© Technion - Israel Institute of Technology, Elyachar Central Library

הטכניון – מכון טכנולוגי לישראל
Technion – Israel Institute of Technology

ספריות הטכניון
The Technion Libraries

בית הספר ללימודי מוסמכים ע"ש ארווין וג'ואן ג'ייקובס
Irwin and Joan Jacobs Graduate School

©
All rights reserved

This work, in whole or in part, may not be copied (in any media), printed, translated, stored in a retrieval system, transmitted via the internet or other electronic means, except for "fair use" of brief quotations for academic instruction, criticism, or research purposes only.
Commercial use of this material is completely prohibited.

©
כל הזכויות שמורות

אין להעתיק (במדיה כלשהי), להדפיס, לתרגם, לאחסן במאגר מידע, להפיץ באינטרנט, חיבור זה או כל חלק ממנו, למעט "שימוש הוגן" בקטעים קצרים מן החיבור למטרות לימוד, הוראה, ביקורת או מחקר. שימוש מסחרי בחומר הכלול בחיבור זה אסור בהחלט.

הדגשת אותות דבור הטבולים ברעש

חבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר

דוקטור למדעים

מאת

יריב אפרים

2033489

הטכניון-מכון טכנולוגי לישראל
הפקולטה להנדסת חשמל
הפקולטה למדעי המחשב
ק ר י ה

000000866781

7.11.88

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

סיון תשמ"ד חיפה יוני 1984

© Technion - Israel Institute of Technology, Elyachar Central Library

המחקר נעשה בהנחית פרופסור דוד מלאך
במעבדה לעבוד אותות בפקולטה להנדסת חשמל בטכניון.

ברצוני להביע את תודתי והערכתי
לפרופסור דוד מלאך על הצעת הנושא,
הנחייתו המסורה ועזרתו הרבה
בכל שלבי המחקר.

כמו כן ברצוני להודות לפרופסורים
ישראל בר דוד, יעקב זיו ומשה זכאי
על שתמיד מצאו פנאי ליעץ לי
בנושאי העבודה.

תודה מיוחדת מגיעה לחברי,
מר שלמה שיץ, על השיחות הרבות
והפוריות בנושאי סיגנון ועל עזרתו
השוטפת.

לבסוף, ברצוני להודות למהנדס
המעבדה לעבוד אותות, מר יורם אור חן
ולתכנתת המעבדה, הגב' צפורה פורטנוי
על העזרה השוטפת בנושאי המעבדה.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

תוכן עניינים

	עמוד
תקציר	1
רשימת סמלים וקיצורים	2
פרק 1 - מבוא	4
1.1 - מהות הבעיה	4
1.2 - תכונות אות הדבור ומערכת השמע	7
1.3 - סקר ספרות	9
1.4 - סקירת העבודה ותוצאותיה	12
1.5 - תרומת המחקר	15
פרק 2 - הדגשת אותות דבור תוך שמוש במשערך השגיאה הריבועית הממוצעת המינימלית של האמפליטודה הספקטרלית לזמן קצר	19
פרק 3 - שלוב הדגשה וקדוד מסתגל בתחום התדר של אותות דבור רועשים	28
פרק 4 - דיון ומסקנות	32
נספח א - הדגשת אותות דבור תוך שימוש במשערך השגיאה הריבועית הממוצעת המינימלית של האמפליטודה הספקטרלית לזמן קצר	36
I - מבוא	38
II - המשערך האופטימלי של האמפליטודה הספקטרלית לזמן קצר	43
- גזירת משערך האמפליטודה	43
- אנליזת שגיאה ורגישות	49
III - המשערך האופטימלי תחת אי-וודאות בקיום האות	53
- גזירת משערך האמפליטודה	56
IV - המשערך האופטימלי במובן השגיאה הריבועית הממוצעת המינימלית של האקספוננט הקומפלקסי	59
- גזירת המשערך האופטימלי של האקספוננט הקומפלקסי	60
- משערך הפאזה האופטימלי	63
V - שערוך הווריאנסים של רכיבי התדר	64
- גישת הסבירות המירבית	64
- גישת "ההחלטה-המכוונת"	66
VI - תאור המערכת והערכת ביצועיה	69
- תאור המערכת	70
- הערכת הביצועים	71
VII - סיכום ודיון	75

© Technion - Israel Institute of Technology, Elyachar Central Library

תוכן העניינים (המשך)

עמוד

נספח A - הוכחת נוסחת משערך האמפליטודה. 77

נספח B - שערוך האמפליטודה בשיטת ההחסרה הספקטרלית הוקטורית. 77

נספח C - השלמה לסעיף III 84

נספח D - הוכחת נוסחאות מסעיף IV. 89

מקורות. 92

נספח ב - הדגשת דבור תוך שימוש במשערך השגיאה הריבועית הממוצעת המינימלית עבור לוגריתם האמפליטודה הספקטרלית 94

I - מבוא. 96

II - גזירת המשערך האופטימלי. 97

III - גזירת המשערך האופטימלי בתנאי אי-וודאות של קיום האות 102

IV - תאור המערכת והערכת הביצועים 103

- תאור המערכת 103

- הערכת ביצועים 104

V - סיכום ומסקנות 106

מקורות. 107

נספח ג - שערוך יחס האות לרעש למטרות הדגשת דבור תוך שימוש באלגוריתם של ויטרבי 108

I - מבוא. 109

II - שערוך MAP של ווארינס רכיב תדר של אות הדבור. 111

- שערוך MAP 113

- אלגוריתם ויטרבי 118

III - תאור המערכת והערכת הביצועים 120

- תאור המערכת 120

- הערכת ביצועים 122

IV - סיכום ומסקנות 122

מקורות. 124

תוכן העניינים (המשך)

	עמוד
נספח ד - שלוב הדגשה וקדוד מסתגל בתחום התדר של אותות דבור רועשים	125
I - מבוא	127
II - קדוד מסתגל בתחום התדר.	128
III - משערך השגיאה הריבועית הממוצעת המינימלית של האמפליטודה הספקטרלית.	133
IV - הערכת ביצועים.	136
V - דיון	137
מקורות	139
תקציר באנגלית	I
תוכן עניינים באנגלית	

© Technion - Israel Institute of Technology, Elyachar Central Library

ת ק צ י ר

מחקר זה עוסק בבעית ההדגשה של אותות דבור הטבולים ברעש אדיטיבי החסר קורלציה עם אות הדבור, בהינתן האות הרועש לבדו. הגישה הבסיסית שננקטה כאן מנצלת את החשיבות המרכזית שיש לאמפליטודה הספקטרלית לזמן קצר של אות הדבור בתפישה שלו ע"י מערכת השמע. אנו מפתחים משערכי שגיאה ריבועית ממוצעת מינימלית של האמפליטודה הספקטרלית לזמן קצר ושל הלוגריתם שלה ובוחנים אותם בהדגשת דבור. כמו כן אנו מרחיבים משערכים אלו בהתאם לעובדה שאות הדבור אינו נמצא כל העת במדידות הרועשות. המשערכים הנ"ל נגזרים על בסיס מודל סטטיסטי המנצל תכונות אסימפטוטיות של הרכיבים הספקטרליים. בפרט אנו מניחים שרכיבי התדר של אות הדבור ושל הרעש ניתנים ליצוג כמשתנים אקראיים גאוסיים בלתי תלויים סטטיסטית.

