקבע את אופי ייצוג אנרגיית האות לפרק זמן קצר.



איור 2.1: קצב חציית האפסים ואנרגיית האות לפרק זמן קצר של אות דיבור.

2.1.4 השבחת אות הדיבור על ידי שיטת ההפחתה הספקטראלית

ההפחתה הספקטראלית הינה שיטה שמטרתה היא שיחזור של ספקטרום ההספק או לחילופין ספקטרום עוצמת האות הנצפה ברעשים מתווספים, באמצעות הפחתה של שיערוך ממוצע ספקטרום הרעש מתוך הספקטרום של האות הרועש. ספקטרום הרעש משוערך ומעודכן, מתוך המחזורים בהם אות הדיבור איננו קיים, כלומר שקיימת נוכחות של רעש בלבד. ההנחה היא שהרעש הוא תהליך בעל תכונות סטציונריות ומשתנה באופן איטי, כמו כן ספקטרום זה לא משתנה באופן משמעותי בין המקטעים של האות הדגום. לצורך שיחזור האות במישור זמן, ספקטרום ההספק הרגעי המשוערך משולב עם הפאזה של האות הרועש ולאחר מכן מומר באמצעות התמרת פורייה הפוכה חזרה למישור הזמן. ראוי לציין שבמונחים של סיבוכיות חישובית שיטה זו במידת מה איננה יקרה, למרות זאת כתוצאה משינויים אקראיים של הרעש, שיטה זו עלולה ליצור שיערוך שלילי של עוצמת האות או הספק האות בזמן קצר. העוצמה וההספק של ספקטרום האות הרועש הם פרמטרים אי שליליים, ובמידה ושיערוך פרמטרים אלו תניב תוצאה שלילית אלו ימופו בתור פרמטרים אי שליליים. פעולת תיקון זו עלולה לעוות את האות המשוחזר, עיוותים אלו באים לידי ביטוי באופן יותר משמעותי ככל שיחס האות לרעש קטן.

אות הדיבור הרועש במישור הזמן מוגדר באופן הבא:

$$y(n) = s(n) + d(n) \tag{2.4}$$

כאשר $s(n)$ הינו אמפליטודת האות, ו-$d(n)$ אמפליטודת הרעש האקוסטי המתווסף.

לפיכך צפיפות ההספק הספקטראלית עבור האות הרועש תוגדר באופן הבא:

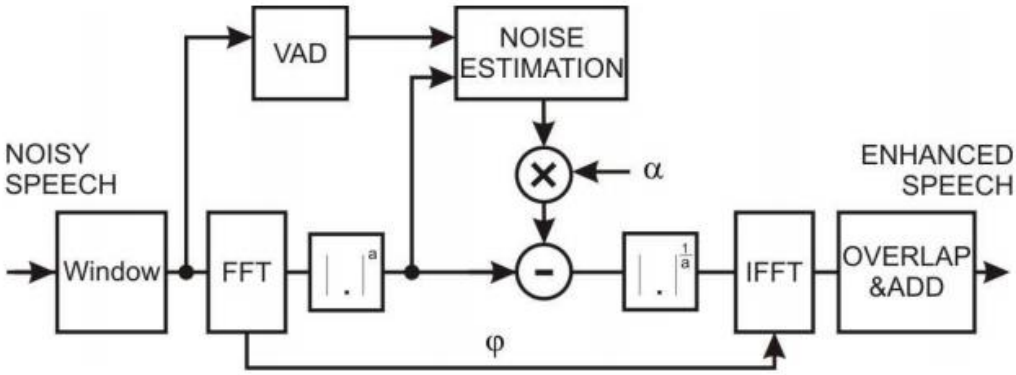
$$P_y(\omega) = P_s(\omega) + P_d(\omega) \tag{2.5}$$

$$|\hat{S}_w(\omega)|^\alpha = |Y_w(\omega)|^\alpha - E[|D_w(\omega)|^\alpha] \tag{2.6}$$

כאשר $E[|D_w(\omega)|^\alpha]$, עבור כל מקטע w, יתקבל על סמך ההנחה שתכונות הרעש ידועות או על ידי מדידת רעש הרקע המתווסף באינטרוולים בהם לא קיים דיבור [4]. אומדן ה- $|\hat{S}_w(\omega)|^\alpha$ יתקבל באמצעות נוסחה מספר 2.6 פיתוח והרחבה ראה נספח C , אולם אומדן זה לא יבטיח ערכים אי שליליים היות ואגף ימין של משוואה זו עלול להיות ערך שלילי. במספר מחקרים הדנים בנושא ההפחתה הספקטראלית ערכים שליליים אלו נקבעים כערכים חיוביים, ואילו במחקרים אחרים, ערכים אלו מאופסים. לבסוף על מנת לשחזר את האות לאחר ביצוע ההפחתה הספקטראלית יש לשלב בחזרה את המידע של הפאזה של האות הרועש המקורי במישור התדר והצגתו במישור הזמן תתקבל על ידי התמרת פורייה הפוכה באופן הבא:

$$\hat{S}_w(\omega) = |\hat{S}_w(\omega)| \cdot exp[j\measuredangle Y_w(\omega)] \tag{2.7}$$

$$\hat{s}_w(n) = F^{-1}\left[\hat{S}_w(\omega)\right] \tag{2.8}$$



איור 2.2: השבחת אות דיבור על ידי הפחתה ספקטראלית של הרעש [2].

2.2 מיצוי מאפיינים ספקטראליים

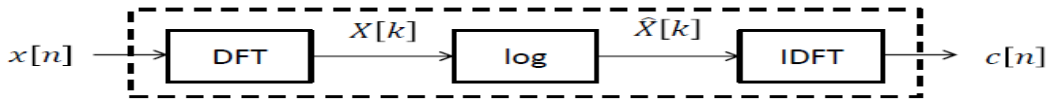
2.2.1 הקפסטרום "Cepstrum":

במציאות אין מערכת שיודעת להבחין בין מאפייני הדיבור של דובר אחד למשנהו בצורה חד משמעית, לעומת זאת הספקטרום של אות הדיבור מספק כלי יעיל לאפיון תכונות אות הדיבור לצורך זיהוי הדובר, הסיבה לכך נובעת מעצם העובדה שהספקטרום משקף את מערכת הקול אשר ייחודית עבור כל דובר, הדבר נעשה על ידי הצגה של העוצמות המתקבלות עבור כל מרכיבי התדרים והרי בכל מערכת לזיהוי קולי השלב הראשון הינו חילוץ אותם התכונות הספקטרליות אשר כוללות בתוכן את התוכן הלשוני ומאפייני רעשי הרקע המתווספים לאות הדיבור.

ה- "Cepstrum" תורם בצורה מקיפה עוד יותר לאנליזה של מאפייני הדיבור שחולצו באמצעות הספקטרום. המונח "Cepstrum" והגדרתו המתמטית נטבעו בשנת 1963 על ידי B.P. Bogert, M. J. R. Healy ו-J. W. Tukey . ההגדרה המתמטית של ה-"Cepstrum" הממשי הינה ההתמרה ההפוכה של לוגריתם הספקטרום של האות ומוצגת באופן הבא:

$$c[n] = \mathcal{F}^{-1}\{\log|x[n]|\} \tag{2.9}$$

האלגוריתם בעבורו מתקבל ה-"Cepstrum" מתואר באופן הבא:



איור 2.4: אלגוריתם ה-"Cepstrum".

אחד ההיבטים החשובים של ה-"Cepstrum" הינה קבלת צורת גל עבור אות הדיבור על ידי חישוב של לוגריתם הספקטרום לצורך המשך אנליזת פורייה, כתוצאה מכך ניתן לקבל דחיסה של הטווח הדינאמי של עוצמות ההרמוניות המתקבלות עבור כל תדר ותדר בטווח התדרים של האות הדגום ולפיכך להפחית את השפעת השינויים החדים ולקבל "מעטפת" חלקה יותר המייצגת את אות הדיבור הדגום.

באופן זה צורת הגל מקבלת מעין מחזוריות ("quasi-periodic") ובנוסף באופן מסוים אפנון תנופה, כדי להפריד בין השניים מבצעים התמרת פורייה הפוכה, כתוצאה מכך אנו נצפה לקבל את המחזוריות שמאפיינת את הדיבור, שזהו תדר הייסוד כמו כן יתקבלו מרכיבים בתחום התדרים הנמוך הנגרמים כתוצאה מאפנון התנופה, את המרכיבים הללו ניתן להפריד על ידי פעולת סינון פשוטה.

נציין שהמשתנה החופשי ב-"Cepstrum" נקרא בשם "Quefrency" , הנמדד ביחידות של זמן למרות שאין לערך זה משמעות במישור הזמן, כמו כן פעולת הסינון הלינארית שצוינה לעיל נקראת בשם "Liftering". שמות המונחים הללו, נטבעו על ידי ההוגים של צורת ההצגה הזו, שרצו להדגיש את ההמרה בין המישורים. במובן מסוים ניתן לייחס ל-"Cepstrum" משמעות של ספקטרום של הספקטרום.

2.2.2 סקלת ה-"Melody (Mel)" [9]:

במסגרת המחקר המדעי העוסק בתפיסה של הצלילים והתדרים המרכיבים אותם, נערכו ניסויים רבים אשר מטרתם הייתה להתאים סקלת תדרים המתאימה לתפיסת מערכת השמיעה של האדם. סקלת תדרים זו נחוצה מכיוון שמבנה האוזן הפנימית משמשת בתור מערכת אנליזה של הספקטרום, כמו כן מורכבות "מכניזם" של האוזן הפנימית ומערכת העצבים בה מרמזים כי תפיסת התכונות והמאפיינים של הצלילים בתדרים השונים לא בהכרח טריוויאלית או בעלי אופי לינארי.

ישנם מספר סטנדרטים לקנה מידה המתאים לתפיסת מערכת השמיעה של האדם והפופולארית ביניהם הינה סקלת ה-"Mel" אשר נטבעה על ידי Stevens, Volkman ו-Newman בשנת 1937. בקנה מידה זה מובלט תדר היסוד (pitch frequency) עבורו מערכת השמיעה של האדם רגישה יותר בתחום התדרים הנמוך יותר מאשר בתחום התדרים הגבוה לפיכך קנה מידה זה מאפיין בצורה טובה יותר את מה שהאדם שומע במציאות. קנה מידה זה הינו לינארי בתחום התדרים שמתחת ל-1 KHz ולוגריתמי מעל תחום תדרים זה מחד, עבור שני התחומים הללו מספר הדגימות נשאר זהה. קנה מידה זה פופולארי במערכות זיהוי דיבור מודרניות ומוגדרת מתמטית באופן הבא:

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right) = 1127\log_{e}\left(1+\frac{f}{700}\right) \qquad (2.10)$$

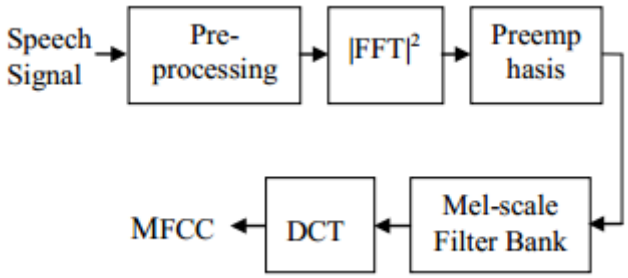
2.2.3 "MFCC":

המטרה המרכזית הניצבת בהבנת אות הדבור נעוצה בצלילים הנוצרים על ידי מערכת הקול של האדם אשר עוברת סינון באמצעות צורת מערכת הקול הכוללת בתוכה את הלשון, השיניים, השפתיים וחלל האף. לפיכך מבנה מערכת הקול קובעת אילו צלילים ייווצרו, לכן אילו נדע להעריך מבנה זה בצורה מדויקת מספיק נוכל לקבל ייצוג מדויק של הפונמות המופקות.

מבנה מערכת הקול בא לידי ביטוי במעטפת עוצמת ספקטרום אות הדיבור בפרקי זמנים לפיכך נרצה לעבוד עם שיטה שיודעת להעריך בצורה נאמנה את המעטפת הזו.

ישנן מספר שיטות שנועדו לייצג את מעטפת עוצמת הספקטרום במערכות זיהוי דיבור אוטומטיות וביניהן: מודל החיזוי הלינארי "Linear Prediction Coding (LPC)". מודל זה מנתח את האות על די שיערוך והסרה של הפורמנטים מאות הדיבור הדגום ולאחר מכן שיערוך העוצמה והתדרים של הזמזום שנותר. אלגוריתם זה פותח בשנת 1978 כאשר הרעיון הבסיסי עליו מתבססת שיטה זו היא היכולת לשערך אות דיבור דגום כקומבינציה לינארית של דגימות אות דיבור קודמות על ידי הקטנת הסכום של ריבוע ההפרשים באינטרוול זמן סופי, בין דגימות אות הדיבור המקורי לבין אלה שהתקבלו על יד החיזוי הלינארי, לבסוף תתקבל סדרה ייחודית של מקדמי חיזוי. שיטת חיזוי זו מעניקה חסינות, אמינות ודיוק עבור שיערוך הפרמטרים המאפיינים מערכות לינאריות משתנות בזמן [1]. בשנת 1980 Davis ו- Mermelstein הציגו אלגוריתם הקרוי בשם "Mel Frequency Cepstral Coefficient" [9] אשר הפך לאלגוריתם הפופולארי ביותר בקרב מערכות זיהוי דיבור אוטומטיות. אלגוריתם זה משלב את יתרונות קנה המידה של "Mel" ואת האנליזה של ה-"Cepstrum" שהוסברה בפרקים הקודמים, להבדיל ממודל החיזוי הלינארי.

<u>אלגוריתם זה מכיל את שלבי הביצוע הבאים:</u>



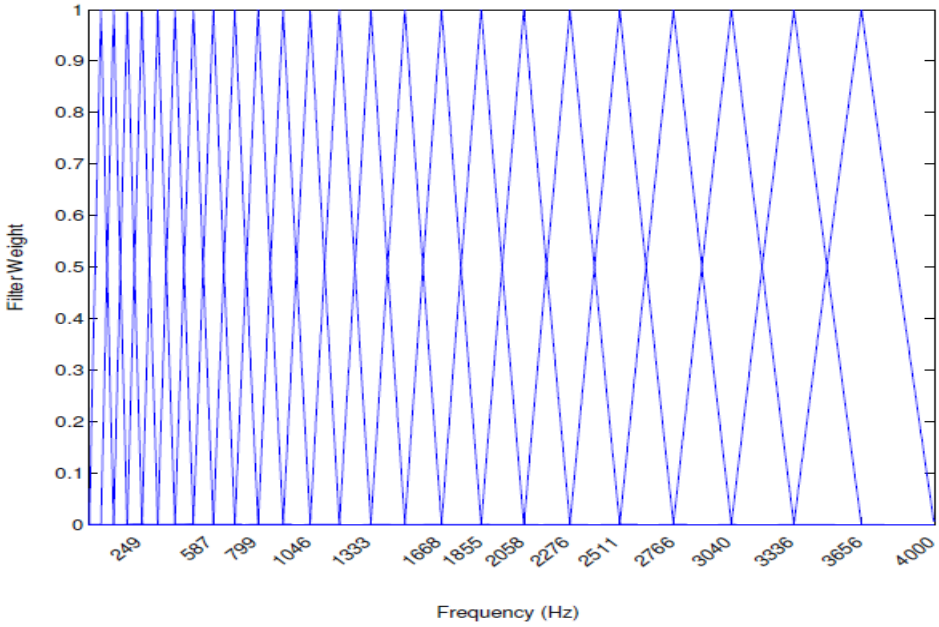
איור 2.5: אלגוריתם מיצוי המאפיינים בשיטת "MFCC".

1. לאחר ביצוע עיבוד מקדים לאות כפי שתואר בהרחבה בפרק 2.1, תחושב העוצמה הספקטראלית המתקבלת עבור כל מסגרת, על מנת לקבל הערכה על התדרים הקיימים בכל מסגרת.
2. עבור כל מסגרת יבוצע תכנון מסננים בעלי סדר N, שיציינו את כמות המקדמים הספקטראליים שיתארו את מעטפת אות הדיבור, המסננים מותאמים לקנה מידה של "Mel" כאשר רוחב הפס של כל מסנן משתנה ככל שיעלה התדר, בנק המסננים הללו מכיל מסנני מעבירי פס משולשים ועוקבים באופן כזה שבו תדר הקטעון הנמוך של מסנן מעביר הפס ימצא במרכז תחום ההעברה של המסנן שקדם לו וכמו כן תדר הקטעון הגבוה ימצא במרכז תחום התדרים של המסנן שבא אחריו וכתוצאה מתקיימת חפיפה בין מסננים עוקבים. לבסוף מסננים אלה מעניקים הערכה של כמות האנרגיה המצויה לאורך כל תחומי התדרים.

