1 ראשי פרקים

1.1 דיבור

1.1.1 תדר

1.1.2 מבנה דיבור (הפקת צליל)

1.1.3 מאפיינים בתדר

1.3 מסווגים

1.4 רגרסיה

1.5 למידה עמוקה

1.1 דיבור

דיבור, תקשורת אנושית באמצעות שפה מדוברת. למרות שבעלי חיים רבים הם בעלי קולות מסוגים שונים ויכולות הטיה, בני האדם למדו לווסת את קולם על ידי ניסוח גווני הגרון לדיבור נשמע. [1]

הדיבור הוא אות פיזיולוגי ייחודי המכיל מידע במספר רב של רמות הנוגעות לתוכן הלשוני כגון, מילים, מסר, מבטא, שפה. וכן לתוכן לא לשוני כגון, מין, גיל, זהות ומצב רגשי.

1.1.1 תדר

ניתן להגדיר תדר ככמות הפעמים שמשהו קורה ביחידת זמן אחת ( תדר = #הפעמים\זמן). כדי לנתח אות באופן תדרי, משתמשים בהתמרת פורייה. התמרת פורייה ממירה מתחום הזמן (אמפליטודה\*זמן) לתחום התדרים (אמפליטודה\*תדר), והיא מחשבת את התדרים השונים מהם מורכב האות.

1.1.2 מבנה דיבור

אדם מפיק אות דיבור בעזרת ששה חלקי גוף שונים - ריאות, תיבת הקול - הנמצאת בסוף קנה הנשימה, לשון, חך רך - הנמצא בחלק האחורי של הפה, חלל האף וחלל הפה. לכל אחד מהחלקים חשיבות גבוהה להפקת הצליל וניתן להפריד אותם לשתי קבוצות - הריאות ותיבת הקול - המשמשים ליצירת אות הקול, וכל שאר חלקי הפה. הריאות ותיבת הקול. הריאות דוחפות אוויר דרך תיבת הקול, בתוך תיבת הקול יש שתי פיסות רקמה שיכולות להתהדק או להרפות. פיסות הרקמה יכולות לאפשר לאוויר לעבור בקלות, או להפריע לאוויר לעבור כלומר להיפתח ולהסגר במהירות. הפעולה הזו יוצרת פולסי אוויר מחזוריים, שגורמים לצליל, בצורה דומה לשריקה עם השפתיים. כל שאר חלקי הפה יכולים להשתייך לקבוצת עיוות האות, כאשר פולסי האוויר (האות הבסיסי) נכנסים לאזור הפה רקמות בחלקים האלה נעות בצורות מסוימות שגורמות למערבולות קטנות, המשפיעות ומעוותות את האות הבסיסי שיצא מתיבת הקול. אותות דיבור אלו אפשר לחלק לשתי קבוצות מאוד כלליות:

1. אותות קוליים voiced)) - אותות הנוצרים מרעידה מחזורית של מיתרי הקול .למשל, התנועות: i, u, o, e, a .
2. אותות א-קוליים unvoiced)) - אותות לא מחזוריים הנוצרים כאשר מיתרי הקול במצב פתוח ואינם רוטטים. המאפיינים של אותות מסוג זה דומים לרעש. למשל, התנועות: ph, sh, s .

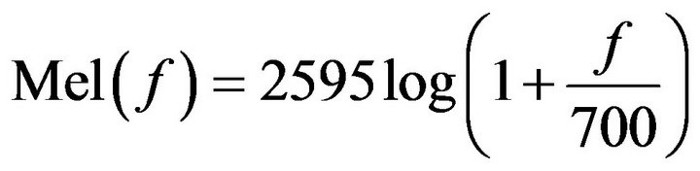
הקול האנושי מייצר צליל באמצעות אוויר נע, שנדחף מהסרעפת, על פני מיתרי הקול, וגורם להם להדהד בצורה שיוצרת צליל הרמוני (בהברות קוליות). דחיסה ושחרור, של מולקולות אוויר גורם להתפשטות תנודות המכונות גל אקוסטי. תנודות אלו נמצאות בטווח תדרים בין 20 ועד כ-20,000 הרץ. הטווח הזה הוא טווח התדרים הנתפסים על ידי בני אדם כצלילים. צליל הרמוני הנוצר בקול האנושי מעוצב על ידי הפה. היווצרות פיזיולוגית של צלילים שונים במיתרי הקול ניכרת כפונקציה. בצורה הזאת בני אדם מעבירים שילובים של צלילים שונים בדיבור ובסגנונות קול שונים בשירה. גובה הצליל או התדר הבסיסי של צליל קולי נקבע על פי קצב תנודת מיתרי הקול, וגם הפסגות הנוספות הנוצרות בכפולות של התדר הבסיסי שנקראות הרמוניות נוצרות מתנודת מיתרי הקול [2]. גישות חישוביות למידול הקול האנושי נופלות לקטגוריות פיזיות או תחזיתיות. הרבה ידוע ומובן במבנה של איברי הקול האנושיים, ולכן בעזרת מודלים פיזיים ניתן להדגים כיצד הקול מפיק צליל.