לשם שחזור האות המודגש תוך ניצול משערכי האמפליטודה הספקטרלית הנ"ל, אנו בוחנים את משערך השגיאה הריבועית הממוצעת המינימלית של האקספוננט הקומפלקסי של הפאזה לזמן קצר. אנו מראים שלמשערך המתקבל ערך מוחלט השונה מיחידה ולכן צרופו למשערך האמפליטודה הספקטרלית משפיע על שערוך אמפליטודה זו. מאידך, משערך השגיאה הריבועית הממוצעת המינימלית של האקספוננט הקומפלקסי, אשר הערך המוחלט שלו מאולץ להיות יחידה, ולכן גם אינו משפיע על שערוך האמפליטודה הספקטרלית, הוא האקספוננט הקומפלקסי של הפאזה הרועשת. מסיבה זו שחזור האות המודגש במערכת המוצעת נעשה תוך ניצול הפאזה הרועשת.

בעיית שערוך יחס האות לרעש של כל רכיב תדר, אשר נחוץ בישום משערכי האמפליטודה הספקטרלית הנ"ל, נחקרה בהרחבה בעבודה זו. אנו מציעים משערך סבירות מירבית, משערך "החלטה-מכוונת" ומשערך סבירות מירבית א-פוסטריורית. המשערך האחרון מיושם תוך שימוש באלגוריתם של ויטרבי.

המערכת המוצעת נבחנה בהדגשת אותות דבור הטבולים ברעש רחב סרט בעלי יחס אות לרעש של 5, 0, ו- ‎-5dB. מערכת זו משפרת באופן משמעותי את האיכות של האות הרועש, ע"י הנחתת רעש הרקע. כמו כן, הרעש הנותר נשמע אחיד ונמצא שהוא פחות מפריע ומרגיז מאשר "הרעש המוסיקלי" המתקבל במערכות אחרות להדגשת דבור. הסיבוכיות של המערכת המוצעת דומה לזו של מערכות אחרות הנמצאות בשימוש.

השיטה הנ"ל של הדגשת דבור הופעלה בהצלחה לשם שיפור האיכות של אות דבור משוחזר, המתקבל ממקדד מסתגל הפועל במישור התדר על אותות דבור רועשים. בישום זה האמפליטודה הספקטרלית לזמן קצר משוערכת בטרם נעשה הקידוד.

© Technion - Israel Institute of Technology, Elyachar Central Library

רשימת סימונים

A_k - אמפליטודת הרכיב הספקטרלי ה-k של אות הדבור.

$\hat{A}_k$ - משערך אופטימלי של A_k.

B_k - מס' סיביות המוקצות לרכיב הספקטרלי ה-k.

D_k - רכיב ספקטרלי k-י של הרעש.

$d(t)$ - תהליך הרעש.

$E\{\cdot\}$ - תוחלת.

$G(\cdot,\cdot)$ - פונקצית הגבר פרמטרית.

H_k^i - השערה ה-i ברכיב התדר ה- k.

$I_n(\cdot)$ - פונקצית בסל מסדר n-י.

$M(a;c;x)$ - confluent hypergeometric function

$p(\cdot)$ - פילוג סגולי.

$P[\cdot]$ - אופרטור חיוביות.

q_k - הסתברות ההשערה H_k^o.

R_k - אמפליטודת הרכיב הספקטרלי ה-k של האות הרועש.

T - אינטרוול אנליזה.

X_k - רכיב ספקטרלי k-י של אות הדבור.

$x(t)$ - אות דבור נקי מרעש.

$y(t)$ - אות דבור רועש.

α - פרמטר מיצוע.

α_k - פאזת הרכיב X_k.

θ_k - פאזת הרכיב Y_k.

γ_k - יחס אות לרעש א-פוסטריורי ברכיב הספקטרלי ה-k.

ξ_k - יחס אות לרעש א-פריורי ברכיב הספקטרלי ה-k.

© Technion - Israel Institute of Technology, Elyachar Central Library

$\lambda_x(k)$ - וואריינס הרכיב הספקטרלי X_k.

$\lambda_d(k)$ - וואריינס הרכיב הספקטרלי D_k.

$\Lambda(\cdot)$ - יחס סבירות מוכלל.

$\Phi(\cdot)$ - פונקצית השגיאה.

$\delta(\cdot)$ - פונקצית דירק.

$\Gamma(\cdot)$ - פונקצית גאמה.

ε_k - עוות ברכיב הספקטרלי ה-k.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

פרק 1 : מ ב ו א

1.1 מהות הבעיה

בעית ההדגשה של אותות דבור (speech enhancement) הטבולים ברעש מהווה מזה זמן רב נושא למחקרים רבים, הן בשל המספר הרב של ישומים והן בשל האתגר הטמון בפתרונה. כמו כן ההתקדמות הטכנולוגית שחלה לאחרונה, המאפשרת לממש בזמן אמיתי אלגוריתמים מורכבים, מאיצה אף היא את המחקר בכוון הנ"ל. המונח "הדגשת דבור" בו נשתמש בעבודה זו, הוא המונח המקובל בספרות לכל אותם הפעולות הנעשות במגמה לשפר היבטים תפישתיים (perceptual aspects) שונים של אות הדבור. פעולות אלו כוללות כמובן את הסינון שהוא המונח היותר מתאים בהקשר של עבודה זו. בהיבט היותר רחב נעשות פעולות הדגשה גם על אות דבור הנקי מרעש.