בנק מסננים בעל מספר M של מסננים (m=1,2,…,M) כאשר מסנן m הינו מסנן משולש אשר מוגדר באופן הבא:

$$H_{[k]=f(x)m} = \begin{cases} 0 & k < f[m-1] \\ \frac{2(k-f[m-1])}{(f[m+1)-f[m-1])(f[m]-f[m-1])} & f[m-1] \le k \le f[m] \\ \frac{2(f[m+1)-k)}{(f[m+1)-f[m-1])(f[m+1]-f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases} \qquad (2.11)$$

מסננים אלו יבצעו חישוב של הספקטרום הממוצע סביב כל תדר מרכזי כאשר רוחב הפס בכל מסנן כזה ילך ויגדל. ההצגה של בנק מסננים נתונה באופן הבא:



איור 2.6: בנק מסננים משולשים לצורך חישוב ה-"Mel-cepstrum" [9] .

3. הפעלת לוגריתם עבור האנרגיות שיתקבלו במוצא המסננים, במטרה להתאים אנרגיות אלה לקנה מידה של מערכת השמיעה של האדם.
4. הפעלת התמרת קוסינוס דיסקרטית (DCT) על הלוגריתם שיתקבל בשלב הקודם לצורך קבלת מספר "N" של מקדמים, לבסוף, בדרך כלל, משתמשים ב- 12-13 המקדמים הראשונים שיתקבלו עבור כל מסגרת ואלה יקראו "Mel Frequency Cepstral Coefficient". התמרת הקוסינוס הדיסקרטית נמצאת בשימוש נרחב עבור עיבוד אות דיבור, כמו כן יש להתמרה זו מספר הגדרות.

ה- DCT2 $C[k]$ עבור אות ממשי $x[n]$ מוגדרת באופן הבא:

$$C[k] = \sum_{n=0}^{N-1} x[n] \cos\left(\frac{\pi k\left(n+\frac{1}{2}\right)}{N}\right) \quad \forall\, 0 \le k < N \qquad (2.12)$$

וההתמרה ההפוכה נתונה באופן הבא:

$$x[n] = 1/N\left\{C[0] + 2\sum_{k=1}^{N-1} C[k]\cos\left(\frac{\pi k\left(n+\frac{1}{2}\right)}{N}\right)\right\} \ \forall\ 0 \le n < N \qquad (2.13)$$

התמרה זו הינה השימושית מבין ההגדרות של התמרת הקוסינוס הדיסקרטי מכיוון שזו מאפשרת דחיסת אנרגיה וכתוצאה מתקבלים מקדמים אשר מרוכזים יותר סביב המרכיבים הספקטראליים בתדרים הנמוכים. מאפיין זה מאפשר הערכה של האות על ידי מספר מצומצם יותר של מקדמים [5].

2.3 מודל שערוך הסתברותי

2.3.1 מודל תערובת הגאוסיאניים "Gaussian Mixture Model (GMM)"

מודל תערובת ההתפלגויות הנורמליות ובלועזית "Gaussian Mixture Model" הינו פונקציית צפיפות הסתברותית פרמטרית המיוצגת בתור סכום משוקלל של צפיפויות המרכיבים המתפלגים נורמלית. ה-GMM נמצא בשימוש נרחב ביישומים בהם דרושות מדידות של מאפיינים או תכונות של מערכות ביומטריות. דוגמא למערכת אחת כזו הינה מערכת קול האדם אשר מפיקה מאפיינים ספקטראליים ייחודיים לכל דובר ואלה מאפשרים זיהוי או אימות של הדובר על ידי מערכות זיהוי קולי.

פרמטרי ה-GMM משוערכים באמצעות בסיס מידע של נתונים שמטרתם הינה אימון המערכת על ידי אלגוריתם "Expectation-Maximization (EM)" הנועד לאמוד את הפרמטרים במידע חלקי בצורה איטרטיבית שעליו נרחיב בהמשך.

מודל תערובת ההתפלגויות הנורמליות הינו סכום משוקלל של M צפיפויות הסתברותיות של מרכיבים אשר מתפלגים בהתפלגות נורמלית. מודל זה נתון במשוואה הבאה:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i g(\vec{x}|\vec{\mu_i}, \Sigma_i) \qquad (2.14)$$

כאשר משתנה $\vec{x}$ הינו ווקטור משתנה המידע הרציף מממד D, המכיל את התכונות או המאפיינים של האות הנמדד, w_i עבור i=1,…,M הינם המשקלים עבור כל תערובת, ובאשר ל- $g(\vec{x}|\vec{\mu_i}, \Sigma_i)$ עבור i=1,…,M הינם מרכיבי צפיפות ההסתברותית של ההתפלגויות הנורמליות. לכל מרכיב מותאמת צפיפות הסתברותית של התפלגות נורמלית מממד D הנתונה במשוואה הבא:

$$g(\vec{x}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu_i})'^{\Sigma_i^{-1}}(\vec{x}-\vec{\mu_i})\right\} \qquad (2.15)$$

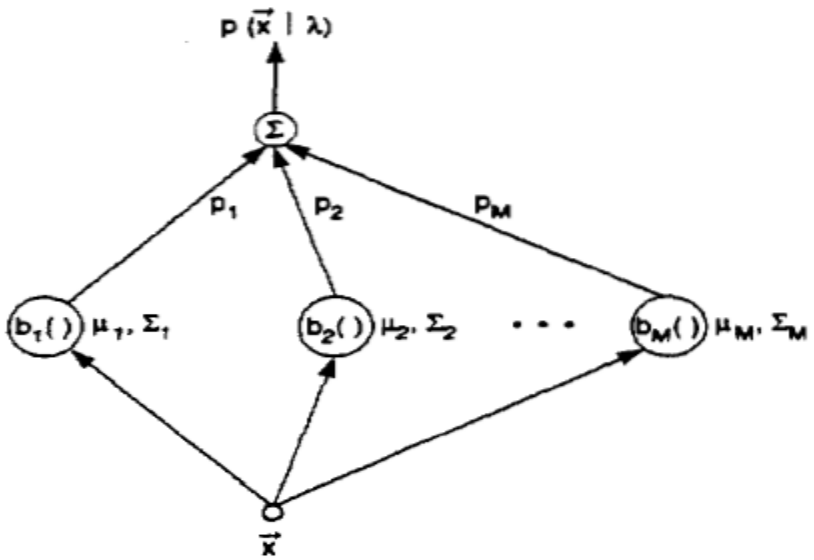
כאשר $\overrightarrow{\mu_i}$ הינו וקטור התוחלת ו-Σ_i הינה מטריצת השונות המשותפת, כמו כן משקל כל תערובת כזו מקיימת את האילוץ:

$$\sum_{i=1}^{M} w_i = 1 \qquad (2.16)$$

לסיכום GMM מיוצג על ידי פרמטר ווקטור התוחלת, מטריצת השונות המשותפת והמשקלים של כל תערובת מסך כל הצפיפויות ההסתברותיות. קבוצת הפרמטרים הללו יסומנו בתור:

$$\lambda = \{w_i, \overrightarrow{\mu_i}, \Sigma_i\} \ \forall\, i = 1,..,M \qquad (2.17)$$

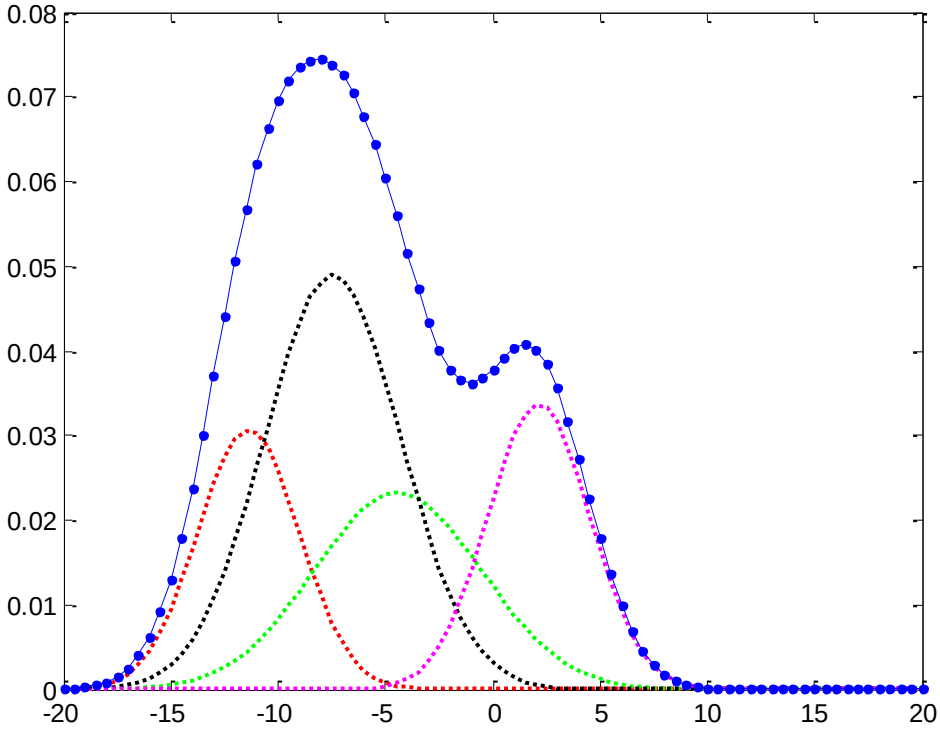
כך שבעבור מערכות לזיהוי דובר המשמעות תהיה שכל דובר יהיה מיוצג על ידי GMM ו-λ המאפיינת אותו/ה.



איור 2.7: סכום משוקלל של M צפיפויות הסתברותיות של מרכיבים אשר מתפלגים בהתפלגות נורמלית [3] .

ל-GMM ישנם מספר תצורות התלויות בדרך בה בוחרים את מטריצת השונות המשותפת. המודל יכול לכלול מטריצת שונות משותפת עבור כל מרכיב התפלגות נורמלית לחוד, מטריצת שונות משותפת עבור כל המרכיבים בהתפלגות הנורמלית עבור מודל הדובר או מטריצת שונות משותפת אשר תהיה משותפת בין כל הדוברים והמודלים המתאימים עבורם. נציין שמטריצה זו יכולה להיות שלמה או אלכסונית. הדרך בה יש לבחור את צורתה של המטריצה מתבססת בעיקר על תוצאות הניסוי הראשוני, למעשה התוצאות מוכיחות שבעבור השונות האלכסונית במטריצת השונות המשותפת מתקבלים ביצועים טובים יותר מאשר במטריצה השונות המשותפת השלמה.

המוטיבציה העומדת מאחורי השימוש במודל תערובת ההתפלגויות הנורמלית לצורך תיאור זהות הדובר נובעת משני סיבות עיקריות, מקור הראשונה מבניהם נובע מהרעיון האינטואיטיבי ובו כל מרכיב צפיפות הסתברותית מסך כל התערובות מסוגל לאפיין קבוצה בסיסית של מאפיינים אקוסטיים. מקובל להניח שהמרחב האקוסטי תואם לקול הדובר המיוצג על ידי קבוצה של מאפיינים אקוסטיים המתוארים על ידי קבוצה רחבה של מאורעות פונטיים וביניהם תנועות, צלילים המופקים דרך דרכי הנשימה לרבות חלל האף וצלילי חיכוך. מאפיינים אקוסטיים אלו מייצגים את מערכת קול הדובר. הצורה הספקטראלית של מאפיין ספקטראלי בודד, כלומר עבור אינדקס i כלשהו יכול להיות מיוצג על ידי ווקטור תוחלת $\vec{\mu_i}$ עבור מרכיב הצפיפות ההסתברותית בעלת האינדקס i בהתאמה. כמו כן השינויים של הצורה הספקטראלית הממוצעת יכולה להיות מיוצגת על ידי מטריצת השונות המשותפת Σ_i. היות וכל בסיס המידע בו טמון אות הדיבור לא מתויג, המאפיינים האקוסטיים החבויים בו לא מוגדרים, לפיכך בהנחה בה ווקטור משתנה המידע הרציף הינו בלתי תלוי, מהתבוננות על הצפיפות ההסתברותית של ווקטור משתנה המידע הרציף, אשר חולץ מהמאפיינים האקוסטיים החבויים, נקבל ייצוג של תערובת ההתפלגות הנורמלית. מוטיבציה נוספת לשימוש במודל תערובת ההתפלגויות הנורמליות נובע מתצפיות ניסיוניות בהן קומבינציה לינארית של מספר פונקציות בסיס מהתפלגות נורמלית יכול לייצג קבוצה גדולה של מדגמים בהתפלגות זו. כמו כן אחד התכונות המועילות של ה-GMM זוהי היכולת ליצור שיערוך חלק עבור צפיפויות הסתברותיות אקראיות.



איור 2.8: דוגמא ל-GMM (הקו הרציף) המורכב מארבעה מרכיבים (הקו המקווקו).

2.3.2 מקסום השיערוך "Expectation Maximization (EM)" [2]

כעת לאחר שסקרנו על מודל תערובת ההתפלגויות הנורמליות, נדון בהרחבה על אלגוריתם הנועד לאמוד את הפרמטרים המיצגים של ה-GMM, $\lambda = \{w_i, \vec{\mu_i}, \Sigma_i\}$, המתאימים ביותר עבור ההתפלגות של ווקטור משתנה המידע הרציף, $\vec{x}$, אלגוריתם זה הינו אלגוריתם "Expectation -Maximization" שהוצג לראשונה בשנת 1977 על ידי Arthur Dempster, Nan Laird ו-Donald Rubin [9], אלגוריתמי EM הם משפחה של אלגוריתמים למציאת מקסימום נראות תחת אינפורמציה לא מלאה. עבור רצף של מספר T של ווקטורים של משתנה המידע הרציף $\vec{x} = \{x_1, \dots, x_1\}$, הסבירות של ה-GMM, בהנחה ומתקיימת אי תלות בין הווקטורים, יכולה להיות מוצגת בצורה הבאה:

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda) \tag{2.18}$$

אולם ביטוי זה הינו פונקציה לא לינארית של קבוצת הפרמטרים λ כמו כן מקסום ישיר של האומדן איננו אפשרי, למרות זאת אומדן הפרמטרים עבור הסבירות המירבית "Maximum Likelihood (ML)" עבור הנראות יכול להיות מושג באמצעות האלגוריתם האיטרטיבי EM. הרעיון המונח בבסיס אלגוריתם זה הוא הנחת מודל התחלתי λ לצורך אומדן חדש עבור מודל זה שיקיים את הביטוי הבא:

$$p(X|\bar{\lambda}) \geq p(X|\lambda) \tag{2.19}$$

המודל החדש לפיכך יקיים את המודל ההתחלתי עבור האיטרציה הבאה והתהליך כולו חוזר על עצמו עד אשר תתקיים התכנסות עבור ערך סף מסוים, כאשר האומדן בין כל איטרציה ואיטרציה, מבטיח עליה מונוטונית בערך הסבירות של המודל λ למול ערכו ההתחלתי, הסיבה לכך טמונה בנוסחאות המחושבות בכל צעד באיטרציה בודדת עבור אומדן $\lambda = \{w_i, \vec{\mu_i}, \Sigma_i\}$, ואלו הנוסחאות הבאות

המשקל עבור כל תערובת:

$$\overline{w_i} = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|\vec{x}_t, \lambda) \tag{2.20}$$

ווקטור התוחלת:

$$\bar{\mu_i} = \frac{\sum_{t=1}^{T} \Pr(i|\vec{x}_t, \lambda) x_t}{\sum_{t=1}^{T} \Pr(i|\vec{x}_t, \lambda)} \tag{2.21}$$

ווקטור השוניות (אלכסון מטריצת השונות המשותפת)-

$$\bar{\sigma_i}^2 = \frac{\sum_{t=1}^{T} \Pr(i|\vec{x}_t, \lambda) {x_t}^2}{\sum_{t=1}^{T} \Pr(i|\vec{x}_t, \lambda)} - \bar{\mu_i}^2 \tag{2.22}$$

ראוי לציין שבשלב אימון ה-GMM בעבור מודל הדובר קיימים שני גורמים מכריעים ואלו כמות התערובות M והנחת המודל ההתחלתי λ בטרם תחילת פעולת האלגוריתם EM למציאת מקסימום הנראות. למעשה אין כלים תאורטיים טובים למדי כדי להנחות הערכה טובה לבחירה נבונה של הפרמטרים הללו, לפיכך קביעת הפרמטרים הללו יתקבלו באופן ניסיוני.