1.1.3 מאפיינים בתדר

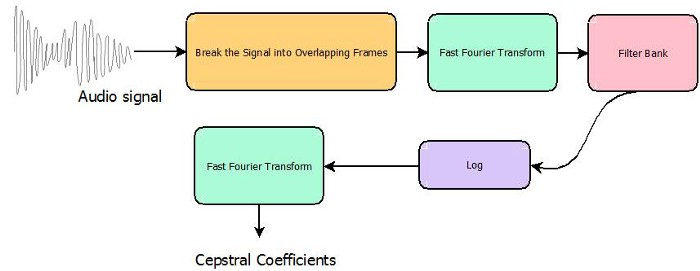
מתוך אות ניתן להוציא מאפיינים כמו pitch (גובה הצליל) ו-amplitude (עוצמת הצליל). מעבר ל-pitch, תדרי התהודה הנוצרים ממערכת הקול נקראים פורמנטים. פורמנטים הם טווחים של תדרים מסוימים שמוגברים על ידי הצורה והגודל של מערכת הקול של האדם. הם שונים עבור אנשים שונים בגלל השוני בצורה ובגודל של מערכת הקול של כל אחד. המבנה הפיזיולוגי של מערכת הדיבור שונה מאדם לאדם והוא גורם לתדר pitch שונה ולתדרי פורמנטים שונים. תדר ה-pitch של דיבור אנושי נע בתחום שבין 50Hz, עבור גברים בעלי טון דיבור נמוך, ל-400Hz עבור ילדים ונשים בעלות טון דיבור גבוה [3].

מאפיינים נוספים שניתן להוציא מתוך אות הינם zero crossing rate) ZCR) ומקדמי MFCC . מאפיין הZCR הוא התדירות בה האות עובר מערכים שלילים לחיובים ולהפך, כלומר התדירות בה האות עובר את נקודת האפס.

מקדמי MFCC הינם מספרים שמראים את הצורה הכללית בה משומשים תדרים שונים באות. המושג MFCC הוא ראשי תיבות של Mel Frequency Cepstral Coefficients. גובה הצליל הוא אחד המאפיינים של אות דיבור ונמדד כתדירות האות. מל (Mel) או סולם מל הוא סולם המייחס בין התדר הנתפס של הטון לתדר הנמדד בפועל. הוא מדרג את התדר כדי להתאים יותר למה שאוזן האדם יכולה לשמוע (בני אדם טובים יותר בזיהוי שינויים קטנים בדיבור בתדרים נמוכים יותר). ניתן להמיר תדר הנמדד בהרץ (f) לסולם מל באמצעות הנוסחה הבאה:



המעטפת של ספקטרום כוח זמן של אות הדיבור מייצגת את מערכת הקול ואת ה- MFCC שמייג את המקדמים המרכיבים את קפסטרום ( cepstrum - התחום בו מקדמי הMFCC נמצאים) תדר מל מייצגת במדויק את המעטפת הזו. [4]



איור : המרה מאות שמע לMFCC

1.3 מסווגים

זיהוי גיל או מין של דובר יכול להתבצע בעזרת מסווג (classifier) מאומן. בסטטיסטיקה ובלמידה חישובית, סיווג הוא פעולה שמחלקת קבוצת עצמים לתת-קבוצות. לדוגמה סיווג מין של דובר לשתי קבוצות (זכר , נקבה) הנקרא גם סיווג בינארי ( סיווג שמחלק עצמים לשתי קבוצות בלבד). סיווג הוא התהליך של חיזוי קבוצה של נקודות נתונים נתונות. שיעורים נקראים לפעמים מטרות / תוויות או קטגוריות. מידול חיזוי סיווג הוא המשימה של קירוב פונקציית מיפוי (f) ממשתני קלט (X) למשתני פלט בדידים[[1]](#footnote-0) (y). סיווג שייך לקטגוריה של למידה מפוקחת, כאשר העצמים מספקים גם את נתוני הקלט. ישנן יישומים רבים בסיווג בתחומים רבים כגון באישור אשראי, אבחון רפואי, שיווק יעד וכו ' [5].

ישנם שני סוגים של לומדים מסווגים לומדים עצלנים ולומדים להוטים (eager learners) .

