

רקע תאורטי בנושא:

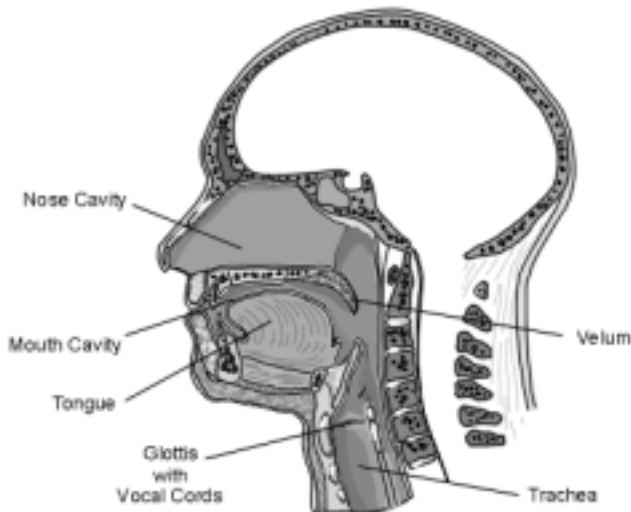
LPC דחיסת דיבור מבוססת אנליזת

קיימים אלגוריתמים רבים לעיבוד אות המתוכננים לפעול על אות דיבור. אלגוריתמים מסוג זה עשויים, למשל, לבצע דחיסה של אות הדיבור או שיפור שלו (ניקוי רעשים, הסרת הדהוד וכד'). כדי להשיג תוצאות טובות, אלגוריתמים אלה מסתמכים לרוב על התכונות הפיסיולוגיות של מערכת הדיבור האנושית.

מטרתנו להכיר את מבנה אות הדיבור, את תכונותיו הסטטיסטיות והמודל פרמטרי לייצוגו. כמו כן, יוכרו שיטות המשמשות לשערוך פרמטרי מודל זה ויודגם השימוש בו לשם ביצוע דחיסה יעילה של האות.

מבנה אות הדיבור ותכונותיו

בציור שלפניכם ניתן לראות כיצד פועלת מערכת יצירת הקול.



ציור 1: מערכת הקול האנושית

לחץ אוויר מהריאות עובר דרך קנה הנשימה (trachea) ומגיע למיתרי הקול (vocal cords) הנמצאים בבית הקול (glottis). מיתרי הקול, שהם שתי פיסות רקמה המוצמדות זו לזו, יכולים לאפשר לאוויר לעבור בחופשיות או שהם יכולים לרטוט, כלומר להיפתח ולהיסגר במהירות. פתיחה וסגירה זו יוצרת פולסי אוויר מחזוריים. המרווח המחזורי בין הפולסים, הנקרא מחזור ה-pitch, מושפע משינויים בלחץ האוויר ובמתיחות של מיתרי הקול.

זרם האוויר מתפשט ממיתרי הקול (הפתחים או הנפתחים-נסגרים לסירוגין) למעבר הקולי (vocal tract), כלומר, אל חלל הפה (mouth/oral cavity)

על פני הלשון, השיניים והשפתיים ואל חלל האף (nose cavity) דרך הנחיריים. החך הרך (velum) הוא שסתום הממוקם בפתח חלל האף והשולט על מעבר האוויר לתוך חלל זה.

אות העירור (excitation signal) למערכת הקול הוא זרם האוויר המגיע אל חללי הפה והאף. החך הרך, הלשון, השפתיים והלסת התחתונה יוצרים מעצורים שונים הגורמים למערבולות בזרם האוויר וכך יוצרים את חיתוך הדיבור (articulation). פונקציה התמסורת של חללי הפה והאף משתנה בזמן, מפני שצורתם של חללים אלה משתנה תוך כדי דיבור. פונקציה תמסורת זו מאפנת את פולסי האוויר ויוצרת את אות הדיבור הנפלט החוצה מהשפתיים ומהנחיריים.