2.3.3 <u>זיהוי הדובר</u>

לאחר שלב אימון הדוברים במאגר, שלב זיהוי הדובר יבוצע עבור קבוצה של S דוברים, דהיינו $S = \{1,2,\dots,S\}$ כאשר קבוצה זו מיוצגת במונחים של מודל תערובות הגאוסייאנים עבור כל דובר, כלומר $\lambda_1, \lambda_2, \dots, \lambda_S$. המטרה בשלב זה תהיה למצוא את מודל הדובר לו ההסתברות הפוסטריורית המקסימאלית עבור רצף הנראות הנתון, האינטרפרטציה במונחים המתמטיים לשלב זה הינה:

$$\hat{S} = \arg\max_{1\le k\le S} \Pr(\lambda_k|X) = arg\max_{1\le k\le S} \frac{p(X|\lambda_k)\Pr(\lambda_k)}{p(X)} \tag{2.23}$$

כאשר המשוואה השניה נובעת מחוק בייס. כעת, בהנחה ומתקיימת הסתברות שווה לכל דובר מהמאגר, דהיינו $\Pr(\lambda_k) = \frac{1}{S}$, כמו גם ש- $p(X)$ זהה לכל הדוברים במאגר, הרי שניתן לפשט את משוואה (2.23) לביטוי הבא:

$$\hat{S} = \arg\max_{1\le k\le S} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_k) \tag{2.24}$$

כאשר $p(\vec{x}_t|\lambda_k)$ נתון במשוואה (2.23) לעיל.

2.3.4 מודל מרקובי חבוי "Hidden Markov Model (HMM)"

שיטה סטטיסטית נוספת מבין השיטות הסטטיסטיות הקיימות [1], [6], לצורך פרמטריזציה של המאפיינים הספקטראליים ליישומי זיהוי או אימות הינה המודל המרקובי החבוי או בלועזית (HMM) "Hidden Markov Model". מודל זה מהווה שיטה סטטיסטית רבת עוצמה לצורך אפיון סדרה של דגימות מידע דיסקרטיות בזמן. לא רק שמודל זה מסוגל להעניק דרך יעילה להרכבת מודל פרמטרי חסכוני, אלא שמודל זה יכול להכיל בתוך ליבתו את עקרונות התכנון הדינאמיים של חלוקה אחידה למקטעים וסיווג תבניות החוזרות על עצמן של מקטעי מידע המשתנים בזמן. ראוי לציין שסדרת הדגימות יכולות להיות רציפות או דיסקרטיות, ווקטוריות או סקלריות. ההנחה היסודית של מודל זה הינה שדגימות המידע יכולות להיות מאופיינות בתור תהליך פרמטרי אקראי, והרי פרמטרים של תהליך סטוכסטי יכולים להיות משוערכים על ידי אלגוריתם מדויק שיוגדר מראש. בסיס התאוריה העומדת מאחורי מודל זה פורסמה לראשונה בסדרה של עיתונים בשנות ה-60 של המאה ה-20 על ידי Leonard E.Baum ועמיתיו לעבודה. תאוריה זו התבססה על התאוריה המתמטית של Andrei Markov שפותחה בשנות ה-20 במאה ה-20. לאורך הזמן מודל זה הפך להיות לאחד המודלים הפופולאריים ביותר לצורך בניית מודל עבור אותות דיבור. את עקרונותיו של המודל ניתן למצוא במגוון רחב של יישומים וביניהם: מערכות אוטומטיות לזיהוי דיבור, מעקב אחר הפורמנטות והתדר היסודי, מערכות להשבחת אות הדיבור, סינתזה של אות הדיבור, בניית מודל סטטיסטי עבור שפות שונות וכדומה.

מודל זה מוגדר באופן רשמי בתור סדרת נתונים המכילה את הפרמטרים (S, V, π, A, B), כאשר $S = \{s_1, \dots, s_N\}$ מציין סדרה אינסופית של N מצבים, $V = \{v_1, \dots, v_m\}$ מציין את סדרת M סימבולים אפשריים מתוך סך כל אוצר המילים, $\pi = \{\pi_i\}$ אלו ההסתברויות המצבים ההתחלתיות, $A = \{a_{ij}\}$ מציין את הסתברויות המעבר בין המצבים השונים, $B = \{b_i(k)\}$ כאשר $b_i(k)$ הינה ההתסתברות לפלט k כאשר נמצא במצב S_i. על מנת לציין את כל הפרמטרים שהוזכרו לעיל ניעזרים בסימון $\lambda = (\pi, A, B)$. משמעותם של הפרמטרים המרכיבים את λ הינם:

- π_i – ההסתברות שהמערכת תתחיל במצב i בתחילת העבודה.
- a_{ij} – ההסתברות למעבר בין מצב i למצב j.
- $b_i(v_k)$ – ההסתברות ליצירת סימבול v_k במצב i.

כאשר על כל אחד מהפרמטרים הללו חלים האילוצים הבאים:

$$\sum_{i=1}^{N} \pi_i = 1 \tag{2.25}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \text{ for } i = 1,2 \dots, N \tag{2.26}$$

$$\sum_{j=1}^{N} b_i(v_k) = 1 \text{ for } i = 1,2 \dots, N \tag{2.27}$$

מודל ה-HMM מכיל מספר סוגיות וביניהם:

1. <u>שיערוך</u>: הערכת ההסתברות עבור סדרה של סימבולים נראים $O = o_1 o_2 \dots o_T$ כאשר $o_i \in V$, בהינתן HMM מסוים, דהיינו $p(O|\lambda)$.

2. <u>פענוח</u>: מציאת מסלול מעבר המצב בעל הסבירות הגבוה ביותר עבור סדרת הנראות. נניח ש- $q = q_1 q_2 \dots q_T$ הינה סדרה של מצבים, המטרה לפיכך תהיה למצוא $q^* = argmax_q p(q, O|\lambda)$, או באופן דומה $q^* = argmax_s p(q|O, \lambda)$.

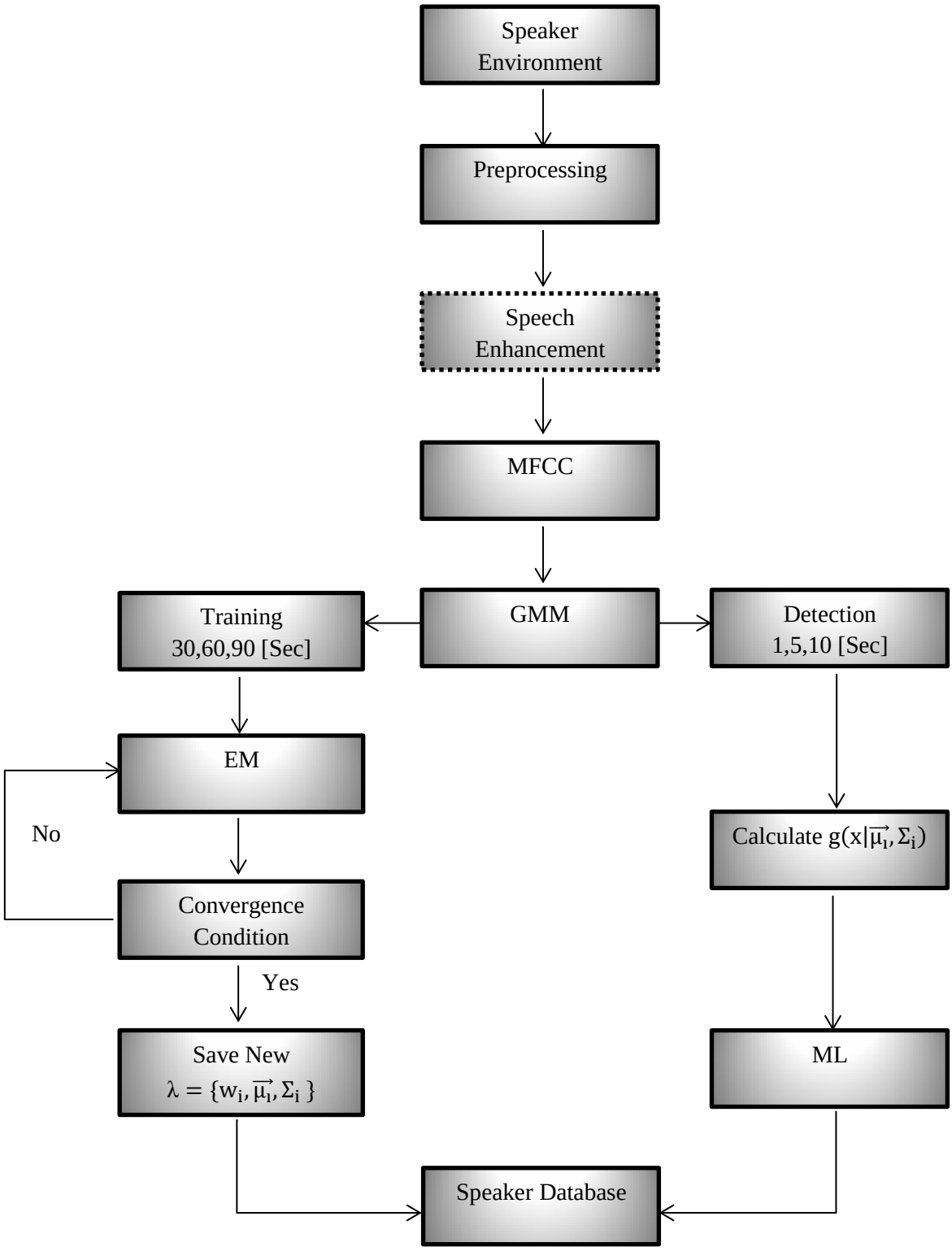
3. <u>אימון</u>: התאמת כלל הפרמטרים המוכלים ב- λ לצורך מקסום ההסתברות ליצירת סדרת הנראות, כלומר מציאת $\lambda^* = argmax_\lambda p(O|\lambda)$.

פתרון לסוגיה הראשונה שמתלווה למודל ה-HMM הינו האלגוריתם האיטרטיבי של "Forward & Backward", ואילו הפתרון לסוגיית הפענוח הינו אלגוריתם "Viterbi". אלגוריתם איטרטיבי זה מספק דרך למציאת מסלול מעבר המצב שיהיה בעל הסבירות הגבוהה ביותר, ולבסוף, הפתרון לסוגיית האתחול הינו אלגוריתם ה-"Baum-Welch", אשר מנצל את הסתברויות ה- "Forward & Backward" לצורך עדכון הפרמטרים באופן איטרטיבי.

לסיכום, בשיטה זו ההנחה היא שאות הדיבור יכול להיות מתואר כתהליך פרמטרי אקראי המשוערך על ידי סדרה סטוכסטית וסופית כווקטור התצפיות של המאפיינים הספקטראליים בה כל תצפית מהווה אירוע קולי או לחילופין דובר כך שכל מצב בווקטור התצפיות תלוי במצבו הקודם. ראוי לציין שלמודל זה יתרון עבור מטלות בהן קיימת תלות בטקסט הנאמר, אולם עבור מטלות בהן אין תלות בטקסט הרי שרצף הדגימות של המאפיינים הספקטראליים המצויים בשלב הקלטת אות הדיבור לאו דווקא ישקפו בצורה נאמנה דיה את הרצף המצוי במאגר המידע המאומן של האלגוריתם [5]. להרחבה ראה נספח D.

3 הצעת תכנון

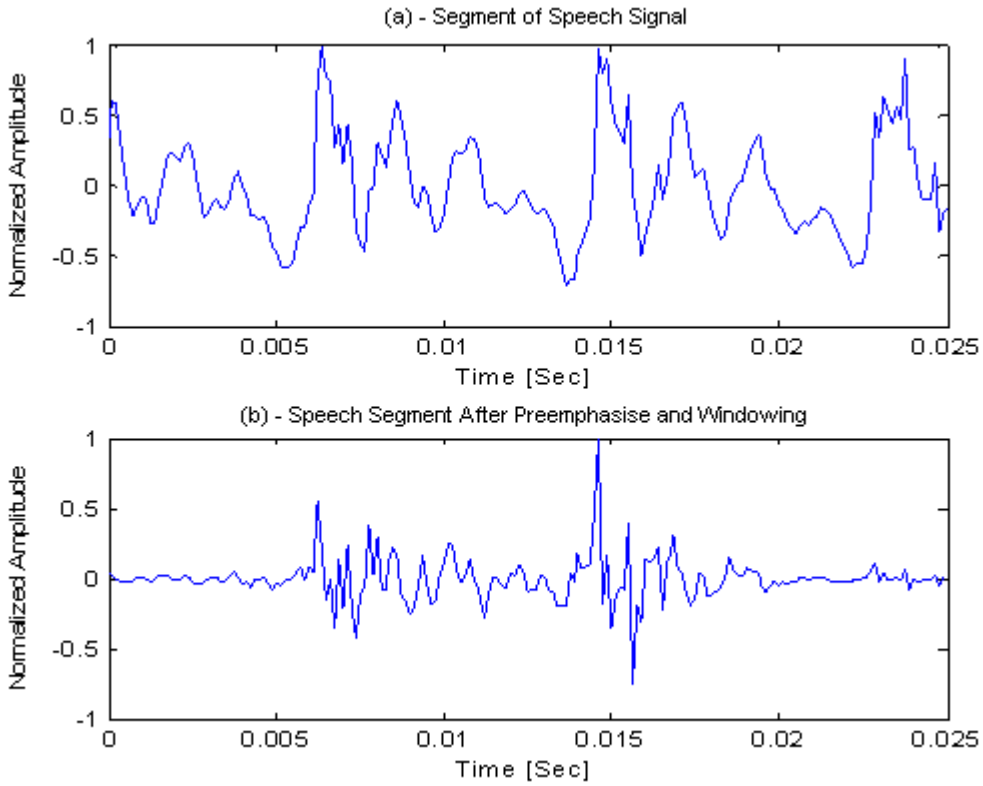
3.1 סכמת דיאגרמת מלבנים



איור 3.1: דיאגרמת מלבנים עבור תכנון הפרויקט.

3.2 הסבר סכימת דיאגרמת מלבנים

1. המערכת לעיל מתחלקת לשני רבדים, כאשר הראשון מביניהם הינו יצירה של מאגר דוברים בסביבה שאינה רועשת לצורך אימון האלגוריתם ויצירת בסיס נתונים ידוע, כאשר אורכי הזמן שנקלטים במערכת הם 30, 60 ו-90 שניות. הרובד השני מתייחס לשלב הזיהוי של הדובר מתוך אותה קבוצת דוברים מהמאגר שאימנו, כאשר אורכי הזמן שנקלטים במערכת הם 1 שניה, 5 ו-10 שניות [3]. ברובד זה המערכת תקלוט את הדובר הנבחן ללא ועם התווספות של רעשי רקע. המטרה של זמני אימון ובדיקה משתנים היא לחקור את ביצועי המערכת בעזרת מדד של אחוזי ההצלחה בזיהוי של הנבדק ולהסיק את התנאים האופטימליים הדרושים לתפעול המערכת.
2. שלב עיבוד מקדים של האות הדיבור, אשר משותף לקלט שצוין לעיל, מתבצע באופן הבא:
   - הורדת רכיב DC מהאות דיבור.
   - חלוקה של האות למסגרות זמן בעל אופי סטציונרי.
   - הכפלה בחלון Hamming להורדת עוצמות של אונות צד.
   - Pre-Emphasize של כל מסגרת באמצעות מסנן מעביר גבוהים.