לומדים עצלנים מאחסנים את נתוני האימון ומחכים לנתוני בדיקה. כאשר זה קורה, הסיווג מתבצע על בסיס הנתונים הקשורים ביותר בנתוני האימונים המאוחסנים. בהשוואה ללומדים להוטים, לומדים עצלנים לוקחים פחות זמן אימון אך יותר זמן בחיזוי, לומדים להוטים לאומת זאת, בדיוק להפך. לומדים להוטים בונים מודל סיווג המבוסס על נתוני האימון הנתונים לפני קבלת נתונים לסיווג. עליו להיות מסוגל להתחייב להשערה אחת המכסה את אפשרויות הקלט. בשל בניית המודל, ללומדים להוטים לוקח הרבה זמן לרכבת ופחות זמן לחיזוי [5].

יש הרבה אלגוריתמים לסיווג, אך שניים המשומשים ביותר הם [5]-

1. עץ החלטות (Decision Tree) - עץ החלטות בונה מודלים של סיווג או רגרסיה בצורת מבנה עץ. היא משתמשת בקבוצת כללי אם-אז (if-then) שהיא בלעדית הדדית וממצה לסיווג. החוקים נלמדים ברצף תוך שימוש בנתוני האימון אחד אחד.
2. השכן הקרוב ביותר ((k-Nearest Neighbor (KNN) - הוא אלגוריתם למידה עצלן אשר מאחסן את כל המופעים המתאימים לנקודות נתוני אימון במרחב n ממדי. כאשר מתקבל נתון דיסקרטי לא ידוע, הוא מנתח את מספר k המופעים הקרובים ביותר שנשמרו (השכנים הקרובים ביותר) ומחזיר את המחלקה הנפוצה ביותר כחיזוי ולנתונים בעלי ערך אמיתי הוא מחזיר את הממוצע של k השכנים הקרובים ביותר.

1.4 רגרסיה

אך כדי לקבל תוצאות טובות יותר בשיערוך גיל עדיף להשתמש ברגרסיה, שבעזרתה אפשר יהיה להעריך גלאים מחוץ לתחום הגילאים של משתני הקלט. רגרסיה היא קבוצת גישות סטטיסטיות המשמשות לקירוב הקשר בין משתנה תלוי למשתנה בלתי תלוי אחד או יותר. הגרסה הנפוצה ביותר של רגרסיה היא רגרסיה לינארית, שמוצאת קו, (hyperplane) המתאים ביותר לנתונים. קו זה מאפשר לאמוד את ערך המשתנה התלוי כאשר המשתנים הבלתי תלויים מתקבלים כקבוצה נתונה של ערכים.למטרת המידול/ האימון, הביצועים של שיטות לימוד מכונה נמדדים באמצעות פונקציות ההפסד[[2]](#footnote-1) (loss function). במקרה של רגרסיה לינארית, זה מציין עד כמה הוא מתאים לנתונים [6].

1.5 למידה עמוקה (deep learning)

בשנים האחרונות התחילו חוקרים להשתמש בלמידה עמוקה שבה אין צורך בחילוץ ידני של מאפיינים. בלמידה עמוקה משתמשים בפונקציות מתמטיות השואבות השראה מהדרך בה המוח האנושי לומד, והן משמשות לזיהוי עצמים, זיהוי דיבור, תרגום שפות וקבלת החלטות, ללא חילוץ ידני של מאפיינים מתוך האות. טכניקה של למידה עמוקה שימושית בתחום הסיווג, מכיוון שהמחשב מצליח למצוא מאפיינים ואת הקשר ביניהם לבד, ללא עזרת אדם.

למידה עמוקה נמצאת בתחום רחב במדעי המחשב שנקרא רשתות ניורוניות. רשתות ניורוניות (neural networks) הם חלק של תחום למידת מכונה (Machine Learning) המאופיין בשימוש במספר רב של חישובים שזורים זה בזה. למידת מכונה נמצאת בתוך תחום הבינה מלאכותית (AI). רשתות ניורוניות מהוות את עמוד השדרה של אלגוריתמי למידה עמוקה. למעשה, מספר השכבות, או העומק, של רשתות נוירונים הוא שמבדיל רשת עצבית אחת מאלגוריתם למידה עמוקה, שחייב להכיל יותר משלוש. רשתות ניורונים מורכבות מכמה רכיבים מרכזיים, אחד מהם הוא הנוירון [8].