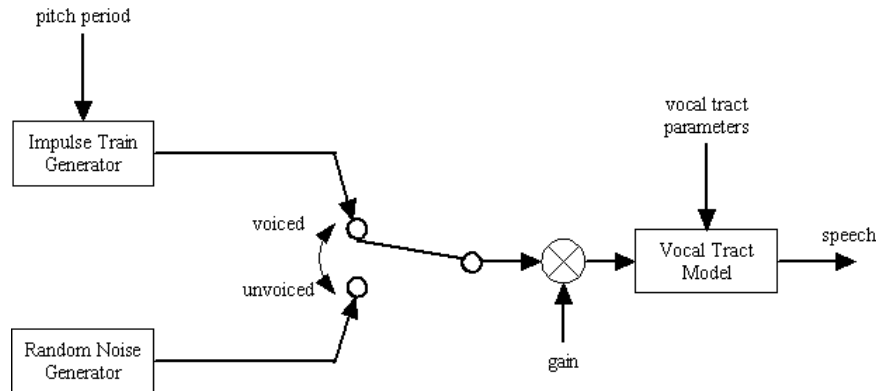
את אות הדיבור ניתן לחלק באופן גס לשני סוגים :

- אותות קוליים (voiced) – אותות מחזוריים הנוצרים מרעידה של מיתרי הקול. למשל, התנועות: a, e, o, u, i. נסו להניח את היד על הגרון בעת השמעת הברות אלה ותרגישו רטיטות.
- אותות א-קוליים (unvoiced) – אותות לא מחזוריים הנוצרים כאשר מיתרי הקול במצב פתוח ואינם רוטטים. המאפיינים של אותות מסוג זה דומים לרעש. למשל, התנועות: s, sh, ph. נסו להניח את היד על הגרון בעת השמעת הברות אלה ולא תרגישו רטיטות.

חללי האף והפה מהווים תיבת תהודה, כלומר, מסנן בעל תדרי תהודה (קטבים) שמשתנים לפי צורת החלל. תדרי תהודה אלו נקראים פורמנטים (formants). השילוב של תדרי פורמנטים שונים ושני סוגי הקולות יוצרים את ההברות השונות.

המבנה הפיסיולוגי 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).

על מנת לנתח אותות דיבור בכלים של עיבוד אות יש צורך למצוא מודל שמתאר את מערכת הדיבור במושגים מתמטיים. מודל קלאסי כזה של מערכת המייצרת אות דיבור הוא :



ציור 2: מודל יצירת אות דיבור

כאשר הדיבור הוא קולי (voiced), מיתרי הקול מייצרים פולסים מחזוריים המיוצגים במודל בעזרת מקור עירור המייצר רכבת הלמים, במרווחי זמן קצובים ביניהם. מרווחי זמן אלה הם משכי מחזור ה-pitch. המודל מניח כי קטעי דיבור קוליים הם קווי-מחזוריים, כלומר, הם משתנים באופן איטי יחסית לזמן המחזור שלהם כך שניתן לייצגם, למשך קטעי זמן קצרים, בעזרת מרווחי זמן קבועים. כאשר הדיבור הוא א-קולי (unvoiced), התהליך מיוצג ע"י מקור אות עירור דמוי רעש רחב סרט.

לפי מודל זה, מערכת הדיבור מפיקה בכל רגע אות קולי לחלוטין או אות א-קולי לחלוטין. זאת בניגוד למודל מורכב יותר, שעשוי לאפשר גם מצבים של הברות מעורבות, המשלבות את שני סוגי הקולות (למשל: z, v). הבחירה בין שני מקורות העירור נעשית ע"י מתג, ובשני המקרים מועבר אות העירור דרך מסנן משתנה בזמן המייצג את חללי הפה והאף. תגובת התדר של מסנן זה, המאופיינת בעיקר ע"י הקטבים (מן האפסים נוטים בדרך כלל להתעלם במקרה זה), מעצבת את צורת המעטפת הספקטראלית של אות העירור ומשנה את עוצמתו בהתאם. הפלט של המערכת הוא אות העירור המעוצב ספקטראלית.