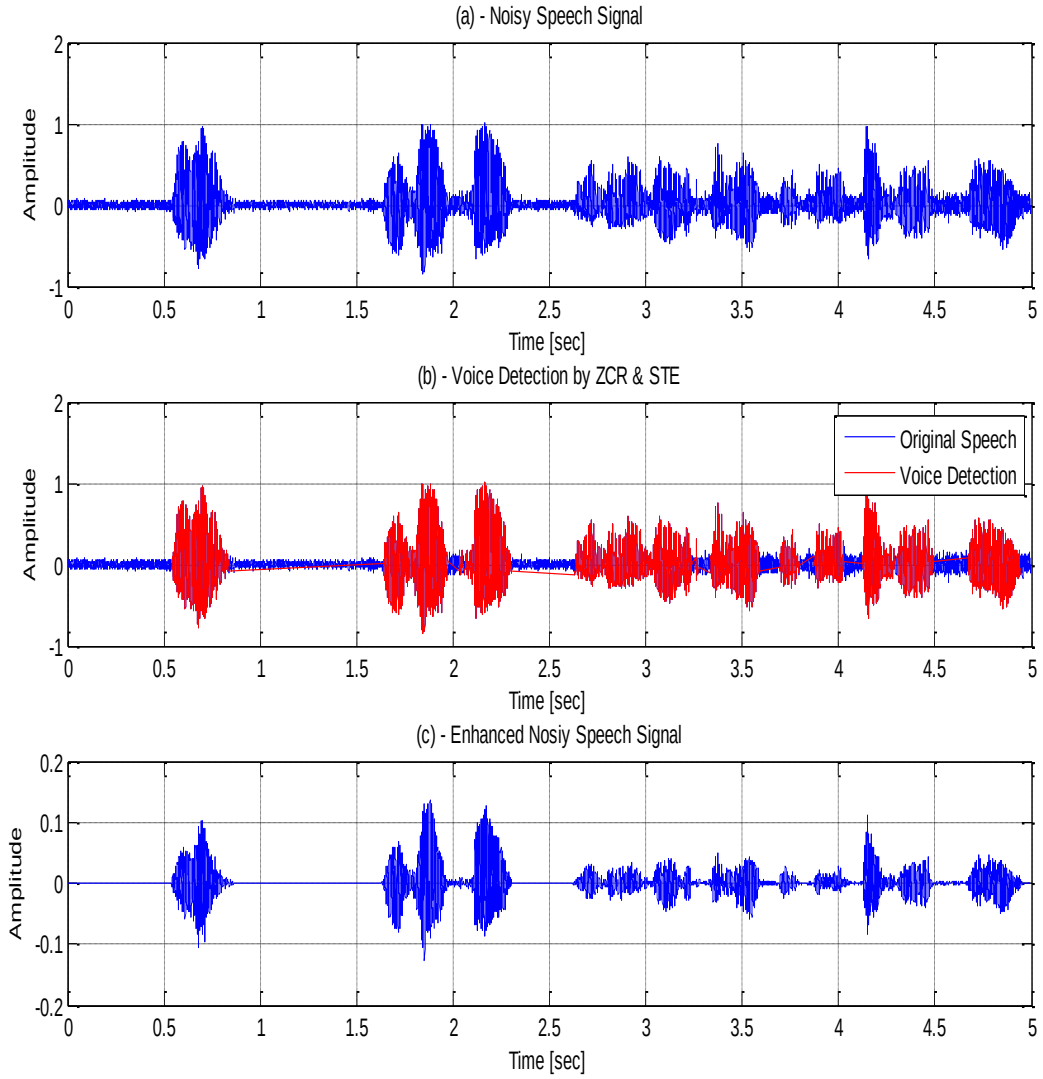
איור 3.2: [a] – מסגרת של אות דיבור באורך של 25 מילי שניות,

[b] – אותה מסגרת אות דיבור לאחר הכפלה בחלון Hamming

והעברה דרך HPF.

3. השבחת אות הדיבור ברובד הזיהוי דובר:

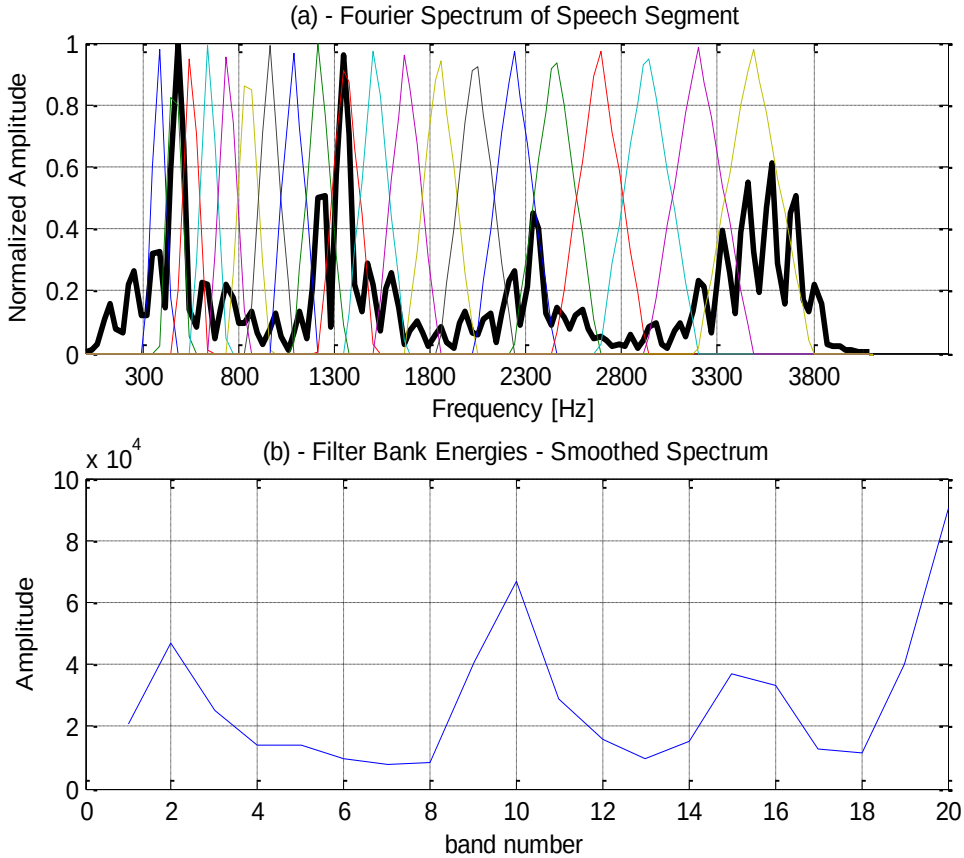
שלב זה נותן לנו אפשרות למדוד את אחוזי הצלחה של המערכת בזיהוי דובר הכולל רעשי רקע מתווספים לאות דיבור, כמו כן, גם את אחוזי הצלחה בזיהוי של הדובר לאחר יישום של אלגוריתם להשבחת האות דיבור. השיטה להשבחת האות דיבור מוסברת בהרחבה בתת-פרק 2.1.2. נציין שגם נעזר באלגוריתם של Voice Activity Detector (VAD) לזיהוי מקטעים קוליים שמהם נוכל באופן איטרטיבי להוריד את רעשי הרקע. מכאן נוכל לתת אומדן עד כמה מועילה או משפרת השיטה להשבחת אות דיבור אם בכלל עבור אחוזי ההצלחה של המערכת לזיהוי דובר.



איור 3.3: [a] – אות דיבור עם התווספות של רעש רקע , [b] – זיהוי מקטעים קוליים באמצעות אלגוריתם VAD, [c] – השבחה של אות דיבור רועש.

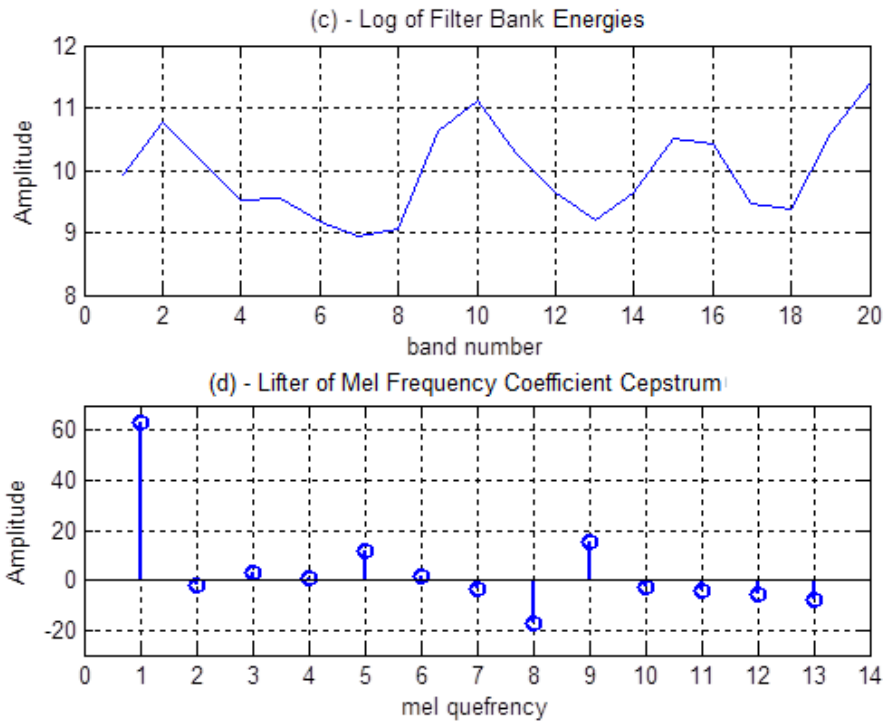
4. שלב מיצוי מאפיינים ע״י אלגוריתם MFCC :

- חישוב עוצמה ספקטרלית עבור מסגרות הזמן של האות דיבור.
- יצירת בנק מסננים משולשים בעל מרווחים המותאמים לקנה מידה של "Mel".
- סכמת אנרגיה ספקטרלית של אות הדיבור באמצעות בנק המסננים.
- הפעלת לוגריתם על הספקטרום וביצוע התמרת קוסינוס דיסקרטית לצורך דחיסה של האות וקבלת מספר מאפיינים ספקטרלים מצומצם.
- ביצוע "Lifter" עבור מקדמי MFCC שקיבלנו לצורך הדגשה עבור המקדמים שמאפיינים את התדרים הגבוהים יותר של זהות הדובר.



איור 3.4: [a] – חישוב FFT ותכנון בנק מסננים מותאם.

[b] – סכמה של האנרגיה הספקטרלית כתוצאה מהבנק מסננים.



איור 3.5: [c] – הפעלת לוגריתם של הספקטרום, [d] – מקדמי MFCC: לאחר DCT ו-"Lifter" של המקדמים שהתקבלו.

5. שלב אימון של מאגר הדוברים ע"י אלגוריתם איטרטיבי EM לחישוב פרמטרים סטטיסטים עבור מודל גאוסיוני הסתברותי עבור כל אחד מהדוברים מתואר באופן הבא:

א. קביעת תנאי התחלה עבור כל אחד מהדוברים במאגר: $\lambda = \{w_i, \vec{\mu_i}, \Sigma_i\}$, עבור כל דובר תנאי ההתחלה יקבעו באופן רנדומלי מתוך וקטורי המאפיינים של אותו דובר ובעל פילוג משקלים שווה עבור מספר מרכיבים גאוסיונים שהוחלט מראש.

ב. חישוב שערוך הסתברותי עבור כל אחד מהמרכיבים $p(\vec{x}|\lambda)$, נוסחה 2.14.

ג. חישוב הסתברות פוסטריורית $p(i|\vec{x_t}, \lambda) = \frac{p_i b_i(\vec{x_t})}{\sum_{k=1}^{M} p_k b_k(\vec{x_t})}$.

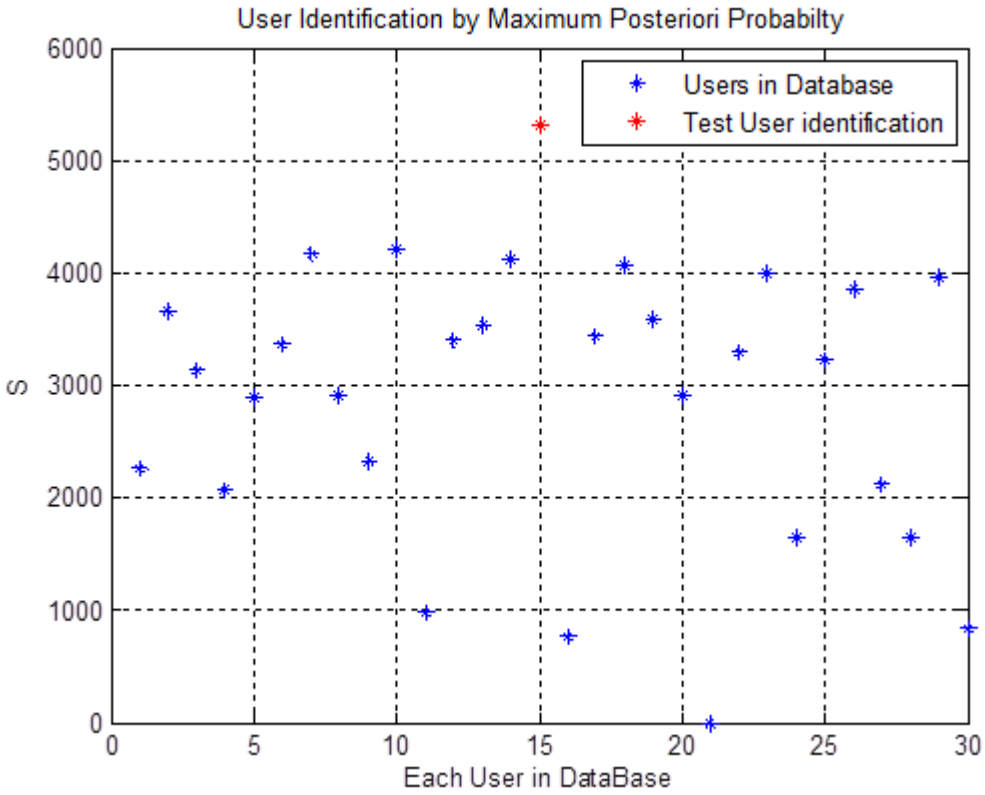
ד. חישוב ערכים סטטיסטיים של מודל תערובת גאוסיונית $\lambda = \{w_i, \vec{\mu_i}, \Sigma_i\}$ לפי נוסחאות: 2.20, 2.21 ו- 2.22.

ה. בדיקת תנאי התכנסות, $p(X|\bar{\lambda}) \geq p(X|\lambda)$ אם התנאי מתקיים יש לחזור על סעיפים ב', ג' ו-ד'.

ו. ברגע שהתהליך האיטרטיבי מסתיים נשמרים הערכים הסטטיסטים שחושבו עבור כל דובר במאגר הנתונים אשר ישמש את המערכת לצורך זיהוי דובר.

6. שלב זיהוי דובר מתוך מאגר דוברים מאומן ע״י שערוך מקסימלי (ML)

בתהליך הזיהוי נלקחים וקטורי המאפיינים של הדובר הנבדק ומוצבים במודל ההסתברותי גאוסיוני (GMM) ביחד עם הפרמטרים ההסתברותיים שחושבו בשלבי האימון עבור כל דובר במאגר ועל ידי השימוש בנוסחה 2.24 כאשר נוסחה זו מאפשרת למצוא את מודל הדובר לו ההסתברות הפוסטריורית המירבית.



איור 3.6: זיהוי דובר מתוך מאגר של 30 דוברים ע״י הסתברות הפוסטריורית המקסימאלית.

4 מערך בדיקות סופיות

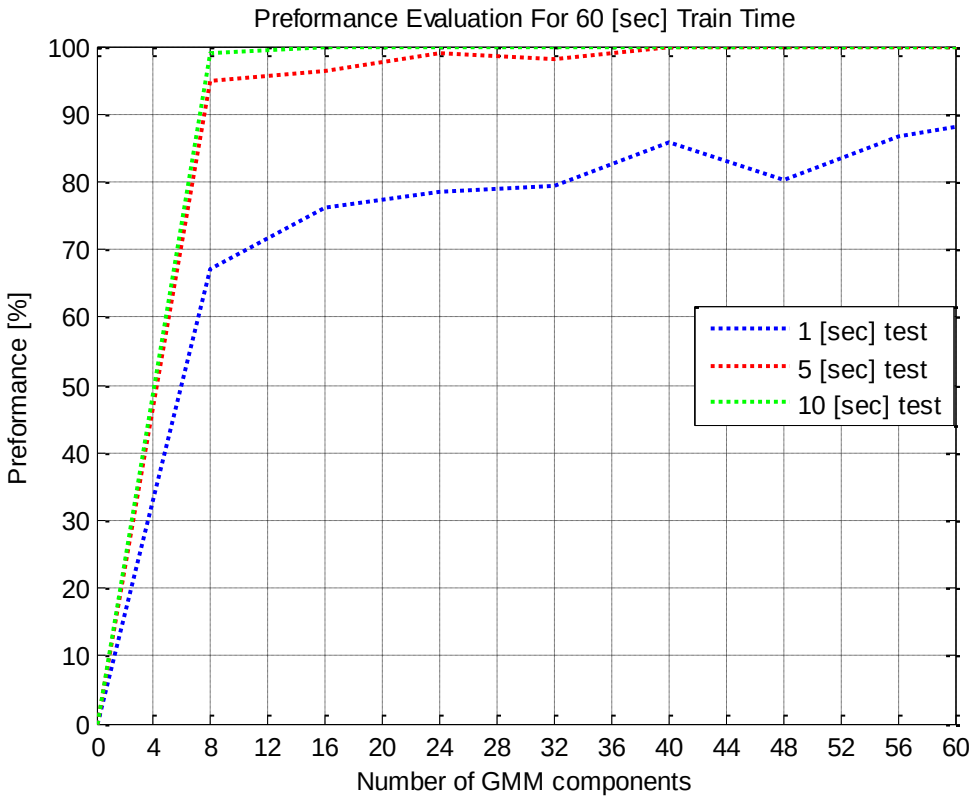
לצורך מדידת ביצועי המערכת נערכת סדרה של השוואות בין קטעים של ווקטור המאפיינים של אות הדיבור הנבחן לבין קטעים של ווקטור המאפיינים של יתר הדוברים הקיימים במאגר. לצורך השוואה של מקטעי הדיבור באורכים שונים של הדובר מול יתר הדוברים במאגר, רצף ווקטור המאפיינים יתחלק למספר T של מקטעים חופפים בערכים שונים של חפיפה בין המקטעים הללו. לבסוף הדובר שיזוהה בכל מקטע יושווה לדובר האמיתי והתהליך יחזור על עצמו עבור יתר הדוברים במאגר.