בעית ההדגשה של אותות דבור הטבולים ברעש מופיעה בהקשרים רבים ובצורות שונות, דבר ההופך אותה להיות כה מורכבת ורחבה [1]. היא מאופינת ע"י סוג הרעש והאופן בו הוא נלווה לאות הדבור. כמו כן המבנה המורכב של אות הדבור ושל מערכת השמע מיחדים בעיה זו מבעיות סינון אחרות. פתרון הבעיה איננו חד משמעי כפי שנראה בהמשך והוא תלוי בתנאי הבעיה ובמטרה אותה מעונינים להגשים. תנאי הבעיה מוגדרים ע"י סוג הרעש, היכן וכיצד הוא נלווה לאות הדבור, האם קיים מקור יחוס לרעש, האם ניתן לפעול על אות הדבור בטרם נלווה אליו הרעש וכו'. המטרה תהיה תלויה בהיבט התפישתי אותו מעונינים להדגיש.

הרעש הנלווה לאות הדבור יכול להיות רעש רחב סרט, מחזורי, קולו של דובר מתחרה, הדים אקוסטיים או חשמליים (בערוץ) הנובעים מקולו של הדובר, צרופים של הרעשים הנ"ל וכו'. הרעש יכול להתווסף לאות הדבור במקום היווצרותו (כגון במקרים של הדים אקוסטיים או של רעש רקע בלתי תלוי באות), בערוץ (כגון במקרים של הדים חשמליים בקווי טלפון), או במקלט. במקרה האחרון הרעש יכול להיות רעש טרמי או לחלופין רעש רקע הנוצר במקום הימצאותו של המאזין. עבור מקרה זה ניתן להביא את הדוגמא של פועלים הנמצאים בחדר מכונות והמקבלים הוראות מחדר שקט. האופן בו הרעש נלווה לאות הדבור יכול להיות דרך תהליך אדיטיבי, מולטיפליקטיבי, קונבולוציה וכו'. לדוגמא, רעש רקע הוא בד"כ אדיטיבי, רעש הנובע מדעיכה (fading) הוא רעש מולטיפליקטיבי ואילו רעש ההדים האקוסטיים הוא רעש קונבולוציה. התלות הסטטיסטית בין האות לרעש מהווה אף היא אחד מאפיוני הבעיה ומשפיעה על דרך פתרונה. למשל, רעש רקע הנובע מרעש רחוב יהיה בלתי תלוי סטטיסטית באות הדבור של שדרן הנמצא ברחוב זה, בעוד שרעש ההדים האקוסטיים יכול להיות תלוי סטטיסטית באות הדבור עצמו. בדוגמא אחרת, רעש הקוונטיזציה הנלווה לאות דבור דגום יהיה בלתי תלוי באות אם מספר רמות הקוונטיזציה הוא מספיק גדול. לעומת זאת אם מספר רמות הקוונטיזציה הוא נמוך, לרעש הרקע תהיה תלות סטטיסטית חזקה עם אות הדבור.

© Technion - Israel Institute of Technology, Elyachar Central Library

המטרה אותה מעונינים להשיג באמצעות ההדגשה היא שיפור בתפישה של אות הדבור הרועש. במובח תפישה מבחינים בשני מושגים: מובנות (intelligibility) ואיכות (quality). מובנות הדבור היא מדד אוביקטיבי כמותי הקשור למידת האינפורמציה הכלולה באות הדבור. ערכו של מדד זה מרמז על המידה הצפויה של הבנת טקסט ע"י מספר רב של מאזינים. לעומת זאת האיכות הינה מדד סוביקטיבי המצביע על המידה בה נוח או נעים להאזין לדבור. בעוד שמובנות אות הדבור נמדדת באחוזים ומוגדרת היטב, האיכות נמדדת במונחים של העדפה סוביקטיבית. היא בד"כ מתוארת ע"י האופי של הרעש ומידת הפרעתו למערכת השמע, האופי של אות הדבור עצמו וכו'. ראוי לציין ששני המדדים הנ"ל הם בד"כ אורתוגונליים. כלומר, אות דבור בעל מובנות גבוהה אינו בהכרח בעל איכות טובה וההפך. לדוגמא, העברת אות דבור דרך מסנן מעביר גבוהים (high pass filter) גורמת לשיפור במובנותו, אולם יחד עם זאת לירידה באיכותו. על הסיבות לכך נעמוד בהמשך.

רוב המערכות להדגשת דבור הקיימות משיגות שיפור באיכות האות הרועש אך לא במובנותו. יתרה מזו, לעתים קרובות השיפור באיכות מושג במחיר של ירידה במובנות. קיים מספר מצומצם ביותר של מערכות בהן מושג שיפור במובנות אך תוך ירידה חריפה באיכות. לשיפור האיכות חשיבות רבה כיוון שהוא מונע את התעייפות המאזין ובכך משפר בעקיפין את מובנות האות.

הדיון לעיל מרמז על קושי בסיסי בבעית ההדגשה של אותות דבור הטבולים ברעש. הוא גם מצביע על דרך טיפול שונה בה יש לנקוט במידה ומעונינים בשיפור באיכות או במובנות של האות הרועש. לאור קושי זה קיימת היות נטיה להסתפק במערכות אשר גורמות לשיפור באיכות אות הדבור הרועש תוך שמירה על מובנותו.

הפתרון לבעית הדגשת אותות דבור הטבולים ברעש יהיה כאמור מותאם לתנאי הבעיה ולאינפורמציה העומדת לרשותנו אודות האות והרעש. כל המערכות להדגשת דבור מנצלות במידה זו או אחרת תכונות של אות הדבור ושל מערכת השמע. על תכונות אלו נעמוד בפרוט בסעיף 1.2. קיימים מצבים בהם עומדת לרשותנו אינפורמציה נוספת פרט לאות הדבור הרועש, או שיש באפשרותנו לפעול על אות הדבור בטרם נלווה אליו הרעש. במקרים כאלו בעית ההדגשה היא קלה יותר ונצול אינפורמצית העזר מוביל לשיפור בטיב ההדגשה. בדוגמת הפועלים שהובאה לעיל ניתן למשל להשיג שיפור במובנות ביחס לזו של האות הרועש במידה ונעביר את אות הדבור הנקי דרך מסנן מעביר גבוהים. בדוגמא אחרת בה טייס השוהה באויר משדר למגדל פקוח, ניתן להשתמש במיקרופון משני שיקלוט בעיקר את הרעש הנמצא בתא הטייס. ניתן לנצל מקור יחוס זה להחסרת רעש המקור ע"י טכניקות אדפטיביות לביטול רעש (adaptive noise cancelling) [2]. ראוי לציין שבמקרה כזה ההדגשה חייבת להיעשות בתא הטייס ומבחינות מסוימות זהו חסרון.