אות דיבור הוא אות מוגבל סרט שרוב האנרגיה שלו מרוכזת בתדרים נמוכים. לרוב מעבירים אות זה דרך מסנן מעביר נמוכים בעל תדר קיטעון נמוך מעט מ-4KHz ודוגמים ב-8KHz. רוחב סרט זה שומר על תכונות המובנות, יכולת הזיהוי והטבעיות של אות הדיבור בצורה טובה.

שערוך תדר ה-pitch

תדר ה-pitch הוא תכונה בסיסית של אות הדיבור. רגישות האוזן האנושית לשינויים בתדר זה היא גדולה יותר, בסדר גודל, מאשר הרגישות שלה לשינויים בפרמטרים אחרים של אות הדיבור. לכן, לשערוך תדר ה-pitch באופן מדויק יש חשיבות גדולה בעת יצירת אלגוריתם המשתמש במודל של אות הדיבור.

שערוך תדר ה-pitch באופן מדויק מאות דיבור דגום הוא בעייתי בשל מספר סיבות:

- תדר ה-pitch משתנה ממחזור למחזור לפי מיקום ההברה במשפט (שינוי זה יכול להגיע עד ל-10% בין שני מחזורים עוקבים).
- קיימים הבדלים באמפליטודת האות בין מחזור למחזור הנובעים מהדגשים שונים בתוך המשפט.
- בעת מעבר מאות קולי לאות א-קולי, או מאות א-קולי לאות קולי, עשוי להתווסף לאות המחזורי רכיב של רעש המקשה על גילוי המחזוריות.
- הרמוניות של תדר ה-pitch או פורמנטים עלולים לעיתים להיראות חזקים יותר מתדר ה-pitch.
- קצב הדגימה הנמוך יחסית והכימוי (קוונטיזציה) הנגרמים בעת הדגימה מוסיפים שגיאות ומקשים על חישוב מדויק של תדר ה-pitch.

קיימים אלגוריתמים רבים לשערוך תדר ה-pitch. בתרגיל זה נכיר שיטה המשמשת בסיס לרב אלגוריתמים אלה – שיטת האוטוקורלציה. בנוסף, נכיר מספר תוספות לשיטה זו שמטרתן לשפר את ביצועיה.

שיטת האוטוקורלציה

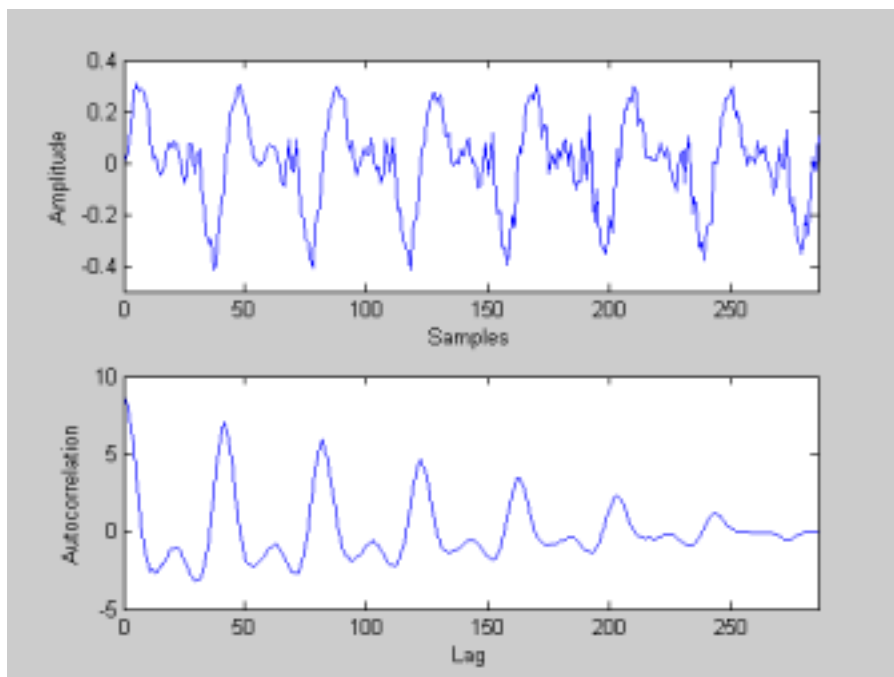
פונקציית הקורלציה מהווה מדד לרמת הדמיון בין שני אותות. פונקציית האוטוקורלציה מודדת את רמת הדמיון של האות לגרסה מוזזת בזמן של עצמו. המקסימום של פונקציית האוטוקורלציה יתקבל במרווחי מחזור ה-pitch של האות המקורי.