ראוי לציין שלא קיימת בידינו בשלב הזה הערכה למדד הביצועים שאנו מצפים לקבל עבור מידת ההצלחה של זיהוי הדובר בסביבה רועשת, אולם לצורך קבלת הערכה ראשונית ונקודת ייחוס נמדוד את רמת הביצועים של המערכת בתנאי סביבה שאינה רועשת ועבורה, כתלות במספר פרמטרים, ביניהם משך זמן ההקלטה וכמות הדוברים במאגר, כבר ידועים לנו [3]. לבסוף מדידת הביצועים תחושב באופן הבא:

$$\text{זיהוי נכון} = \frac{\text{מספר המקטעים שהזיהוי היה נכון}}{\text{מספר המקטעים הכולל}} * 100\% \qquad (4.1)$$

4.1 חקר ביצועים של זיהוי הדובר בסביבה שאינה רועשת

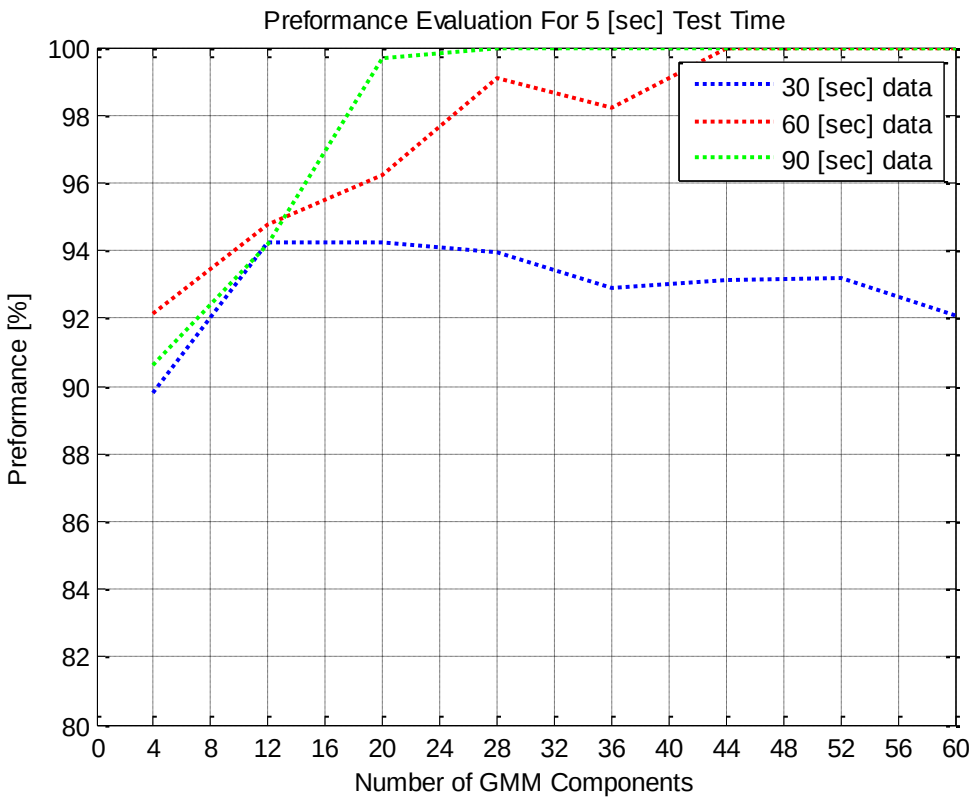
על מנת להתחיל לחקור את רמת אחוזי ההצלחה של האלגוריתם שמימשנו, תחילה, השווינו את אחוזי ההצלה שהתקבלו ב-[3] למול התוצאות שהתקבלו בסימולציות שערכנו:



איור 4.1: אחוזי ההצלחה של זיהוי דובר כפונקציה של מספר המרכיבים הגאוסיאניים.

בגרף זה בוצעו מדידות רמת הביצועים של האלגוריתם עבור מאגר של 30 דוברים, מחציתם גברים ומחציתם נשים, דוברי השפות העברית והאנגלית, שאומנו על ידי האלגוריתם לפרק זמן של 60 שניות, ביחס לכמות המרכיבים הגאוסייאנים שערכם נע בין 4 לבין 60 באינטרוול של 4 מרכיבים. כפי שניתן לראות אכן מתקיימות מגמות זהות בין התוצאות שהתקבלו ב- [3] לבין התוצאות שהתקבלו בסימולציה, דהיינו ככל שזמן הבדיקה עלה, כך עלו אחוזי ההצלחה של הזיהוי. ראוי לציין שעבור דוברים שונים במאגר התקבלו תוצאות שונות, לפיכך מדידות אלו חזרו על עצמן עבור דוברים שונים ולבסוף חושב ממוצע התוצאות שהתקבלו.

בשלב הבא נבדקו אחוזי ההצלחה של האלגוריתם כתלות בזמן אימון ביחס לכמות המרכיבים הגאוסייאנים עבור זמן בדיקה קבוע של 5 שניות, להלן התוצאות שהתקבלו:



איור 4.2: אחוזי הצלחת הזיהוי כפונקציה של מספר המרכיבים הגאוסיאניים עבור זמן אימון משתנה.

גם תוצאות אלו מתאימות לתוצאות שהתקבלו ב- [3], כפי שציפינו, המסקנה שעולה מהתוצאות שהתקבלו הינה שבעבור זמני אימון ארוכים יותר אחוזי ההצלחה הולכים וגדלים.

Amount of Training Speech	Model Order	Test Length		
		1 [sec]	5 [sec]	10 [sec]
30 [sec]	M=8	69.82%	96.08%	96.49%
	M=16	72.41%	93.83%	98.64%
	M=32	71.55%	92.59%	100%
60 [sec]	M=8	74.13%	97.41%	100%
	M=16	74.13%	96.08%	100%
	M=32	76.72%	99.23%	100%
90 [sec]	M=8	75.86%	95.65%	99.13%
	M=16	78.44%	96.08%	100%
	M=32	74.13%	97.13%	100%

טבלה 4.1: אחוזי הצלחת זיהוי הדובר עבור ערכים שונים של זמן אימון, בדיקה וכמות המרכיבים הגאוסיאניים.

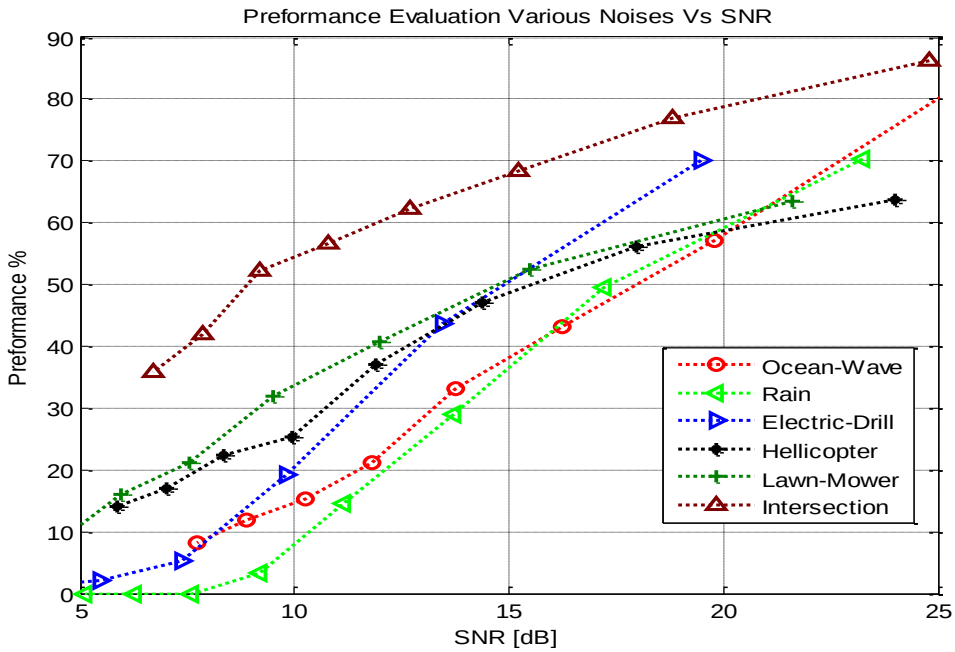
מתוצאות אלו ניתן להבחין בעליה חדה ברמת אחוזי ההצלחה של זיהוי הדוברים החל מ-4 ועד ל-8 מרכיבים גאוסיאניים, ומגמת התייצבות לאחר 16 מרכיבים גאוסיאניים, מה שמעיד שקיים ערך סף תחתון לכמות התערובות הנחוצות לצורך ביצוע זיהוי יעיל. ערך זה, כפי שעולה ממצאי הסימולציה נוטה להיות סביב ה-16 מרכיבים, והחל מערך זה, תוצאות אחוזי ההצלחה של זיהוי הדוברים נוטה להיות זהה עבור זמני בדיקה של 5 ו-10 שניות, בניגוד לזמן הבדיקה של השנייה, בה אחוזי ההצלחה של זיהוי הדובר משתפרים במגמה מסוימת.

תוצאות אלו מעידות כיצד הוספה של מרכיבים גאוסיאניים, שמעניקה מודל הסתברותי יותר עשיר במאפיינים האקוסטיים של הדובר תגרום לאחוזי הצלחה גבוהים יותר עבור זמני בדיקה קצרים. גורם חיוני נוסף אשר יקבע את רמת אחוזי ההצלחה של מלאכת הזיהוי הינו אוכלוסיית מאגר הדוברים, ככל שזו תגדל, תגדל בהתאם כמות הדוברים שהמערכת תיאלץ להבחין ביניהם ולפיכך תגדל ההסתברות לזיהוי לא נכון. הדמיון בין אוכלוסיית הדוברים גם היא צריכה להילקח בחשבון, מכיוון שדוברים בעלי מאפיינים שונים של קול, למשל אוכלוסייה המחולקת לגברים ונשים, בדרך כלל תניב רמת ביצועים גבוהה יותר מבחינת אחוזי ההצלחה של הזיהוי מאשר אוכלוסייה הומוגנית, למשל אוכלוסייה של גברים בלבד.

4.2 חקר ביצועים של זיהוי דובר בסביבה רועשת

בשלב זה נחקור את אחוזי ההצלחה של אלגוריתם לזיהוי דובר עם רעשי רקע אקוסטיים, לרבות: גשם, פרופלור של מסוק, מדורה, גלים בים, תנועה סואנת של רכבים, מכסחת דשא ומקדחה, עבור המאגר הזהה למאגר שאומן קודם לכן. לצורך ביצוע מדידות אלו, אנו נשתמש בערך זמן בדיקה קבוע של 10 שניות עבור מאגר דוברים שאומן 60 שניות בעל 20 מרכיבים גאוסיאניים. כמות מקדמי ה-MFCC נשארה זהה כמקודם, דהיינו 12 המקדמים הראשונים. ערכים אלו נבחרו על סמך ממצאי חקר הביצועים שבוצע קודם לכן עבורם נצפו אחוזי הצלחה גבוהים. ראוי לציין שבעבור ערכים גבוהים מאלו, אחוזי ההצלחה של זיהוי הדובר משתפרים באחוזים בודדים אולם גורמים לזמני החישוב של האלגוריתם לגדול באופן משמעותי.

תחילה, נציג את תוצאות אחוזי ההצלחה של זיהוי הדובר תחת תנאי יחס אות לרעש משתנה עבור הרעשים האקוסטיים שצוינו קודם לכן. להלן תוצאות אחוזי ההצלחה שנמדדו:



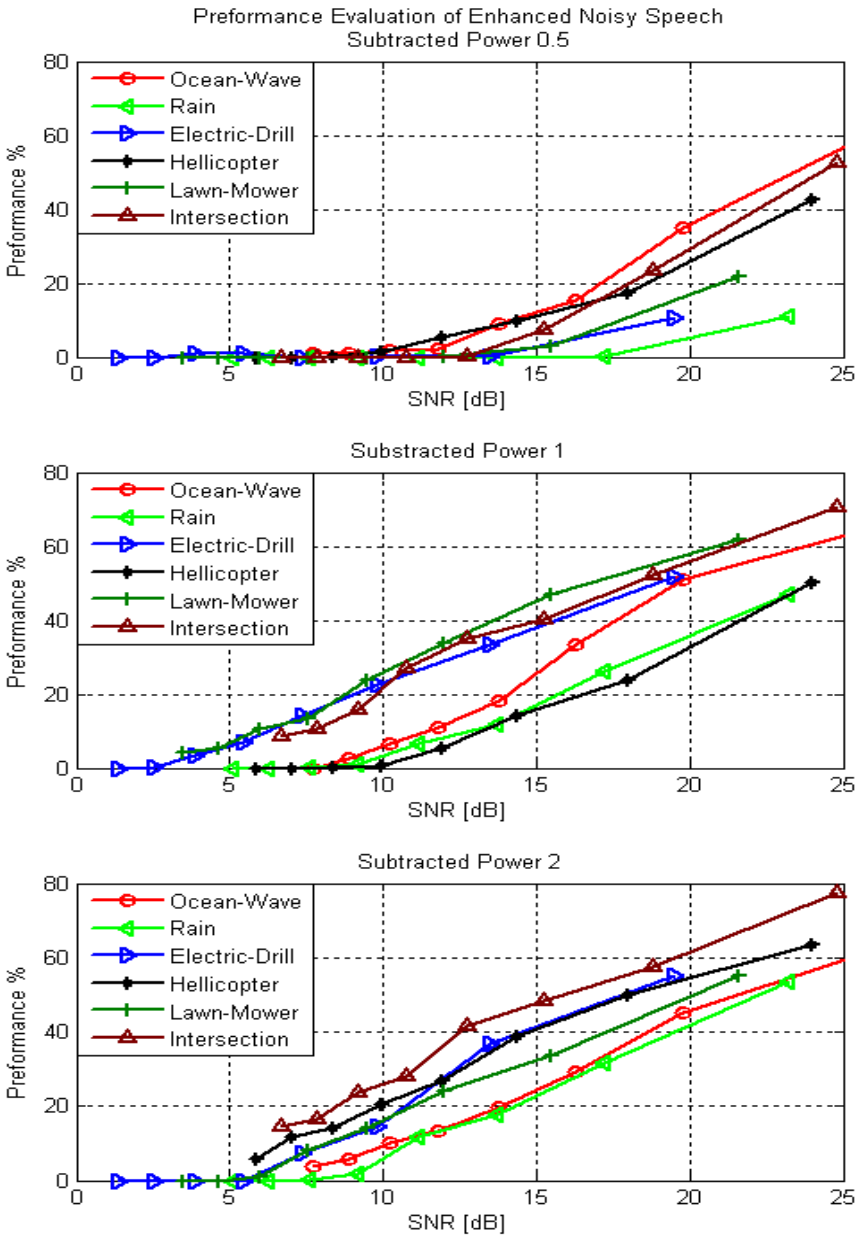
איור 4.3: אחוזי הצלחת הזיהוי כפונקציה של יחס אות לרעש עבור רעשי רקע אקוסטיים שונים.

כפי שניתן לראות אחוזי ההצלחה של זיהוי הדובר הולכים וקטנים באופן די לינארי ככל שיחס האות לרעש הולך וקטן. מאגר הרעשים עבורם בוצעו המדידות בעלי אופי שונה בזמן ובתדר, לפיכך, רעשים אלו, תחת אותם ערכי יחס האות לרעש, ישפיעו בצורה שונה על רמת אחוזי ההצלחה של אלגוריתם זיהוי הדובר. ראוי לציין שרמת אחוזי ההצלחה של זיהוי הדובר עבור אות הדיבור ללא הרעש המתווסף הינה 98.002% כמו כן בשלב ביצוע המדידות הללו דאגנו לשמור על קנה מידה מנורמל עבור עוצמות הן של אות הדיבור והן של הרעש האקוסטי המתווסף. כעת כשבידנו מדד הביצועים עבור אות דיבור עם רעשי רקע אקוסטיים מתווספים, נהיה מעוניינים לחקור עד כמה, אם בכלל, ישתפרו אחוזי ההצלחה של זיהוי הדובר באמצעות השבחת אות

הדיבור הרועש על ידי שיטת ההפחתה הספקטראלית שהורחבה בפרק 2.1.2. על מנת לבחון שיטה זו, נבצע את אותן מדידות עבור מספר ערכים שונים של חזקות [10], [11].