© Technion - Israel Institute of Technology, Elyachar Central Library

עיקר הדיון עד כה נסב סביב הבעיה של שיפור האיכות ו/או המובנות של אות דבור הטבול ברעש בטרם השמעתו בפני מאזינים. קיימים היבטים נוספים לבעית ההדגשה של אותות דבור. אחד ההיבטים המרכזיים הוא זה של שיפור פעולתם של מקדדי אותות דבור הפועלים בתנאי רעש [3-5]. חשיבות הנושא נובעת מהשימוש ההולך וגובר בתקשורת ספרתית הן לשם שיפור האמינות והן לצורך הצפנת אותות דבור. כמו כן קיימת נטיה לאינטגרציה של מקדדי דבור עם רשתות תקשורת מחשבים לשם העברה או אחסון של אותות דבור. בעיקר מדובר על מקדדים לערוצים צרי סרט המתאימים להעברת מידע ספרתי בקצב של עד 2400 סיביות לשניה. כיוון שמקדדי אות דבור הפועלים בקצבים הנ"ל מתבססים במידה ניכרת על מודל ליצירת אות הדבור (מודל אוטורגרסיבי), פעולתם משתבשת כאשר אות הכניסה אליהם הוא אות רועש. הסיבה לכך היא שהמודל האוטורגרסיבי איננו תקף יותר עבור כניסה רועשת. נהוג לתקוף בעיה זו בשתי שיטות. בראשונה מתאימים את שיטת השערוך של פרמטרי המודל לנוכחות הרעש בכניסה [6-10], בשיטה השניה משתמשים במעבד-קדם שמתפקידו לסלק את הרעש בטרם מופעל המקדד [11-13]. החסרון העיקרי של השיטה הראשונה הוא בכך שבעית השערוך הופכת להיות בעיה לא ליניארית עבורה לא קיימים פתרונות פשוטים ומשביעי רצון. לעומת זאת לשיטה השניה יש יתרון בכך שהיא מאפשרת שימוש במקדדים נפוצים המתאימים לעבודה עם אות דבור נקי. נציין שבישום הנוכחי המטרה אותה מבקשים להגשים ע"י פעולת ההדגשה תהיה שיפור במובנות ובאיכות של אות הדבור במוצא המקדד ולאו דוקא במוצא מערכת ההדגשה.

היבטים אחרים של הבעיה להדגשת אותות דבור קשורים לפתוח עזרי שמע לאנשים כבדי שמיעה (לדוגמא ראה [14-16]), הסרת העוות מקולם של צוללנים המשדרים ממעמקי הים [17] וכו'.

הדיון לעיל ממחיש את היקף הבעיה של הדגשת אותות דבור הטבולים ברעש. בעבודה זו נטפל בהדגשת אותות דבור הטבולים ברעש אדיטיבי קוואזי-סטציונרי החסר קורלציה עם אות הדבור. כמו כן נניח שלרשותנו עומד האות הרועש בלבד. בפתוח האלגוריתם לא נגביל את הרעש להיות מסוג מסוים, אולם נבחן אותו בעיקר עבור רעש רחב סרט. רעש זה נחשב לבעייתי ביותר כיוון שהוא תוקף את מלוא רוחב הסרט של אות הדבור. הבעיה כפי שהוגדרה לעיל היא מספיק רחבה להקיף מקרים רבים, כיוון שהיא מניחה הנחות מינימליות אודות הרעש והדרך בו הוא נלווה לאות הדבור. איננו מניחים למשל קיום מקור יחוס לרעש או שניתן לפעול על אות הדבור בטרם נלווה אליו הרעש. בעיה זו טופלה רבות בספרות (כפי שנפרט בסעיף 1.3) ועדיין מהווה אתגר לחוקרים רבים.

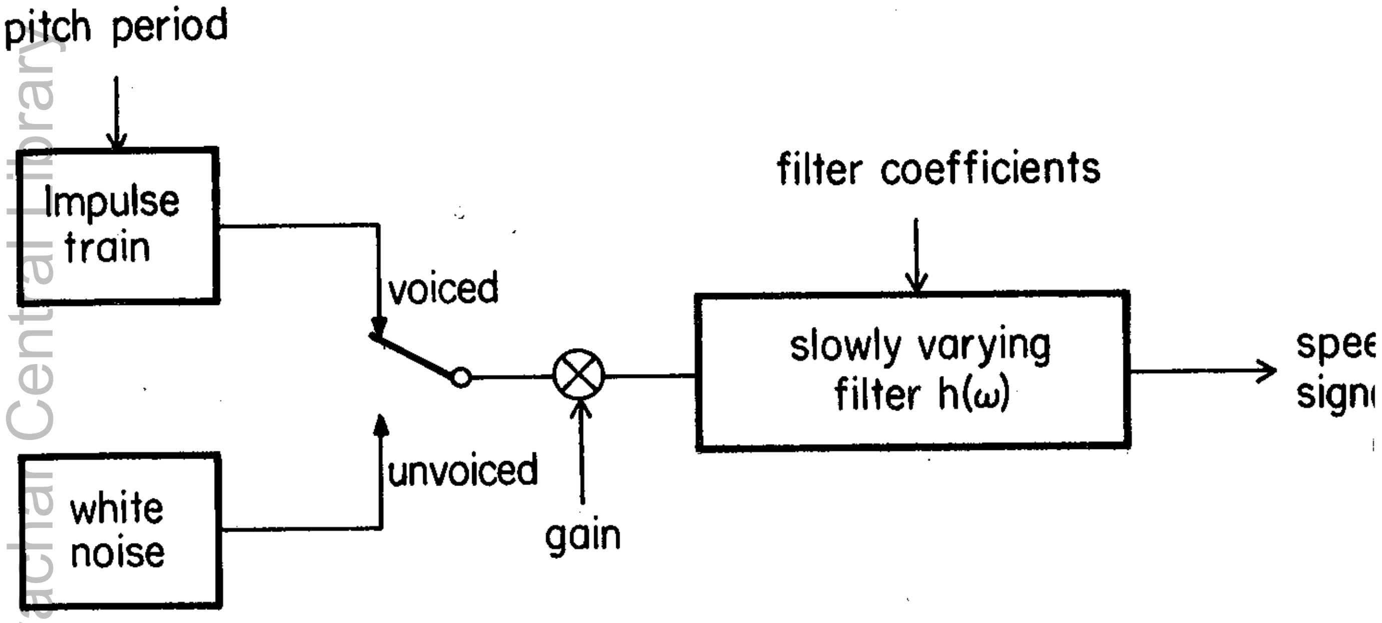
בהמשך פרק זה נתאר מספר תכונות חשובות של אות הדבור ושל מערכת השמע. כמו כן נסקור בקצרה עבודות קודמות שנעשו בתחום תוך הדגשת החומר הרלוונטי לעבודתנו. לבסוף נסקור את העבודה הנוכחית ונסכם בקצרה את תרומת המחקר.