נתון קטע אות דיבור $x(n)$ באורך N . פונקציית האוטוקורלציה לזמן קצר של קטע אות זה מוגדרת באופן הבא:

$$r_k = \sum_{n=0}^{N-k-1} x_n x_{n+k} \quad k = 0, \dots, N-1$$

k הוא ההזזה יחסית לאות המקורי.

בשרטוט הבא אנו רואים קטע אות דיבור קולי ואת סדרת האוטוקרלציה שלו :



ציור 3: קטע אות דיבור קולי ופונקציית האוטוקרלציה המתאימה

ניתן לראות בציור כי פונקציית האוטוקרלציה מקבלת מקסימום מקומי עבור ערכי k השווים למחזור ה-pitch ולכפולות שלמות שלו. מכאן ניתן לשערך את מחזור ה-pitch ע"י מציאת המקסימום המקומי הגדול ביותר של פונקציית האוטוקרלציה שאינו באפס.

לבחירת אורך החלון N יש שני שיקולים סותרים – מצד אחד, החלון צריך להיות ארוך מספיק כך שבכל מקרה יכיל לפחות שני מחזורי pitch שלמים. מצד שני, האות הוא קווי-סטציונרי ולכן רצוי לקחת חלון קצר מספיק כדי שהנחת הסטציונריות תתקיים. ערך מקובל לאותות דיבור, שעונה על שתי דרישות סותרות אלה הוא חלון באורך 20-40msec.

עיבוד מקדים

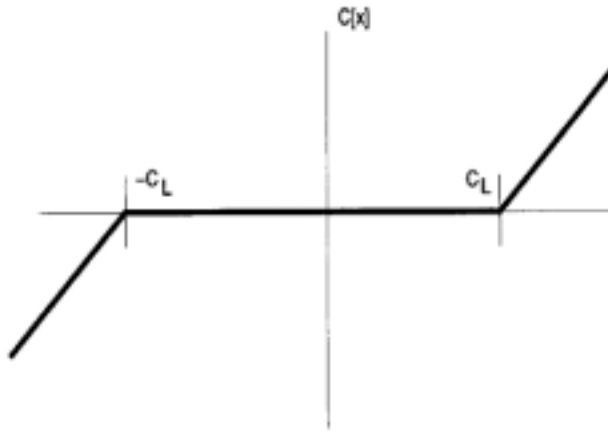
בפונקציית האוטוקרלציה ניתן להבחין, בנוסף לשיאים המופיעים בכפולות של זמן מחזור ה-pitch, גם בשיאים נוספים הנובעים מהשפעת הפורמנטים. שיאים נוספים אלה מקשים על זיהוי נכון של תדר ה-pitch. כדי להקל על פעולת הזיהוי, ניתן להעביר את האות דרך מסנן מעביר נמוכים. מסנן זה מנחית תדרים גבוהים הנמצאים מחוץ לתחום החוקי של תדר ה-pitch. בנוסף, ניתן לקטום את האות. תפקידה של פעולת הקטימה הוא להנחית שיאים הנובעים מפורמנטים גבוהים המקשים על זיהוי זמן מחזור ה-pitch.