4.3 חקר ביצועים של זיהוי דובר לאחר השבחה ספקטראלית

לצורך ביצוע מדידות אחוזי ההצלחה של זיהוי הדובר לאחרת השבחת אות הדיבור נבחרו מספר רעשים עבורם התקבלו אחוזי הצלחת זיהוי נמוכים ללא השבחת אות הדיבור. לצורך ביצוע מדידות אלו כמקודם, אנו נשתמש בערך זמן בדיקה קבוע של 10 שניות עבור מאגר דוברים שאומן 60 שניות בעל 20 מרכיבים גאוסיאניים. להלן תוצאות המדידה של אחוזי ההצלחה של זיהוי הדובר לאחר השבחת אות הדיבור הרועש :



איור 4.4: אחוזי הצלחת הזיהוי כפונקציה של יחס אות לרעש עבור אות דיבור רועש לאחר ההשבחה הספקטראלית.

תוצאות המדידה לעיל מראות כי רמת אחוזי ההצלחה של הגילוי לאחר ההשבחה הספקטראלית, תחת יחס אות לרעש זהה, פוחתות במידה ניכרת ביחס לרמת אחוזי ההצלחה שנמדדו קודם לכן, דהיינו שיטת ההפחתה הספקטראלית של הרעש פוגעת בביצועים של אלגוריתם זיהוי הדובר. נציין ונאמר שרמת אחוזי ההצלחה הגבוהה ביותר עבור חלק ניכר מהרעשים עבורם נמדדו אחוזי הצלחת הגילוי מתקבלת עבור החזקה ה-2 עבורה מתקבלת ההפחתה המשמעותית ביותר, בהנחה והרעש דומיננטי מספיק ואות הדיבור איכותי. טבלה 4.2 ממחישה כמותית את ההבדלים ברמת אחוזי ההצלחה של זיהוי הדובר עם ובהיעדר השבחת האות הרועש.

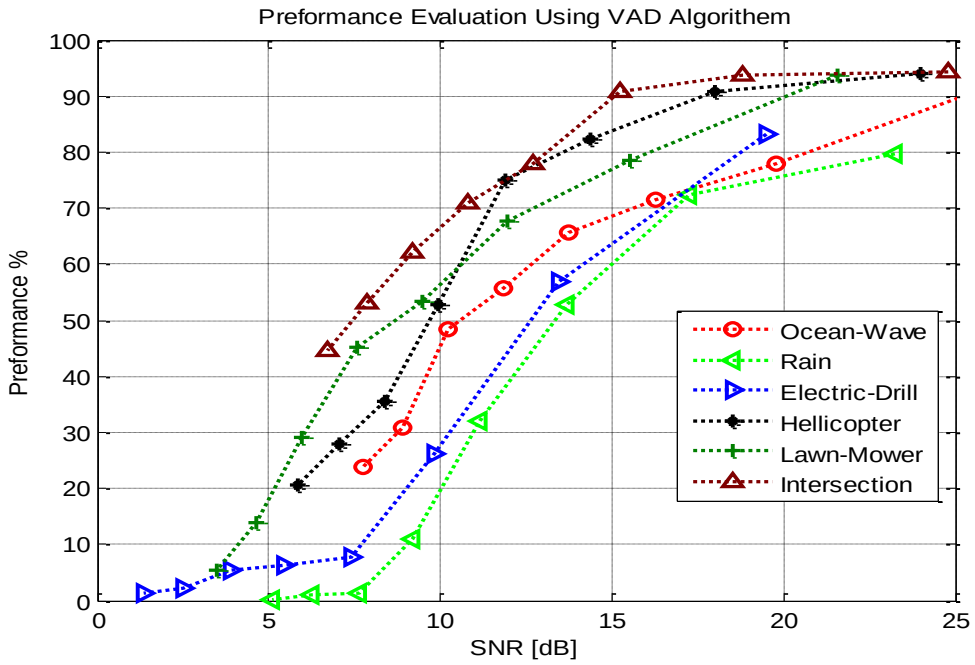
Noise	Signal to Noise Ratio [dB]	Noisy Speech Detection Evaluation	Enhanced Noisy Signal With Various Subtracted power		
			0.5	1	2.0
Ocean-Wave	11.82	21.25%	2.06%	10.8%	13.24%
Rain	11.13	14.59%	0%	6.56%	11.69%
Electric-Drill	11.63	31.48%	0.02%	28.03%	25.62%
Helicopter	11.88	36.96%	5.27%	5.29%	26.84%
Intersection	11.74	59.54	0.1%	30.87%	34.71%
Lawn-Mower	11.98	40.84%	0.2%	33.78%	24.13%

טבלה 4.2: השוואה בין אחוזי הצלחת זיהוי הדובר עם ובהיעדר השבחת אות הדיבור עבור 3 רעשים תחת יחס אות לרעש זהה.

אלגוריתם השבחת אות הדיבור באמצעות שיטת ההפחתה הספקטראלית אמנם מותאם לשיפור מובנות המלל שנאמר על ידי הדובר עבור המאזין, אולם לא מעניק רמת ביצועים גבוה עבור יישומי זיהוי דובר אוטומטי.

4.4 חקר ביצועים של זיהוי דובר עבור אלגוריתם ה-VAD

בשלב זה התעניינו לחקור מהם אחוזי הביצועים אשר יתקבלו אילו נעזר באלגוריתם זיהוי הדיבור "Voice Activity Detector (VAD)" לזיהוי מקטעים קוליים, ובעבור מקטעים קוליים אלו נחזור על המדידות שבוצעו לעיל עבור הרעשים האקוסטיים בתת פרק 4.3 עם אותו מאגר הדוברים. להלן תוצאות המדידה שהתקבלו:



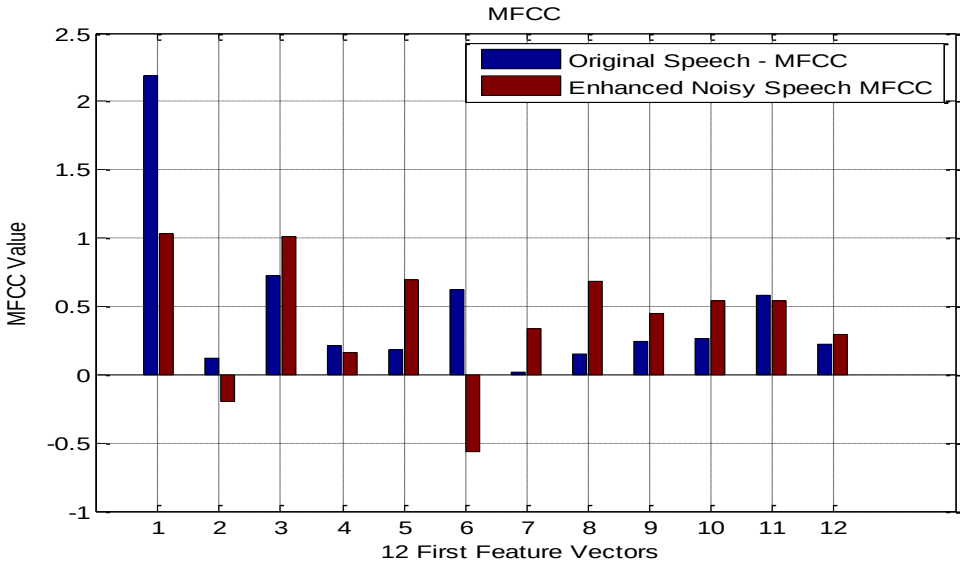
איור 4.5: אחוזי הצלחת הזיהוי כפונקציה של יחס אות לרעש עבור מקטעי אות הדיבור הרועש הקולי.

מתוצאות אלו ניתן להבחין במגמת עליה של עשרות אחוזים ברמת ההצלחה של זיהוי הדוברים ביחס לתוצאות שנמדדו באמצעות שיטת ההשבחה הספקטראלית שנצפו קודם לכן. האותות הקוליים לחוד אכן מכילים מידע ספקטראלי עשיר של המאפיינים האקוסטיים של הדובר וזאת למרות נוכחות רעשי הרקע המתווספים אל אות הדיבור המקורי. יחד עם זאת, בדומה למגמת הירידה של רמת אחוזי ההצלחה כפונקציה של ערך יחס אות לרעש יורד שנצפתה קודם לכן, גם שיטה זו לא מעניקה חסינות גבוהה למדי עבור ערך יחס אות לרעש הולך וקטן. טבלה 4.3 ממחישה כמותית את ההבדלים בין אחוזי הצלחת זיהוי הדובר עם ובהיעדר אלגוריתם ה-VAD.

Noise	Signal to Noise Ratio [dB]	Noisy Speech Detection Evaluation	VAD Algorithm Detection Evaluation
Ocean-Wave	11.82	21.25%	55.56%
Rain	11.13	14.59%	32.12%
Electric-Drill	11.63	31.48%	41.59%
Helicopter	11.88	36.96%	74.93%
Intersection	11.74	59.54	74.34%
Lawn-Mower	11.98	40.84%	67.81%

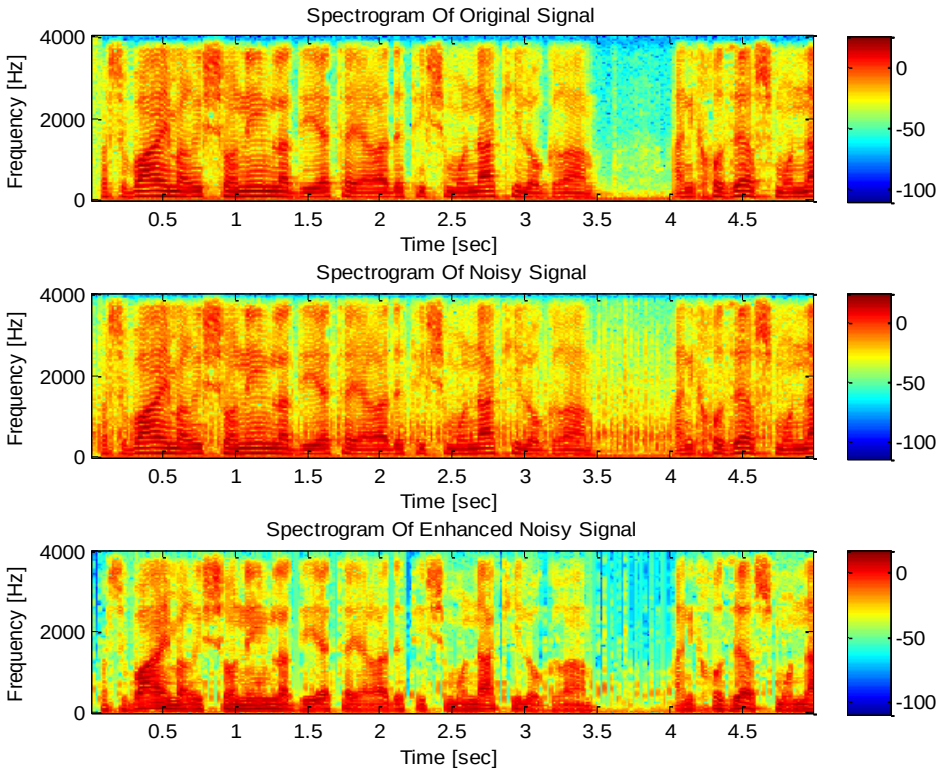
טבלה 4.3: השוואה בין אחוזי הצלחת זיהוי הדובר עם ובהיעדר אותות דיבור א-קוליים עבור שלושה רעשים תחת יחס אות לרעש זהה.

בשלב זה מתעוררת השאלה מדוע בכל זאת נפגעים אחוזי ההצלחה של אלגוריתם זיהוי הדובר לאחר ההשבחה הספקטראלית, שהרי מטרת שיטה זו היא להפחית מערך הרעש המתווסף ולפיכך להפיק אות דיבור איכותי ומובן, כאשר בפועל אכן פעולת ההשבחה תרמה להפחתה, ובמקרים מסוימים אף לסינון הרעשים המתווספים לאות הדיבור. כדי לחקור בכל זאת כיצד משפיע פעולת ההפחתה הספקטראלית של הרעש מאות הדיבור, מדדנו את ערכי 12 הערכים הראשונים של ווקטור מיצוי המאפיינים MFCC לפני ואחרי ההשבחה וקיבלנו את התוצאה הבאה:



איור 4.6: 12 המקדמים הראשונים של ווקטור מיצוי המאפיינים עם ובהיעדר השבחת אות הדיבור, ההשבחה בוצעה עבור אות עם תוספת רעש.

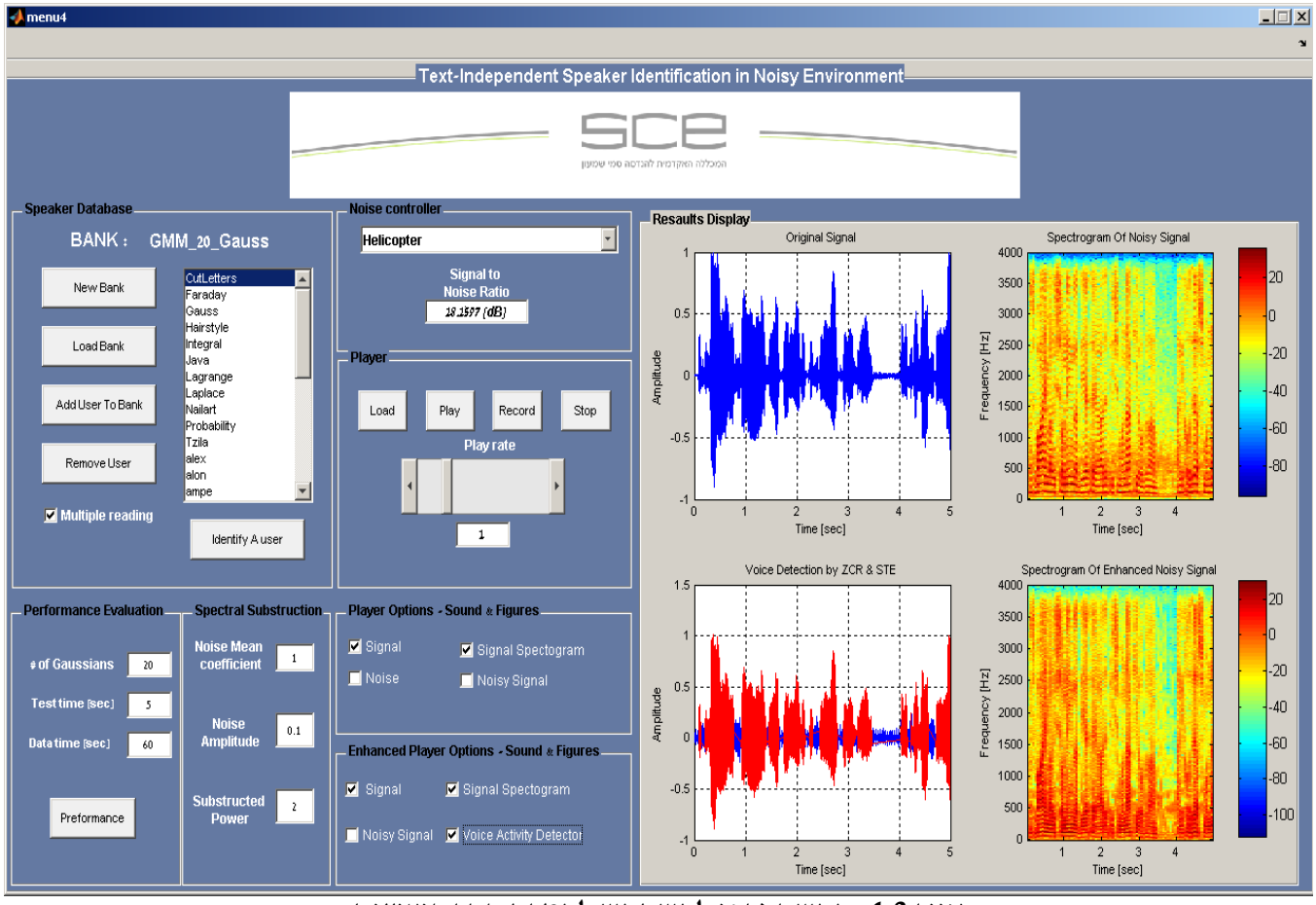
מיד ניתן להבחין בשוני בין מאפייני המעטפת הספקטראלית שהתקבלה בשני המקרים, דהיינו ווקטור המאפיינים האקוסטיים של הדובר לאחר ההשבחה והפחתה של הרעש האקוסטי המתווסף יהיו שונים למדי ביחס לווקטור מיצוי המאפיינים של האות ללא הרעש, דבר אשר עלול ליצור חריגה משמעותית בשלב אומדן זיהוי הדובר ביחס למודל הסטטיסטי שהתקבל בשלב האימון שכאמור גם הוא נאמד על סמך מיצוי המאפיינים האקוסטיים הללו. ראוי לציין שעלול להיווצר מצב בו ההפחתה תגרום להופעת "רעש מוזיקלי" [4] שאכן נצפה במהלך המדידות הללו, רעש זה נגרם כתוצאה מהפחתת יתר של תוחלת הרעש $E[|D_w(\omega)|^\alpha]$ מהאות, רעש זה יגרום לשינויים ועיוותים נוספים בערכים של ווקטור מיצוי המאפיינים כיוון שאלו כלל אינם מאפיינים את מאפייני הקול של הדובר. נציין שקיים כלי שימושי לצורך ייצוג אות הדיבור במונחים של האנרגיה המוכלת באות והוא הספקטוגרמה. הספקטוגרמה הינו ייצוג תלת ממדי של המרכיבים הספקטראליים של אות הדיבור כאשר הציר האופקי בייצוג זה מציין את מישור הזמן והציר האנכי מציין את מישור התדר, כך, עבור כל מקטע במונחים של הזמן, רמת האנרגיה של כל מרכיב ספקטראלי מקבלת גוון בהירות בהתאמה. ככל שרמת האנרגיה תהיה גבוה יותר, כך תהיה רמת הבהירות גבוהה יותר. איור 2.2 ממחיש את אופן פעולת האלגוריתם של ההשבחה הספקטראלית המתואר לעיל, באיור זה ניתן לראות את השפעת ההפחתה של אנרגיית המרכיבים הספקטראליים של הרעש המתווסף לאות הדיבור ושחזורו בקירוב לצורתו המקורית.