1.2 תכונות אות הדבור ומערכת השמע

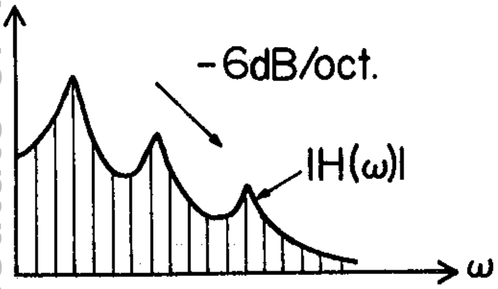
אות הדבור הוא תוצאה של ערור המעבר הקולי (vocal tract) ע"י דפקי אויר הנוצרים כתוצאה ממעבר אויר מהריאות דרך מיתרי הקול, או ע"י מערבולות אויר הנוצרות באילוצים צרים של המעבר הקולי. במקרה הראשון אות הדבור הנוצר הוא אות קולי (voiced) ובמקרה השני הוא אות אל-קולי (unvoiced). ציור 1.1 מתאר מודל מקובל להדגמת התהליך הנ"ל. דפקי האויר מיוצגים ע"י סדרת הלמים ואילו מערבולות האויר מיוצגות ע"י מקור הרעש הלבן. מחזור סדרת ההלמים הוא כמחזור התדר היסודי (pitch) של אות הדבור הקולי. המסנן הליניארי מייצג את המעבר הקולי, השפתיים, ועבור אותות דבור קוליים גם את מקור הערור (glottal source). הוא נתן לאפיון ע"י פונקצית תמסורת רציונלית כאשר קטביה מתארים את מערכת התהודה של המעבר הקולי. תדרי התהודה של המערכת נקראים פורמנטים.

לאור העובדה שהמערכת ליצירת אות דבור משתנה לאט בזמן, נהוג לייחד את הטיפול באותות דבור בקטעי זמן קצרים שאורכם 20-40 msec ושעבורם ניתן להתיחס למערכת הנ"ל כקבועה בזמן. על בסיס התבוננות זו נהוג למשל לאפיין בפרקי זמן קצרים אותות דבור קוליים כאותות הרמוניים עם מחזור יסודי השווה למחזור ה-pitch. כמו כן כאשר מייצגים את אות הדבור כתהליך אקראי, מתיחסים אליו כאל תהליך קוואזי-סטציונרי. כלומר רואים כל קטע קצר ממנו באורך 20-40 msec כחלק מתהליך סטציונרי. גם הטרמינולוגיה הקשורה לאותות דבור מבחינה במפורש בין אנליזה שנעשית על פני זמן קצר לבין זו שנעשית על פני זמן ארוך. כך למשל מבחינים בין אנליזה ספקטרלית לזמן קצר לבין אנליזה ספקטרלית לזמן ארוך. קיימת גם סיבה עמוקה יותר לטיפול באותות דבור לזמן קצר. סיבה זו קשורה בהוכחות כמעט וודאיות שקיימות היום ולפיהן מערכת השמע מבצעת אנליזה ספקטרלית לזמן קצר כבר בשלבים הראשונים של עבוד האות [18,19]. ציור 1.1-(b) מתאר באופן עקרוני ספקטרום לזמן קצר של אות דבור קולי ושל אות דבור אל-קולי. מהתאור לעיל אודות הדרך בה אות הדבור נוצר, ניתן להסיק שהעוטפת הספקטרלית לזמן קצר נקבעת ע"י המסנן הליניארי ואילו שאר פרטי הספקטרום נקבעים ע"י מקור הערור. ראוי לשים לב שהספקטרום של אות דבור קולי יורד בכ-6 dB/oct, בעוד שעיקר האנרגיה של אות אל-קולי מרוכזת בתדרים הגבוהים. לא נרחיב יותר את הדבור בנושא זה ונפנה את הקורא המתעניין למקורות המצויינים הבאים [19-21].

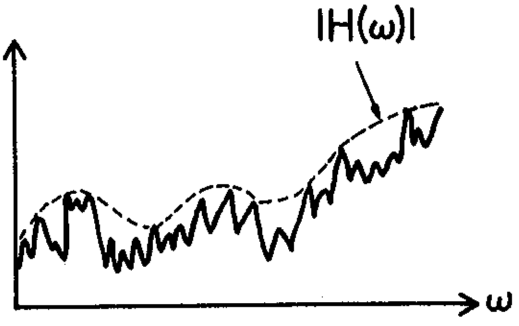
© Technion - Israel Institute of Technology, Elyachar Central Library



(a) - A speech production model



(b) - voiced



(c) - unvoiced

SPEECH SPECTRA

ציור 1.1: (a) - דיאגרמת בלוקים של המערכת ליצירת דיבור.
(b) - ספקטרום אופייני של אות דבור קולי.
(c) - ספקטרום אופייני של אות דבור אל-קולי.

Fig. 1.1: (a) - A block diagram of Speech Production.
(b) - Typical spectrum of voiced speech.
(c) - Typical spectrum of unvoiced speech.

© Technion - Israel Institute of Technology, Elyachar Central Library

ההיבטים הקשורים בתפישה של אות הדבור מורכבים הרבה יותר ופחות מובנים מתהליך יצירתו. נמנה כאן מספר עובדות חשובות: (1) - לאמפליטודה הספקטרלית (הערך המוחלט של התמרת פוריה) לזמן קצר, בפרט באזורי הפורמנטים, חשיבות מרכזית בתפישה (מובנות ואיכות) של אות הדבור. לפאזה לזמן קצר חשיבות משנית ואף זניחה [19]. (2) - רכיבי התדר בקרבת הפורמנט השני נושאים אינפורמציה רבה למרות שהם בעלי אנרגיה נמוכה מזו של רכיבי התדר הנמצאים בקרבת הפורמנט הראשון. לכן תרומתם למובנות אות הדבור היא רבה יותר. רכיבי התדר בקרבת הפורמנט הראשון תורמים בעיקר לאיכות אות הדבור, טבעיות הצליל, זהוי הדובר וכו' [22]. (3) - לעיצורים (consonants) חשיבות רבה למובנות למרות האנרגיה הנמוכה שלהם. (4) - למעברים בין פונמות (פונמה היא היחידה הבסיסית הקטנה ביותר של אות דבור, בדומה לאטום בכימיה) חשיבות רבה ביותר למובנות אות הדבור [20]. (5) - תרומת רכיבי התדר מעבר ל- 4 kHz למובנות אינה גדולה ודיכויים לא גורם לירידה גדולה באיכות אות הדבור. (6) - למערכת השמע היכולת למסך אות אחד ע"י אות אחר. ניתן לנצל תכונה זו למיסוך רעשים שאינם נוחים לאוזן ע"י רעשים אחרים נוחים יותר.