קטימה (center clipping) היא הפעולה הלא-ליניארית הבאה :

$$\text{clip}(x_n) = \begin{cases} x_n - C_L & x_n > C_L \\ 0 & |x_n| \leq C_L \\ x_n + C_L & x_n < -C_L \end{cases}$$

C_L הוא קבוע, הנקבע באופן אדפטיבי לכל חלון. ערכים אופייניים ל- C_L הם בתחום 60%-80% מאנרגיית האות המקסימלית.

פעולת הקטימה נראית כך :



צוור 4 : פונקציית הקטימה

החלטה על קיום דיבור ועל voiced/unvoiced

ניתן לחלק כל אות דיבור דגום לקטעים בהם קיים דיבור ולקטעי שקט. לפי המודל שהוצג ליצירת אות דיבור, את הקטעים בהם קיים דיבור ניתן לחלק לקטעים המכילים אות קולי ולקטעים המכילים אות א-קולי.

לאלגוריתם המזהה קיום דיבור או קיום שקט נוהגים לקרוא VAD (Voice Activity Detector). בנייתו של VAD איכותי היא מלאכה מורכבת המערבת לרוב שקלול של פרמטרים שונים. הפרמטר החשוב ביותר הוא פונקציית האנרגיה לזמן קצר של האות (short-time energy function) המוגדרת באופן הבא :

$$E_n = \sum_{m=-\infty}^{\infty} [x_m w_{n-m}]^2$$

עבור חלון מלבני :

$$w_n = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{elsewhere} \end{cases}$$

כפי שניתן להבין באופן אינטואיטיבי, קיים סיכוי גבוה כי הקטעים שבהם האנרגיה לזמן קצר של האות הדגום נמוכה מאוד הם קטעי שקט.

גם זיהוי קטעי אות קולי וקטעי אות א-קולי הוא פעולה עדינה המערבת שקלול של מספר פרמטרים. פרמטר המביא לרוב לתוצאות טובות הוא פונקצית האנרגיה לזמן קצר, שזה עתה הוצגה. קטעים קוליים מאופיינים באנרגיה גבוהה, לעומת קטעים א-קוליים המאופיינים באנרגיה נמוכה.

פרמטר שימושי נוסף המבדיל בצורה טובה בין קטעים קוליים וקטעים א-קוליים הוא קצב חציות האפס (zero-crossing rate). קצב חציות האפס הוא מספר הפעמים בו דגימות עוקבות של האות משנות את סימןן האלגברי משלילי לחיובי או להפך. פונקציה זו מוגדרת באופן הבא:

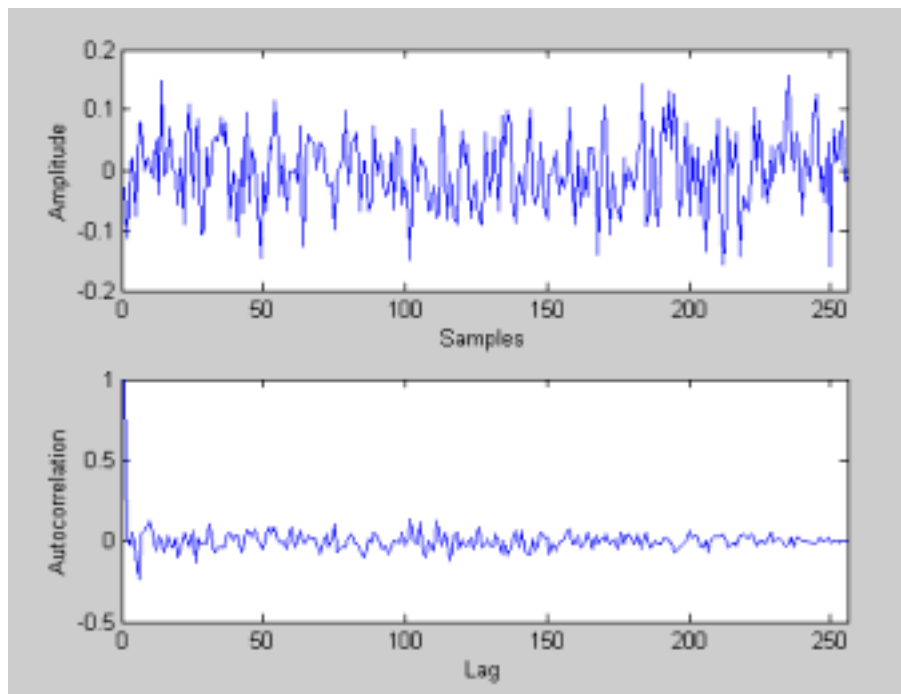
$$Z_n = \frac{1}{2} \sum_{m=n-N+1}^n |\text{sgn}(x_m) - \text{sgn}(x_{m-1})|$$