הנוירון מהווה אבן בניין לרשת הנוירונים, והוא בנוי מקלט שבו ניתן ערך מסוים, משקל לערך וגם הטיה (bias). הערך והמשקל יכולים להינתן כקלט ראשוני או כפלט מנוירון אחר. לנוירון יכולים להיות מספר של ערכי קלט עם משקלים שונים. בנוסף לנוירון יש פלט שהוא התוצאה של החישובים בתוך הנוירון. תפקידו של נוירון בשכבה נסתרת הוא לקחת את סכום מכפלת ערכי הקלט ומשקולותיהם ולהעביר ערך זה לפונקציית הפעלה. זה יהיה הערך שהועבר כקלט לנוירון הבא (הפלט). מרכיב שני ברשת נוירונים הינם הקשרים בין הנוירונים, לכל קשר בין נוירון יש משקל (המשקל של הערכים הנקלטים לתוך הנוירונים) [9].

פונקציית הפעלה ברשת עצבית מגדירה כיצד הופך הסכום המשוקלל של הקלט ליציאה מצומת או צמתים בשכבה של הרשת. לפעמים פונקציית ההפעלה מכונה "פונקציית העברה". אם טווח הפלט של פונקציית ההפעלה מוגבל, אז זה יכול להיקרא "פונקציית סקווש" (squashing function). פונקציות הפעלה רבות אינן ליניאריות וניתן לכנות אותן כ"לא לינאריות "בשכבה או בתכנון הרשת. לבחירת פונקציית ההפעלה יש השפעה רבה על היכולת והביצועים של הרשת העצבית, וניתן להשתמש בפונקציות הפעלה שונות בחלקים שונים של המודל. מבחינה טכנית, פונקציית ההפעלה משמשת בתוך או אחרי העיבוד הפנימי של כל צומת ברשת, אם כי רשתות נועדו להשתמש באותה פונקציית הפעלה לכל הצמתים בשכבה [10,11]. שלושת סוגי פונקציות ההפעלה השכיחות ביותר הינם-

* (Rectified Linear Activation (ReLU
* (Logistic (Sigmoid
* (Hyperbolic Tangent (Tanh

רשתות נוירונים מארגנות נוירונים בשכבות. לרב רשתות נוירונים מחולקים לשלושה חלקים, השכבה הראשונה שבא נמצא הקלט נקראת שכבת הקלט, החלק השני שהיא השכבה הנסתרת (hidden layer) שבה נעשים כל החישובים (יכולה להיות יותר משכבה אחת של שכבה נסתרת) , והשכבה האחרונה נקראת שכבת הפלט שכבה בה כל נוירון מחובר לכל נוירון אחר בשכבה הבאה שלו נקראת שכבה צפופה [9].

למידה עמוקה מאפשרת למודלים חישוביים המורכבים משכבות עיבוד מרובות ללמוד ייצוגים של נתונים עם רמות מופשטות מרובות. שיטות אלה שיפרו באופן דרמטי זיהוי דיבור, וזיהוי אובייקטים חזותיים. למידה עמוקה מגלה מבנה מורכב במערכי נתונים גדולים על ידי שימוש באלגוריתם (backpropagation ) כדי לציין כיצד מכונה צריכה לשנות את הפרמטרים הפנימיים שלה המשמשים לחישוב הייצוג בכל שכבה מהייצוג בשכבה הקודמת. רשתות עוויתות עמוקות הביאו לפריצות דרך בעיבוד תמונות, וידאו, דיבור ושמע, ואילו רשתות חוזרות והאירו אור על נתונים עוקבים כמו טקסט ודיבור.

האלגוריתם (backpropagation) משמש לאימון יעיל של רשת נוירונים באמצעות שיטה הנקראת כלל שרשרת. במילים פשוטות, לאחר כל מעבר קדימה ברשת, (backpropagation) מבצעת מעבר לאחור תוך התאמת פרמטרי המודל (משקולות והטיות). מטרת האלגוריתם היא להקטין הסתייה של פלט המודל לערכים הרצויים [12].

ישנם סוגים שונים של למידה עמוקה, המשומשים ביותר מבינהם הם [13] -

1. (Classic Neural Networks) - יעיל ביותר בבעיות רגרסייה וסיווג עם מספרים ממשיים.
2. (Convolutional Neural Networks \ CNN) - רשתות ה- CNN מתאימות למשימות, כולל זיהוי תמונות, ניתוח תמונות, פילוח תמונות, ניתוח וידאו ועיבוד שפה טבעית.
3. (Recurrent Neural Networks \ RNN) - תוכננו לראשונה כדי לסייע בחיזוי רצפים. בעיות כמו סיווג תמונות, סיווג וידאו ותרגום שפה.

1. משתנים מתוך סט מוגדר של ערכים אפשריים [↑](#footnote-ref-0)
2. פונקציית ההפסד היא שיטה להערכת מידת ביצועי האלגוריתם במערך הנתונים [7]. [↑](#footnote-ref-1)