1.3 סקר ספרות

בסעיף זה נתאר בקצרה מערכות להדגשת דבור שהוצעו בספרות. המערכות שיתוארו מתאימות לבעיה בה נעסוק בעבודה הנוכחית. דגש יושם על המערכות הקשורות ישירות לגישה שננקטה בעבודה זו.

המערכת הפשוטה ביותר להדגשת אותות דבור הטבולים ברעש היא זו המנצלת מסנן מעביר נמוכים בעל רוחב סרט של 4 kHz. מסנן זה יסלק רכיבי רעש הנמצאים בתחום התדרים הפחות חשוב בתפישת הדבור.

במטרה להרחיב את העקרון הנ"ל, הוצע בעבר להשתמש במסנן וינר משתנה בזמן, המבוסס על שערוך הצפיפויות הספקטרליות של אות הדבור ושל הרעש [23-25]. בה בשעה שגישה זו נראית הגיונית היא אינה המתאימה ביותר להדגשת אותות דבור. הסיבה לכך היא שקריטריון השגיאה הריבועית הממוצעת המינימלית בין אות המקור לאות המשוערך לפיו נגזר מסנן וינר, איננו נמצא בקורלציה עם התפישה של אותות דבור ע"י מערכת השמע. עובדה ידועה היא ששגיאה ריבועית ממוצעת קטנה איננה מצביעה בהכרח על איכות או מובנות טובים. ההיפך גם נכון. כלומר שגיאה ריבועית ממוצעת גדולה אינה בהכרח אינדיקציה לכך שהאיכות או המובנות של אות הדבור ירודים. ע"מ להמחיש נושא חשוב זה נעיין בבעיה בה הרעש הנלווה לאות הדבור

© Technion - Israel Institute of Technology, Elyachar Central Library

הוא לבן. עקב הירידה של ספקטרום אות הדבור ב- 6 dB/oct, יחס האות לרעש בתדרים השונים ילך ויקטן ככל שעולים בתדר. מסנן וינר במקרה כזה יעביר כמעט ללא פגע את רכיבי התדר הנמצאים בקרבת הפורמנט הראשון ושלהם אנרגיה גבוהה, אך ידכא רכיבי תדר הנמצאים בקרבת הפורמנט השני להם אנרגיה נמוכה יחסית. התוצאה תהיה אות דבור המלווה ברעש שהספקו נמוך יותר מזה של אות הכניסה, אולם גם בעל מובנות נמוכה יותר. המובנות תקטן כי כפי שציינו קודם רכיבי התדר הנמצאים בקרבת הפורמנט השני חשובים יותר למובנות מאשר אלו הנמצאים בקרבת הפורמנט הראשון. הראינו אם כך דוגמא שבה הקטנת השגיאה הריבועית הממוצעת לא גוררת עליה במובנות. נראה כעת גם ששגיאה ריבועית ממוצעת נמוכה אינה בהכרח מדד לאיכות טובה. לשם כך נציין שלעתים נהוג להוסיף לאות המודגש רעש רחב סרט ע"מ למסך בעזרתו רעשים אחרים אשר מפריעים יותר לתהליך התפישה של אות הדבור. אין ספק שהוספת הרעש מגדילה את השגיאה הריבועית הממוצעת ואולם היא מובילה לאיכות טובה יותר. ניתן גם להביא דוגמאות נגד המורות ששגיאה ריבועית ממוצעת גדולה אינה מצביעה על כך שהמובנות או האיכות של האות המודגש ירודים. למשל העברת אות דבור דרך מסנן all-pass בעל תגובה להלם קצרה יחסית תגרום להגדלת השגיאה הריבועית הממוצעת בין מוצא המסנן לכניסתו, אולם שני האותות ישמעו כמעט זהים. הסיבה לתופעה זו היא שמערכת השמע אינה רגישה לפאזה לזמן קצר בעוד שהשגיאה הריבועית הממוצעת כן רגישה.

למרות כל האמור לעיל מסנן וינר נמצא היום בשימוש נרחב בהקשר של סינון אותות דבור הטבולים ברעש. הסיבה לכך נעוצה בעובדה שלמסנן זה (בניח הלא סיבתי) פונקצית תמסורת ממשית ולכן הוא משפיע על האמפליטודה הספקטרלית לזמן קצר של אות הדבור מבלי לשנות את הפאזה הרועשת. לכן קיימת נטיה בספרות לראות בפעולה זו כפעולה נכונה ורצויה, אשר עולה בקנה אחד עם ההיבט התפישתי לפיו האמפליטודה הספקטרלית לזמן קצר חשובה בעוד שהפאזה לזמן קצר אינה חשובה. הטעות בהילך מחשבה נפוץ זה היא בעובדה שמסנן וינר אינו משערך אופטימלי (גם בהנחות גאוסיות) של האמפליטודה הספקטרלית אלא של האות הזמני.

על בסיס קו המחשבה הנ"ל פותחה השיטה הפופולרית ביותר היום והידועה בשם "שיטת ההחסרה הספקטרלית" [24-26]. בשיטה זו משערכים את האמפליטודה הספקטרלית לזמן קצר ומצרפים לה את הפאזה של האות הרועש לשם שחזור האות המודגש. משערך האמפליטודה לזמן קצר מתקבל כשורש משערך הסבירות המירבית של הווארינס של כל אחד מרכיבי התדר של אות הדבור. משערך הווארינס נגזר תחת הנחות גאוסיות ומתקבל כהפרש שבין ריבוע הערך המוחלט של הרכיב הספקטרלי הרועש ומשערך ווארינס רכיב הרעש. משערך ווארינס הרעש מתקבל מקטעי שקט הסמוכים ביותר לקטע אות הדבור המעובד. מסיבה זו השיטה טובה בתנאי שהרעש הוא קוואזי-סטציונרי.