כאשר:

$$\text{sgn}(x) = \begin{cases} 1 & x_n \geq 0 \\ -1 & x_n < 0 \end{cases}$$

קצב חציות האפס הוא גבוה יותר בקטעים א-קוליים מאשר בקטעים קוליים.

אמצעי הבחנה נוסף בין קטעים קוליים לקטעים א-קוליים הוא צורת פונקצית האוטוקורלציה. לפונקצית האוטוקורלציה צורה שונה עבור אות קולי ועבור אות א-קולי מכיוון שבאות קולי קיימת קורלציה גבוהה בין מחזורי ה-pitch (כפי שניתן היה לראות בציור 3) ובאות א-קולי, שהוא דמוי רעש ובעל אקראיות גבוהה, הקורלציה בין אזורים שונים באות היא נמוכה (כפי שרואים בציור 5).



ציור 5: קטע אות דיבור א-קולי ופונקצית האוטוקורלציה המתאימה

חיזוי ליניארי

חיזוי ליניארי או LPC (Linear Predictive Coding) היא שיטה נפוצה לייצוג המעטפת הספקטרלית של אותות דיבור. שיטה זו עומדת בבסיסם של רוב מקודדי הדיבור המודרניים. היא משמשת במקודדים אלה לשערוך מקדמי המסנן המייצג את מערכת הקול במודל ליצירת אות הדיבור שהוצג. חיזוי ליניארי מאפשר לייצג את מעטפת הספקטרום של קטע דיבור קצר בעזרת מספר מועט של פרמטרים.

הנחה בסיסית הנובעת ממודל הדיבור היא שאות הדיבור ניתן לייצוג ע"י מודל אוטורגרסיבי מסדר p (AR(p) – autoregressive model of order p), כלומר, שכל דגימה שלו מקיימת:

$$x_n = \sum_{i=1}^p a_i x_{n-i} + u_n$$

כאשר $\{a_i\}_{i=1}^p$ הוא סט של מקדמים קבועים המיוצג ע"י הווקטור \underline{a} , ו- u_n הוא דגימה מתהליך רעש לבן בעל ממוצע אפס. לכן, ניתן להתייחס ל- x_n כאל מוצא של מערכת ליניארית בעלת פונקצית תמסורת $H(z)$ מסוג all-pole המעוררת ע"י רעש לבן:

$$H(z) = \frac{X(z)}{W(z)} = \frac{1}{A(z)}$$

כאשר:

$$A(z) \triangleq 1 - \sum_{i=1}^p a_i z^{-i}$$

אנו מעוניינים לשערך את הדגימה x_n מתוך סט של דגימות קודמות עוקבות: $[x_{n-1}, x_{n-2}, \dots, x_{n-p}]$, ע"י קומבינציה ליניארית שלהן:

$$x'_n = \sum_{i=1}^p a'_i x_{n-i}$$

כלומר, חוזים את ערכו של x_n ע"י חזאי ליניארי מסדר p . שגיאת החיזוי היא:

$$e(n) \triangleq x_n - x'_n = x_n - \sum_{i=1}^p a'_i x_{n-i}$$

כך שאם $\underline{a}' = \underline{a}$, הרי ששגיאת החיזוי תהיה לבנה. בחירה נכונה של סדר המודל p תביא לקיום תכונה זו בקירוב.