איור 4.7: ספקטוגרמה של אות דיבור, אות הדיבור עם רעש רקע מתווסף ומתחתיו האות הרועש לאחר ההשבחה.

4.5 ממשק גרפי למשתמש למטרת חקר ביצועים

להלן ממשק גרפי למשתמש שעיצבנו באמצעות תוכנת ה-MATLAB:

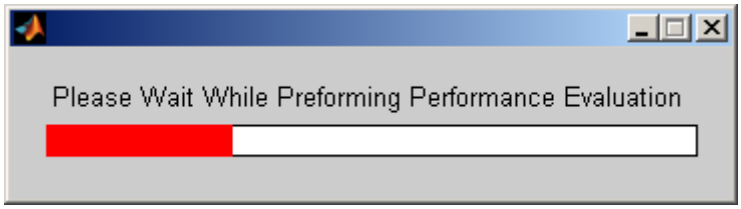


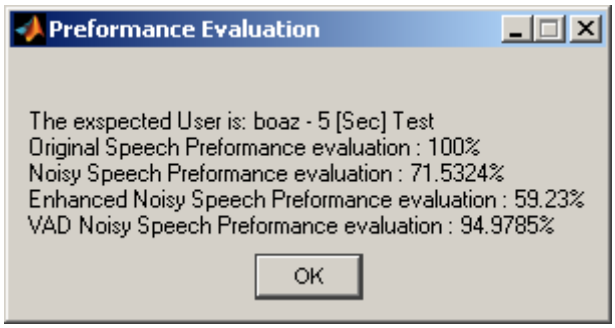
איור 4.8: ממשק גרפי למשתמש למטרת חקר ביצועים.

הממשק מאפשר הקלטה והשמעה של אותות דיבור, יתרה על כך הממשק מאפשר לטעון מגוון רחב של רעשי רקע אקוסטיים לרבות: מדחף מסוק, גשם, רוח וכו'. כמו כן הממשק מעניק יכולת צפייה באותות הדיבור המוקלטים עם וללא תוספת הרעש ואף מאפשר סימון המקטעים הקוליים באמצעות אלגוריתם ה-VAD. בנוסף ניתנת היכולת לצפות בספקטוגרמה המתקבלת בעבור האות.

הממשק מאפשר לבצע השבחה עבור אות הדיבור עם ובהיעדר הרעש, כמו כן קיימות מספר דרגות חופש אשר המשתמש יכול להגדיר בטרם ביצוע ההשבחה, לדוגמא הגדרת החזקה α (Subtracted Power), או עוצמת הרעש בקנה מידה מנורמל (Noise Amplitude). גם בעבור ההשבחה ניתנת יכולת ההצגה גרפית של האות עם ובהיעדר הרעש, כמו גם הספקטוגרמה בהתאמה.

הממשק מאפשר טעינה וביצוע הליך אימון אוטומטי עבור הדוברים המתווספים למאגר כפונקציה של הפרמטרים הניתנים להגדרה על ידי המשתמש, מבין פרמטרים הללו: זמן האימון ומספר המרכיבים הגאוסיאניים. לאחר הגדרת מאגר המשתמשים, הממשק מאפשר לערוך סדרה של מדדי ביצועים של אחוז הצלחת זיהוי הדובר עבור אות הדיבור הנבדק, דהיינו הדובר הנבחן, תחת אילוצים שונים כלומר עם ובהיעדר השבחת אות הדיבור ובעבור קטימה של אותות קוליים בלבד באמצעות אלגוריתם ה-VAD.





איור 4.9: תוצאות מדדי הביצועים המוצגות על ידי הממשק הגרפי למשתמש.

5 סיכום

לאור כל הממצאים לעיל אנו אכן מקבלים מגמות זהות בין [3] לבין אחוזי ההצלחה של האלגוריתם לזיהוי דובר מקבוצת דוברים ללא מעורבות של רעש אקוסטי מתווסף כפי שניתן לראות בטבלה 4.1, אף על פי כך, בניגוד לצפיותינו, אלגוריתם השבחת אות הדיבור באמצעות שיטת ההפחתה הספקטראלית איינה תורמת לאחוזי זיהוי גבוהים יותר של אות הדיבור עם הרעשים האקוסטיים המתווספים אלה גורמת לרמת ביצועים נמוכים יותר מאלו שהתקבלו ללא השבחת האות כפי שניתן לראות בטבלה 4.2 הסיבה לירידה בביצועים נובעת כתוצאה משינוי ווקטור מיצוי המאפיינים הספקטראליים של האות הנגרמת כתוצאה מההשבחה, למרות שבפועל האלגוריתם מעניק שיחזור איכותי ומובן בהשמעתו של אות הדיבור לאחר הפחתת הרעש המתווסף. דרך נוספת שהצענו לצורך שיפור הביצועים של אלגוריתם זיהוי הדובר היא מדידת אחוזי ההצלחה עבור קטימה של המרכיבים הא-קוליים של אות הדיבור בתוספת רעש הרקע, דהיינו מדידת אחוזי ההצלחה עבור המקטעים הקוליים בלבד באמצעות אלגוריתם ה-VAD, כפי שניתן לראות בטבלה 4.3 אכן התקבל שיפור ניכר של מספר עשרות אחוזים ברמת אחוזי ההצלחה. אחת הצעותינו לצורך שיפור הביצועים של האלגוריתם הוא על ידי שימוש במסננים אדפטיביים, כדוגמת מסנן Wiener, Comb, Kalman ו- Kernel לצורך הפחתת רעשי הרקע האקוסטיים ושיפור איכות אות הדיבור. הממשק הגרפי שהרכבנו יכול לשמש כאמצעי הדרכה שימושי לצרכי קורסים העוסקים בעיבוד אותות דיבור ספרתי.

6 מקורות

[1] L.Rabiner, "Fundamentals of speech recognition", Prentice-Hall, 1993. Ch. 3 pp. 69-121.

[2] Vaseghi S., "Advanced Digital Signal Processing and Noise Reduction," Brunel University, UK, 2008 4th edition.

[3] Reynolds D., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE, Jan. 1995. Vol. 3 pp. 72-83.

[4] Lim J., "Enhancement and Bandwidth Compression of Noisy Speech," IEEE, Dec. 1979. Vol. 67 pp. 1590-1593

[5] Xuedong H., Alex A. and Hsia-Wuen H., "Spoken Language Processing", Prentice-Hall, 2001.

[6] Campbell and Joseph P., "Speaker Recognition : A Tutorial". : IEEE, September 1997, Proceeding, Vol. 85, pp. 1437-1462.

[7] Judith Markowitz and Bill Scholz ., "Advances in Speech Recognition Mobile Environments, Call Centers and Clinics ". Springer 2010.

[8] Gerard Chollet, Maria Gabriella Di Benedetto, Anna Esposito and Maria Marinaro,."Speech Processing, Recognition and Artificial Neural Networks" Springer 1999.

[9] Homayoon Beigi, "Fundamentals of Speaker Recognition " Springer 2011.

[10] M.R Weiss, E. Aschkenasy, and T. W. Parsons, "Study and development of the INTEL technique for improving Speech inteligibility" Nicoloet Scientific Corp., Final Rep. NSC-FR/4023, Dec. 1974.

[11] M.R Weiss et al., "Processing speech signals to attenuate interference," presented at IEEE symp. Speech Recognition, Apr. 1974.

7 נספחים

7.1 נספח A

1. רעשים של רכיבים אלקטרוניים – אלו רעשים הכוללים:

1.1 רעשים טרמיים הנוצרים כתוצאה מתנועה אקראית של חלקיקים הטעונים על ידי אנרגיה טרמית בתוך המוליך החשמלי. רעשים אלו קיימים בכל סוגי המוליכים אפילו ללא ערעור חשמלי.

1.2 רעשי שוט הנוצרים כתוצאה מתנודות אקראיות של הזרם החשמלי במוליך. רעש זה נגרם למעשה מעצם העובדה שהזרם נשא על ידי מספר דיסקרטי של מטענים, ובעבור כל אחד מהם זמני הגעה ותנודות אקראיות שונים.

1.3 רעשי "Flicker" עבורם הספקטרום משתנה ביחס הפוך לתדר, הסיבה לכך נובעת כתוצאה מזיהומים במוליך או כתוצאה מגנרציה ורקומבינציה של רעש בטרנזיסטור הנובע ממתח הייחוס.

1.4 רעשי "Burst" המתוארים בתור עליה פתאומית של מספר מאות מילי-וולט, בזמנים ומשכים אקראיים, רעשים אלו מתרחשים ברכיבי מוליכים למחצה, בעיקר בתחום התדרים הנמוך, דהיינו פחות מ- 100 Hz ומאופיינים במערכות שמע בתור רעש פופקורן כצליל הפיצפוצים.

2. רעשים אלקטרומגנטיים: אלו מצויים לאורך כל תחום התדרים ובעיקר בתחום תדרי הרדיו, רעשים אלו מורכבים מצרוף של רעשים הנובעים מהרכיבים החשמליים ורעשי סביבה הנובעים מהאטמוספרה והקרינה הקוסמית.

3. רעשים אלקטרוסטטיים: אלו נוצרים כתוצאה מנוכחות מפל מתח ללא זרם חשמלי, תאורת פלורסנט הינה אחת מהמקורות הנפוצים ליצירת רעש זה.

4. עיוותי הערוץ, "Multipath", הד ו-"Fading": אלו נוצרים כתוצאה מהאי אידיאליות של ערוצי התקשרות. ערוצי הרדיו המצויים בתחומי ה-[GHz] כמו למשל מערכות ההפעלה של הסלולר למיניהם, רגישות במיוחד למאפייני ההתפשטות האות בסביבת הערוץ, בו האות מעוצב, מושהה ומתעוות כתוצאה מהתגובה לתדר המאפיינת את הערוץ. שני העיוותים הבולטים מבין העיוותים הנגרמים על ידי הערוץ אלה עיוותי הפאזה ועיוותי האמליטודה. נוסף על כך האות המשודר מתחנת המקור עלול לעבור מספר מסלולים עד הגעתו למקלט כך שהאות המשודר עלול להגיע במספר גרסאות, דהיינו בהפרשי פאזה ואפליטודה שונים.

5. הפרעה בין ערוצים: הפרעה שנוצרת כתוצאה משני מקורות שידור שונים המשדרים בתדרים זהים, דבר היוצר את תופעת הערב-דיבור. סיבות נוספות לערב דיבור הינם תנאים אטמוספריים מסוימים אשר עלולים לגרום להחזרות האותות המשודרים מהטרוטוספרה או כתוצאה מצפיפות יתר בספקטרום הרדיו.

6.דגימות חסרות: מקטעים של האות עלולים להיות חסרים כתוצאה ממספר סיבות וביניהם עליה גבוהה ופתאומית של רעש, הצפת האינפורמציה באות או כתוצאה מאיבוד חבילות המידע המשודרות דרך ערוץ התקשורת.

7. רעש הנגרם כתוצאה מעיבוד האות: נגרם על ידי מערכות ההמרה של האות האנלוגי לדיגיטלי בשלב עיבוד האות, למעשה זהו רעש הכימות במערכות קידוד דיגיטליות עבור אות דיבור או תמונה.

סוגים שונים של צורות רעש

בהתבסס על ספקטרום התדרים או על המאפיינים שלו בזמן, ניתן לסווג את תהליך הרעש במספר הקטגוריות הבאות:

1. רעש לבן: הינו רעש אקראי טהור, פונקציית האוטוקורולציה המתקבלת עבור רעש זה הינו הלם וספקטרום הספק קבוע, שבאופן תיאורטי מכיל את כל ספקטרום התדרים.

2. רעש לבן מוגבל פס: זהו רעש בעל ספקטרום הספק קבוע ורוחב פס המוגבל לתחום התדרים של הציוד האלקטרוני. פונקציית האוטוקורולציה המתקבלת עבור רעש זה הינו פונקציית ה-"Sinc".

3. רעש צר סרט: זהו תהליך רעש בעל רוחב סרט צר בתחום ה-50/60 Hz הנגרם על ידי ספק החשמל.

4. רעש צבעוני: זהו כל רעש שאיננו רעש לבן, או רעש רחב סרט שספקטרום ההספק עבורו איננו קבוע.

5. רעש אימפולסיבי: רעש זה מכיל פולסים לאורך פרקי זמן קצרים בעלי אמפליטודות, וזמני הופעה אקראיים.

6. רעש פולסים חולפים: רעשים אלו מכילים פולסים לאורך פרקי זמן יחסית ארוכים מבין הרעשים הללו מצויים רעשי "הקליק" שנוצרים בזמן השידור, או רעשי ה-"Burst" שצויינו קודם לכן.

7.2 נספח B

7.2.1 פונקציית חלון

פונקציות החלון אלו אותות מרוכזים ומוגבלים בזמן, מבין הנפוצים ביניהם בתחום עיבוד האותות אלו פונקציות חלון מסוג משולש, מרובע, Kaiser, Barlett, Hanning, Haming. פונקציות אלו מצויות בתחום התדרים הנמוך.

7.2.2 חלון מרובע:

הגדרה:

$$h_\pi[n] = u[n] - u[n-N] \tag{7.1}$$

התמרת Z של החלון:

$$H_\pi(z) = \sum_{n=0}^{N-1} z^{-n} \tag{7.2}$$

ע"י הכפלה משני צדי משוואה (2) ב- z^{-1} נקבל:

$$Z^{-1}\, H_\pi(z) = \sum_{n-1}^{N} z^{-n} = H_\pi(z) - 1 + z^{-N} \tag{7.3}$$

ניתן לתאר את הטור הגאומטרי שהתקבל לעיל ע"י:

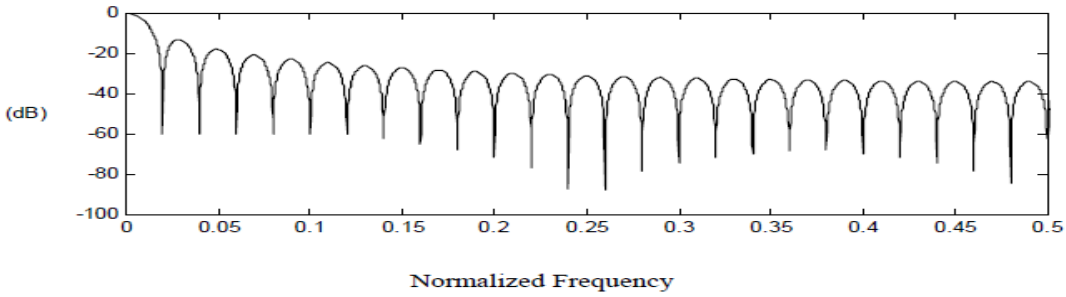
$$H_\pi(z) = \frac{1 - z^{-N}}{1 - z^{-1}} \tag{7.4}$$

התמרת פורייה של חלון ריבועי המוגדר ע"י משוואה מספר 6.4:

$$H_\pi\left(e^{j\omega}\right) = \frac{1-e^{-j\omega N}}{1-e^{-j\omega}} = \frac{\left(e^{\frac{j\omega N}{2}} - e^{-\frac{j\omega N}{2}}\right)e^{-\frac{j\omega N}{2}}}{\left(e^{\frac{j\omega}{2}} - e^{-\frac{j\omega}{2}}\right)e^{-\frac{j\omega}{2}}} \tag{7.5}$$

$$= \frac{\sin(\frac{\omega N}{2})}{sin(\frac{\omega}{2})} e^{-j\omega(N-1)/2} = A(\omega)e^{-j\omega(N-1)/2}$$

כאשר $A(\omega)$ הוא ממשי וזוגי.