היתרון הגדול של שיטת ההחסרה הספקטרלית הוא פשטותה ובכך שהיא מהווה שיטת שיערוך לא פרמטרית. חסרונה העיקרי מתבטא בכך שהאות המודגש המתקבל בשיטה זו מלווה ברעש מוסיקלי חזק (ז.א. טונים שעוצמתם ותדרם משתנים בזמן) המפריע מאוד לתפישת הדבור. יחד עם זאת אות הדבור עצמו הוא די ברור. מהסיבות הנ"ל נעשו בספרות מאמצים גדולים על מנת להלבין את הרעש הנותר באות המודגש או לפחות להחליש את עוצמתו. בין היתר הוצעה השיטה של שימוש בטכניקת ה-spectral floor [27] שמעיקרה היא שיטה למיסוך הרעש המוסיקלי. שיטות אחרות שהוצעו היו בכיוון של שיפור משערך ווארינס האות ע"י שימוש במיצוע המשערכים מקטעי דבור סמוכים [26]. ב-[25] הוצעה שיטה נוספת המנצלת את העובדה שאות הדבור איננו קיים כל הזמן במדידות הרועשות על מנת להשיג הנחתה נוספת של הרעש המוסיקלי. תאור מפורט של שיטת וינר ושיטת ההחסרה הספקטרלית, כולל הגירסאות השונות המקובלות, ניתן למצוא ב-[24].

קיימות גישות נוספות להדגשת אותות דבור הטבולים ברעש אותן נזכיר כאן בקצרה כיוון שאינן קשורות ישירות לגישה שננקטה בעבודה זו. נתחיל בעבודות המנצלות את הייצוג ההרמוני לזמן קצר של אות הדבור הקולי [28-34]. ב-[28-30] מציעים לסנן את האות הרועש ע"י "מסנן מסרק". מסנן זה מהווה סדרה של מסננים מעבירי סרט שמרכזיהם נמצאים סביב ההרמוניות של אות הדבור. בשיטה זו, אותות אל-קוליים עוברים הנחתה מסוימת שנקבעת מראש. ב-[31] מוצע מימוש של מסנן המסרק מהסוג הנ"ל ע"י שימוש במסנן האדפטיבי לביטול רעש. ב-[32] מתוארת דרך להפרדת אות דבור של שני דוברים מאות המכיל את סכומם. בסיס ההפרדה הוא תדר ה-pitch השונה שיש לשני הדוברים. ראוי לציין ששלושת הגישות האחרונות דורשות ידיעה מדויקת של מחזור ה-pitch אותו קשה לשערך במהימנות מאות דבור רועש. ב-[33] מתואר שימוש של המסנן האדפטיבי לביטול רעש לשם הדגשת אותות דבור הטבולים ברעש רחב סרט ואילו ב-[34] מנצלים את המסנן האדפטיבי לסילוק הדים אקוסטיים הנלווים לאות דבור.

מחלקה אחרת של מערכות להדגשת דבור היא זו בה מנצלים את המודל ההנדסי ליצירת אות הדבור המתואר בציור 1.1. לפי גישה אחת משערכים את פרמטרי המודל מהאות הרועש ויוצרים בעזרתם מחדש את אות הדבור בתהליך של סינטזה. גישה זו ננקטה לראשונה ב-[35]. דיון מפורט בשערוך מקדמי המסנן ניתן למצוא ב-[6-10], שערוך תדר ה-pitch ב-[36] ואילו זהוי החלטות קוליות ב-[37]. גישה אחרת הוצעה ב-[6] בה מתבצע שערוך איטרטיבי של מקדמי המסנן ושל אות הדבור בקטע נתון. שערוך אות הדבור נעשה ע"י שימוש במסנן וינר בו הצפיפות הספקטרלית של אות הדבור מתקבלת באמצעות מקדמי המסנן המשוערכים.

© Technion - Israel Institute of Technology, Elyachar Central Library

כל המערכות שתוארו עד כה בסעיף זה (פרט ל-[33-34]) נבחנו ב-[23,24,30,38] לקביעת מידת שיפור המובנות המושגת בהן. פרט ל-[23] בה דווח על שיפור גבולי במובנות, נמצא ששאר המערכות אינן משפרות את המובנות על אות הדבור הרועש. לכל היותר מתקבל שמובנות האות המודגש זהה לזו של האות הרועש. אולם כל המערכות הנ"ל אכן משפרות את האיכות של האות הרועש ובכך חשיבותם (ראה סעיף 1.1). בספרות הוצעו שלוש מערכות המשיגות שיפור משמעותי במובנות האות הרועש [39-41]. מבין השלוש [41] היא המערכת המעשית אותה נתאר. במערכת זו מציעים להעביר את האות הרועש (רעש רחב סרט) דרך מסנן מעביר גבוהים ואת יציאתו דרך מגבל (Hard limiter). תפקיד המסנן הוא להדגיש את רכיבי התדר שבקרבת הפורמנט השני ע"י דיכוי רכיבי התדר שבקרבת הפורמנט הראשון. פעולה זו גורמת לאי מיסוך הרכיבים החשובים למובנות על ידי אלו שפחות חשובים. המגבל עושה פעולה דומה בתחום הזמן. הוא דואג להדגשת אותות חלשים כגון העיצורים ביחס לאותות חזקים כגון ההגאים. בעוד שהמערכת הנ"ל משפרת את המובנות היא גורמת לירידה באיכות.

1.4 סקירת העבודה ותוצאותיה

הגישה הבסיסית שננקטה בעבודה זו מתוארת בפרק 2. גישה זו מנצלת את החשיבות המרכזית שיש לאמפליטודה הספקטרלית לזמן קצר של אות הדבור בתהליך תפישתו ע"י מערכת השמע. אנו מציעים לראשונה לשערך באופן אופטימלי את האמפליטודה הספקטרלית לזמן קצר של אות הדבור מתוך האות הרועש הנתון בקטע שאורכו T. זאת בניגוד למשערכים המקובלים היום (וינר וההחסרה הספקטרלית) אשר אינם משערכי אמפליטודה ספקטרלית אופטימליים עבור המודל הסטטיסטי והקריטריון לפיהם הם נגזרו. תחת מודל סטטיסטי די כללי המנצל תכונות אסימפטוטיות (עבור $T \to \infty$) של רכיבי התדר, אנו גוזרים בעבודה זו את המשערך האופטימלי במובן השגיאה הריבועית הממוצעת המינימלית של האמפליטודה הספקטרלית. במודל הסטטיסטי הנ"ל אנו מניחים שרכיבי התדר של האות ושל הרעש ניתנים לייצוג כמשתנים אקראיים גאוסיים בלתי תלויים סטטיסטית. הנחת הגאוסיות מתבססת על משפט הגבול המרכזי בהיות כל אחד מרכיבי התדר סכום (או אינטגרל) משוקלל של משתנים אקראיים.