הבעיה שעומדת לפנינו היא לשערך את וקטור הפרמטרים \underline{a} ונרצה למצוא את וקטור המקדמים \underline{a}' האופטימאלי במובן של מינימום שגיאה ריבועית ממוצעת (MMSE – Minimum Mean Squared Error), כלומר:

$$\underline{a}'_{opt} = \min_{\underline{a}'} \mathcal{E}^2 = \min_{\underline{a}'} E\{e_n^2\}$$

מטרתנו היא לשערך את וקטור המקדמים כך ששונות שגיאת החיזוי ε^2 תהייה מינימלית. לאחר גזירת הביטוי המתקבל עבור ε^2 לפי כל אחד מהפרמטרים והשוואה לאפס נקבל את הסט הבא של משוואות לחישוב \underline{a} (בהנחה ש- $e(n)$ לבנה):

$$\sum_{i=1}^p a_i \cdot r_{j-i} = r_j, \quad j = 1, 2, \dots, p$$

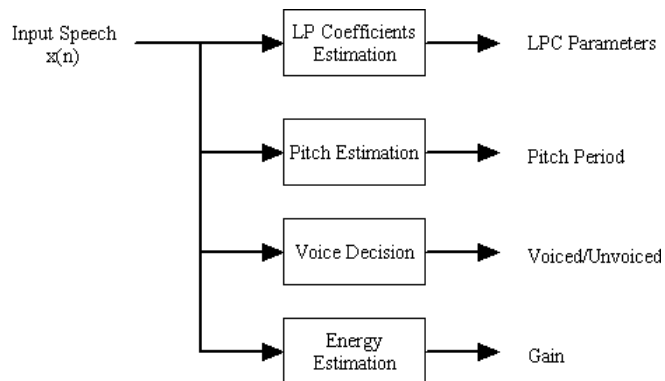
כאשר:

$$r_j \triangleq R_{yy}(j) = E\{x_n x_{n-j}\}$$

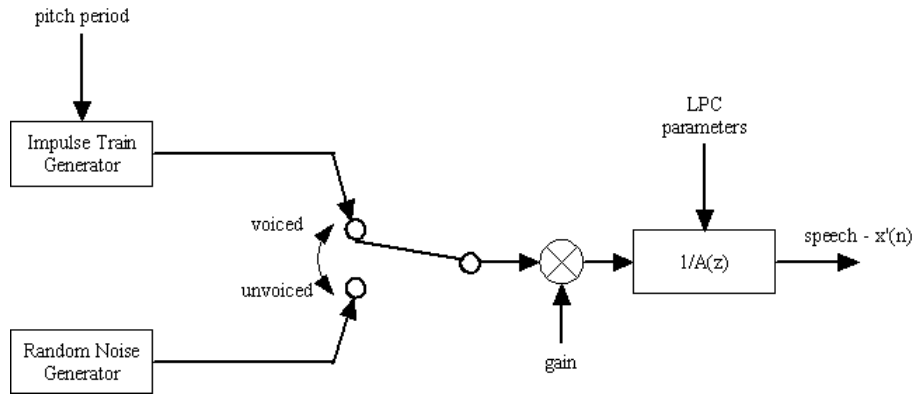
מערכת משוואות זו נקראת משוואות Yule-Walker או משוואות נורמליות.

לשם פתירת מערכת המשוואות יש לשערך תחילה את מקדמי פונקציית הקורלציה מתוך הדגימות הנתונות של האות. שתי השיטות השימושיות ביותר לשערך מקדמי הקורלציה הן שיטת האוטוקורלציה, שאותה כבר הכרנו, ושיטת הקווריאנס. בנוסף, קיימות ווריאציות שונות על שיטות אלו. לפתירה ללא היפוך ישיר של המטריצה של מערכת המשוואות משתמשים באלגוריתם חישוב רקורסיבי הנקרא אלגוריתם Levinson-Durbin. שיטות אלה הן מעבר להיקף התרגיל, אך ניתן למצוא עליהן מידע במקורות לעיון נוסף המופיעים בסוף החוברת.