איור 7.1: תגובת התדר של חלון מרובע עבור N=50 [5].

7.2.3 חלון Hamming ו-Hanning:

הגדרה:

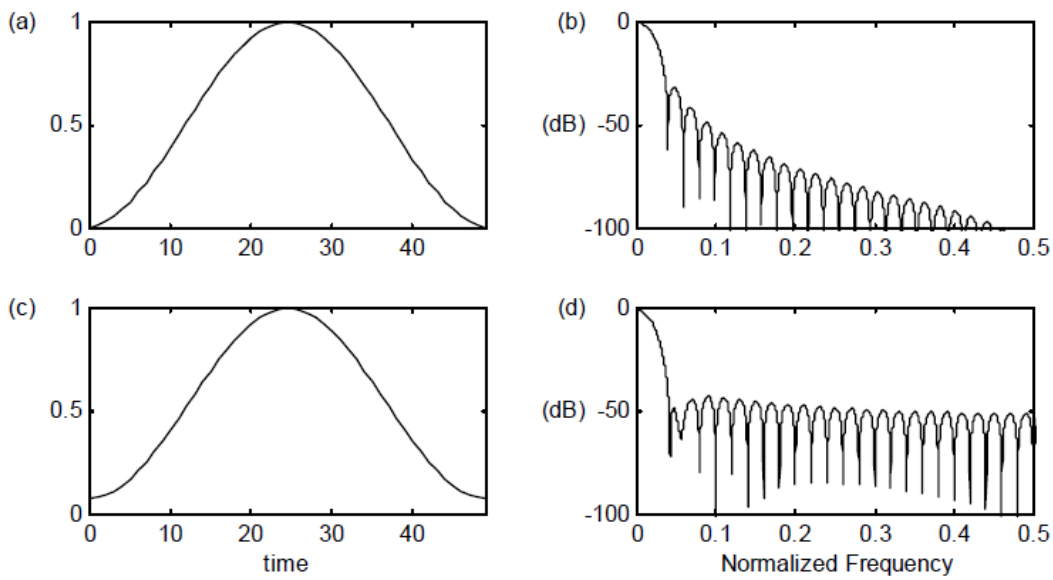
$$h_h[n] = \begin{cases} (1-\alpha) - \alpha\cos\left(\frac{2\pi n}{N}\right), 0 \le n \le N \\ 0, \text{אחרת} \end{cases} \qquad (7.6)$$

התמרת הפורייה של החלון:

$$H_h(e^{j\omega}) = (1-\alpha)H_\pi(e^{j\omega}) - \left(\frac{\alpha}{2}\right)H_\pi\left(e^{\frac{j(\omega-2\pi)}{N}}\right) - \left(\frac{\alpha}{2}\right)H_\pi\left(e^{j(\omega+2\pi)/N}\right) \qquad (7.7)$$

כאשר הערך של $\alpha = 0.5$ החלון יקרא חלון Hanning, ובמקרה בו $\alpha = 0.46$ החלון יקרא חלון Hamming. ראוי לציין שרוחב האונה הראשית כמעט וכפולה מזו של החלון המרובע אולם ההנחתה של אונות הצד המתקבלת באמצעות שימוש בחלונות אלו הרבה יותר משמעותית ביחס לחלון המרובע.

אונות הצד המשנית המתקבלת עבור חלון Hanning הינה ברמה של 31dB מתחת לאונה הראשית לעומת חלון Hamming בו רמת אונת הצד המשנית הינה 44dB אולם הניחות של אונות הצד של חלון Hanning דועכות יותר מהר עם עליית התדר, במקרה של חלון Hamming רמת אונות הצד נשארת קבועה לאורך כל התדרים.



איור 7.2 תגובת התדר של חלון Hanning ו-Hamming עבור N=50 [5].

7.3 נספח C

7.3.1 השבחת אות דיבור

אות הדיבור הרועש במישור הזמן מוגדר באופן הבא:

$$y(n) = s(n) + d(n) \tag{7.8}$$

לפיכך צפיפות ההספק הספקטראלית עבור האות הרועש תוגדר באופן הבא:

$$P_y(\omega) = P_s(\omega) + P_d(\omega) \tag{7.9}$$

אות הדיבור איננו סטציונארי, אולם על ידי הכפלה בפונקציית חלון המוגבל למסגרות זמן קצרות מאפשרת לשמור על התכונות הסטטיסטיות של האות ועבורם התמרת הפורייה הבאה:

$$|Y_w(\omega)|^2 = |S_w(\omega)|^2 + |D_w(\omega)|^2 + S_w(\omega) \cdot D_w^*(\omega) + S_w^*(\omega) \cdot D_w(\omega) \tag{7.10}$$

כאשר $D_w^*(\omega)$ ו $S_w^*(\omega)$ מייצגים את הצימוד המרוכב של $D_w(\omega)$ ו- $S_w(\omega)$. הפונקציה $|S_w(\omega)|^2$ באה לידי ביטוי כאנרגיית הספקטרום של אות הדיבור לפרק זמן קצר. עבור השבחה של אות דיבור המבוסס על ספקטרום האמפליטודה לפרק זמן קצר, נדרש לחלץ את האות הממוצע $|\widehat{S}_w(\omega)|^2$ ועל ידי כך ניתן יהיה לשערך את הממוצע של האות במישור הזמן.

מהאות הנצפה $y_w(n)$ ניתן לחשב באופן ישיר את $|Y_w(\omega)|^2$ כאשר הביטויים $|D_w(\omega)|^2$, $S_w(w) \cdot D_w^*(\omega)$ ו- $S_w^*(\omega) \cdot D_w(\omega)$ לא ניתנים לשיערוך מדויק לפיכך על ידי טכניקת ההפחתה הספקטראלית אלו נאמדים על ידי $E[|D_w(\omega)|^2]$, $E[S_w(\omega) \cdot D_w^*(\omega)]$ ו- $E[S_w^*(\omega) \cdot D_w(\omega)]$ כשאר אופרטור $E[\cdot]$ מציין את ממוצע האות, והרי ההנחה היא שממוצע הרעש שווה לאפס, כמו כן לא מתקיימת קורולציה בין האות $s(n)$ לבין הרעש $d(n)$ לפיכך הביטויים $E[S_w(\omega) \cdot D_w^*(\omega)]$ ו- $E[S_w^*(\omega) \cdot D_w(\omega)]$ מתאפסים, על סמך ההנחות הללו נוכל לרשום את הביטוי הבא:

$$|\widehat{S}_w(\omega)|^2 = |Y_w(\omega)|^2 - E[|D_w(\omega)|^2] \tag{7.11}$$

כאשר $E[|D_w(\omega)|^2]$ יתקבל על סמך ההנחה שתכונות הרעש ידועות או על ידי מדידת רעש הרקע המתווסף באינטרוולים בהם לא קיים דיבור [6]. אומדן ה- $|\widehat{S}_w(\omega)|^2$ יתקבל באמצעות נוסחה מספר 6 אולם אומדן זה לא יבטיח ערכים אי שליליים היות ואגף ימין של משוואה זו עלול להיות ערך שלילי. במספר מחקרים הדנים בנושא ההפחתה הספקטראלית ערכים שליליים אלו נקבעים כערכים חיוביים, ואילו במחקרים אחרים, ערכים אלו מאופסים. לבסוף על מנת לשחזר את האות לאחר ביצוע ההפחתה הספקטראלית יש לשלב בחזרה את המידע של הפאזה של האות הרועש המקורי והצגתו במישור הזמן תתקבל על ידי התמרת פורייה הפוכה באופן הבא:

$$\hat{S}_w(\omega) = \left|\hat{S}_w(\omega)\right| \cdot exp[j\measuredangle Y_w(\omega)] \tag{7.12}$$

$$\hat{s}_w(n) = F^{-1}\left[\hat{S}_w(\omega)\right] \tag{7.13}$$

7.4 נספח D

7.5 אלגוריתם "Forward":

נגדיר $a_t(i) = p(o_1, \dots, o_t, q_t = s_i|\lambda)$ כהסתברות של כל הסימבולים שנוצרו עד לפרק הזמן t וכאשר המערכת תמצא במצב s_i בזמן זה. פרמטרי ה- a יכולים להיות מחושבים על ידי התהליך האינדוקטיבי הבא:

1. אתחול: $a_1(i) = \pi_i b_i(o_1)$ כאשר המצב ההתחלתי הוא s_i והסימבול המיוצר הינו o_1.

2. אינדוקציה: $a_{t+1}(i) = b_i(o_{t+1}) \sum_{j=1}^{N} a_t(j) a_{ji}$, עבור $1 \le t < T$ כאשר עוברים ממצב s_j למצב s_i בהסתברות a_{ji} כאשר מיוצר הסימבול o_{t+1}.

3. סיום: $p(O|\lambda) = \sum_{i=0}^{N} a_t(i)$.

7.6 אלגוריתם "Backward":

ערכי ה- a אשר חושבו באמצעות אלגוריתם ה- "Forward" לעיל למעשה עונים על הסוגיה הראשונה של מודל ה-HMM, דהיינו חישוב $p(O|\lambda)$, למרות זאת לצורך פתרון סוגיית האימון של מודל ה-HMM, נדרשת סדרת הסתברויות נוספת, נסמנה ב- β.

נגדיר $\beta_t(i) = p(o_{t+1}, \dots, o_T|q_t = s_i, \lambda)$ בתור ההסתברות של יצירת כל הסימבולים לאחר זמן t, בהינתן והמערכת תמצא במצב s_i בזמן זה. בדומה ל- a, גם הסתברות β יכולה להיות מחושבת על ידי תהליך רקורסיבי באופן הבא:

1. אתחול: $\beta_T(i) = 1$ לא מיוצר סימבול, כמו כן ייתכן כל מצב מבין כלל המצבים האפשריים.

2. רקורסיה: $\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(o_{t+1})$, עבור $1 \le t < T$, כאשר כל מצב s_j שהוא יכול לקיים את היווצרות הסימבול o_{t+1}.

3. סיום: $p(O|\lambda) = \sum_{i=1}^{N} a_t(i)\beta_t(i)$, עבור $1 \le t < T$.

7.7 אלגוריתם "Viterbi":

אלגוריתם תכנון דינאמי זה מאפשר את חישוב מסלול מעבר המצבים בעלי הסבירות הגבוהה ביותר בהינתן סדרה נתונה של סימבולים. למעשה אלגוריתם זה דומה מאוד לאלגוריתם ה-"Forward" שתואר קודם לכן, ההבדל בין שני האלגוריתמים הוא שבעבור אלגוריתם זה נלקח המקסימום ולא הסכום עבור כל אחת מהדרכים האפשריות להגיע למצב הקיים. אף על פי כך התיאור הרשמי של האלגוריתם מכיל מספר סימונים מסורבל.

נניח $q = q_1 q_2 \dots q_T$ הינה סדרה של מצבים, המטרה לפיכך תהיה למצוא $q^* = argmax_q p(q, O|\lambda)$ שלמעשה זהה למציאת $q^* = argmax_s p(q|O,\lambda)$, מכיוון ש- $p(q|O,\lambda) = p(q|O,\lambda)p(O|\lambda)$ ו- $p(O|\lambda)$ לא משפיעים כלל על בחירת ה- q . אלגוריתם זה לפיכך מחפש את המסלול האופטימלי עבור q^* באופן הדרגתי בזמן שנערכת סריקה של כלל הסימבולים שהתקבלו.

בזמן t, האלגוריתם יעקוב אחר כל המסלולים האופטימליים המסתיימים עבור כל אחד מ- N המצבים השונים.בזמן $t+1$, האלגוריתם מעדכן את אותם- N המסלולים האופטימליים. נניח q_t^* הינו המסלול האופטימלי לתת סדרה של סימבולים $O = o_1 o_2 \dots o_t$ עד לזמן t, ו- $q_t^*(i)$ הינו המסלול בעל הסבירות הגבוהה ביותר המוביל למצב s_i בעבור תת סדרת הסימבולים O, כמו כן נניח ש- $VP_t(i) = p(O(t), q_t^*(i)|\lambda)$ הינה ההסתברות המתקבלת במעקב אחר המסלול $q_t^*(i)$ עבור רצף הסימבולים $O(t)$. נאמר $q_t^* = q_t^*(k)$ כאשר $k = argmax_i VP_t(i)$ ו- $q^* = q_T^*$, אלגוריתם ה-"Viterbi" אם כך פועל באופן הבא:

1. אתחול: $VP_1(i) = \pi_i b_i(o_1)$ ו- $q_1^*(i) = (i)$.

2. רקורסיה: $VP_{t+1}(i) = \max_{1\le j\le N} VP_t(\mathrm{j}) a_{ji} b_i(o_{t+1})$ כמו כן $q_{t+1}^*(i) = q_t^*(k).(i)$, עבור $1 \le t < T$, כאשר $k = argmax_{1\le j\le N} VP_t(j) a_{ji} b_i(o_{t+1})$ ו- "." הינו אופרטור השירשור של המצבים ליצירת המסלול, כמו כן $q^* = q_T^* = q_T^*(k)$ כאשר $k = argmax_{1\le i\le N} VP_T(i)$.

7.8 אלגוריתם "Baum-Welch":

הפתרון לסוגיית האימון שמלווה למודל ה-HMM, דהיינו מציאת $\lambda^* = argmax_\lambda p(O|\lambda)$, הינו למעשה בעיית השיערוך בעל הסבירות המירבית. אילו ניתן היה לעקוב אחר מסלול מעבר המצבים אשר יצרו את הסימבולים הנצפים הרי שתהליך השיערוך יהיה פשוט, אולם מסלול המעברים איננו נראה ולפיכך, באופן כללי, לא ניתן למצוא באופן אנליטי את השיערוך בעל הסבירות המירבית. אף על פי כך, בדומה למקרה של GMM שסוקר בהרחבה, ניתן לנצל את

אלגוריתם ה-EM לטובת מודל ה-HMM, אלגוריתם מותאם זה קרוי בשם "Baum-Welch". כמו בעבור יתר המקרים בהם משתמשים באלגוריתם EM תחילה יש לקבוע סדרה של ערכים אקראיים עבור הפרמטרים של λ, לאחר מכן, במהלך כל איטרציה באלגוריתם, מחושבת ההסתברות של כל המצבים החבויים האפשריים במסלול המעבר, בהמשך הפרמטרים משוערכים פעם נוספת בהתבסס על החישוב ההסתברויות עד אשר תתקבל התכנסות עבור ערך הסף המירבי.
כדי לתת ביטויים מתמטיים עבור אלגוריתם זה מגדירים $\gamma_t(i) = p(q_t = s_i|O,\lambda)$ כהסתברות להמצא במצב s_i בזמן t, ו- $\xi_t(i,j) = p(q_t = s_i, q_{t+1} = s_j|O,\lambda)$ כהסתברות למעבר ממצב i למצב j בעבור רגע נתון t, כך שמתקיים $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$ עבור $t = 1, \dots, T$,

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{7.14}$$

עבור $t = 1, \dots, T-1$, $\xi_t(i,j)$ נתון באופן הבא:

$$\xi_t(i,j) = \frac{\alpha_t(i)\alpha_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} = \frac{\gamma_t(i)\alpha_t(i)\alpha_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\beta_t(i)} \tag{7.15}$$

נוסחאות עידכון הפרמטרים של λ מוגדרות באופן הבא:

$$\pi_i' = \gamma_1(i) \tag{7.16}$$

$$a_{ij}' = \frac{\sum_{T}^{T-1} \xi_t(i,j)}{\sum_{j=1}^{N} \sum_{t=1}^{T-1} \xi_t(i,j)} \tag{7.17}$$

$$b_i(v_k)' = \frac{\sum_{t=1, o_t=v_k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{7.18}$$