מקודד ומפענח מבוססי LPC עשויים להראות כך:



צור 6: מקודד LPC



ציוור 7 : מפענת LPC

מדדים אובייקטיביים לאיכות דיבור

כאשר מתכננים אלגוריתם לעיבוד דיבור קיים צורך להעריך את איכות הדיבור לאחר העיבוד לעומת איכותו לפני העיבוד. לא קיימת נוסחה המסוגלת לתת הערכה מושלמת של איכות דיבור משום שאיכות זו מושפעת מגורמים פסיכולוגיים וסביבתיים רבים וקשורה באופן הדוק לתפיסה הסובייקטיבית של כל אדם ואדם. לכן, במקרים רבים משתמשים לצורך הערכת איכות דיבור במדדים סובייקטיביים המבוססים על ניתוח תוצאות האזנה מבוקרת ודירוג ע"י אנשים שונים.

עם זאת, קיים מאמץ בלתי פוסק למציאת נוסחה שתיתן תוצאות קרובות ככל האפשר לבדיקה סובייקטיבית. נציין בהקשר זה את מדד ה-PESQ (Perceptual Evaluation of Speech Quality) עפ"י תקן ITU P.862 הנותן תוצאות קרובות מאוד למדדים הסובייקטיביים. אנו נעזר בשני מדדים אובייקטיביים פשוטים לחישוב – SNR ו-Segmental SNR.

מדד ה-SNR (Signal to Noise Ratio) מחשב את היחס בין האנרגיה של האות המקורי לבין האנרגיה של הרעש (ההפרש בין האות המעובד לאות המקורי):

$$SNR = 10 \log_{10} \left(\frac{\sum_{n=0}^{N-1} x_n^2}{\sum_{n=0}^{N-1} (x_n - x_n')^2} \right)$$

כאשר x הוא האות המקורי ו- x' הוא האות המשוחזר. N הוא אורך מקטע האות שעליו מתבצע החישוב.

מדד ה-SNR נותן לאזורים באות בעלי אמפליטודה גדולה משקל גדול יותר מאשר לאזורים בעלי אמפליטודה נמוכה. עם זאת, שגיאות קטנות באזורים בעלי אמפליטודה נמוכה עלולות להיות חשובות. מדד Segmental SNR פותר בעיה זו ע"י חלוקת האות למקטעים קצרים וחישוב ה-SNR בכל אחד מהם בנפרד. התוצאה הסופית היא ממוצע ה-SNR בכל המקטעים. באופן זה, השגיאה מחושבת באופן יחסי לאנרגיה בכל מקטע וכך מקטעים בעלי אנרגיה נמוכה תורמים באופן שווה למקטעים בעלי אנרגיה גבוהה. החישוב מתבצע לפי הנוסחה הבאה:

$$Segmental_SNR = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \left(\frac{\sum_{l=0}^{L-1} x_{mL+l}^2}{\sum_{l=0}^{L-1} (x_{mL+l} - x'_{mL+l})^2} \right)$$

כאשר x הוא האות המקורי, x' הוא האות המשוחזר ו- ML הוא אורך האות. האות מחולק ל- M מקטעים באורך L כל אחד. ערכים אופייניים ל- L הם בין 10ms ל-20ms.

לעתים מגבילים את הסכימה רק למקטעים בהם עוצמת האות לא נופלת מערך מסוים, וזאת כדי למנוע ארגומנטים קטנים של פונקצית הלוגריתם.

מקורות לעיון נוסף:

3. Rabiner Lawrence R., Schafer Ronald W., Digital Processing of Speech Signals, Prentice Hall, 1978
4. Goldberg Randy, Riek Lance, A Practical Handbook of Speech Coders, CRC Press, 2000
5. Barnwell Thomas P., Nayebi Kambiz, Richardson Craig H., Speech Coding – A Computer Laboratory Textbook, Wiley, 1996