לצורך שיחזור אות הדבור תוך ניצול המשערך האופטימלי של האמפליטודה הספקטרלית, בחנו את השאלה הבאה: האם ניתן גם לשערך באופן אופטימלי, תחת אותו קריטריון ומודל סטטיסטי, את הפאזה לזמן קצר של אות הדבור? כאן הבחנו שלמעשה מספיק לשערך את האקספוננט הקומפלקסי של הפאזה ולא את הפאזה עצמה. בדרך זו

פישטבו לאין ערוך את הבעיה. מסתבר (כצפוי) שלא ניתן לשערך באופן אופטימלי ובו זמני הן את האמפליטודה הספקטרלית והן את האקספוננט הקומפלקסי. הסיבה לכך היא שלמשערך האופטימלי של האקספוננט הקומפלקסי ערך מוחלט השונה מיחידה. לכן צרוף משערך זה למשערך האופטימלי של האמפליטודה הספקטרלית, נותן משערך חדש לאמפליטודה הספקטרלית שכמובן איננו אופטימלי. כיוון שברצוננו להשתמש במשערך אמפליטודה ספקטרלית אופטימלי (עקב חשיבות האמפליטודה הספקטרלית ביחס לפאזה), בחנו את המשערך האופטימלי של האקספוננט הקומפלקסי תחת אילוץ שהערך המוחלט שלו ישווה ליחידה. כאן קבלנו שהמשערך האופטימלי המאולץ של האקספוננט הקומפלקסי זהה לאקספוננט הקומפלקסי של הפאזה הרועשת. אומנם המשערך הנ"ל של האקספוננט הקומפלקסי זהה לזה הנמצא בשימוש בשיטת ההחסרה הספקטרלית, אולם כאן הנחנו את הבסיס התאורטי לשימוש במשערך זה. עד כה מקובל היה לנמק את השימוש בפאזה הרועשת מסיבות של קושי במציאת משערך טוב יותר ובכך שממילא תרומת הפאזה לתפישה איננה גדולה [24,42].

מן הראוי לחזור ולהדגיש שפנינו לשערוך נפרד של האמפליטודה הספקטרלית ושל הפאזה, במקום לשערך את הרכיבים הספקטרליים במישרין, כי שערוך ישיר של הרכיבים הספקטרליים היה מחזיר אותנו לשערוך הוינרי. כפי שצויין בפרוט בסעיף 1.3, השערוך הוינרי איננו מתאים בהקשר של הדגשת אותות דבור, היות והקריטריון עליו הוא מבוסס איננו נמצא בקורלציה עם התפישה של אות הדבור. בעבודה זו השווינו את משערך האמפליטודה הספקטרלית האופטימלי שקבלנו עם משערך האמפליטודה הספקטרלית המתקבל משערוך וינרי. השוואה זו היא בעלת ענין כיוון ששני המשערכים נגזרים תחת אותו מודל גאוסי וכן מכיוון שמשערך וינר נמצא בשימוש נרחב במערכות להדגשת דבור. כמו כן בחנו את הרגישות של כל אחד מהמשערכים לאי דיוק בפרמטר של המודל הסטטיסטי (ז.א., יחס האות לרעש של כל רכיב תדר). מסתבר שהמשערך האופטימלי ומשערך וינר מתלכדים עבור יחסי אות לרעש גבוהים. אולם למשערך האופטימלי עדיפות רבה ביחסי אות לרעש נמוכים. תוצאה זו היא בעלת חשיבות ותומכת בהעדפת המשערך האופטימלי שקבלנו כאן על פני המשערך הוינרי. זאת מכיוון שרכיבי התדר החשובים למובנות אות הדבור הם דווקא אלו הנמצאים בקרבת הפורמנט השני והמאופיינים ע"י אנרגיה נמוכה. המסקנה הנ"ל אומתה גם בניסויי שמיעה לא פורמליים שבצענו במהלך עבודה זו.

עבודת מחקר זו החלה למעשה בנסיון לשפר את משערך האמפליטודה בשיטת ההחסרה הספקטרלית. לצורך זאת בחנו ניצול של משערך פאזת השגיאה בין רכיב התדר הרועש לרכיב המקורי ע"מ לשפר את שערוך האמפליטודה הספקטרלית. לשיטה שקבלנו קראנו שיטת "החסרה ספקטרלית וקטורית". מסתבר שמשערך זה מתלכד עם המשערך האופטימלי של האמפליטודה הספקטרלית.

© Technion - Israel Institute of Technology, Elyachar Central Library

בהמשך העבודה בחנו מספר הרחבות של משערך האמפליטודה הספקטרלית האופטימלי. ראשית גזרנו את המשערך תחת הנחת אי וודאות בקיום אות הדבור ברכיבים הספקטרליים הרועשים. לשם כך פיתחנו שני מודלים סטטיסטיים לאי קיום האות ברכיבי התדר הרועשים ובחנו אותם. אחד המודלים הוביל לשיפור משמעותי בתוצאות ההדגשה של אות דבור רועש. המודל השני היה פחות מוצלח בהקשר של הדגשת דבור ואולם הוא נמצא שימושי למטרות גילוי-מאוחר (post-detection) של קיום אות דבור. כל הנושאים הנ"ל מתוארים בפרוט בנספח א' של העבודה.

בעבודה זו בחנו גם את משערך האמפליטודה הספקטרלית המביא למינימיזציה של השגיאה הריבועית הממוצעת בין לוגריתם האמפליטודה הספקטרלית המקורית ולוגריתם משערך האמפליטודה הספקטרלית. השימוש בקריטריון זה נובע מהטענה שקריטריון שגיאה המבוסס על הפרש הלוגריתמים של האמפליטודות הספקטרליות נמצא בקורלציה גבוהה יותר עם התפישה של אות הדבור מאשר קריטריון המבוסס על הפרש האמפליטודות הספקטרליות עצמן [43]. נספח ב' של עבודה זו דן בפרוט בגזירת המשערך החדש ובהשוואתו למשערך הקודם. הסתבר שהמשערך החדש הוא מאוד יעיל בהדגשת אותות דבור הטבולים ברעש.

משערכי האמפליטודה הספקטרלית שנגזרו בעבודה זו תלויים בפרמטרי המודל הסטטיסטי עליו הם מבוססים. פרמטרים אלו הם הווריאנס של כל אחד מרכיבי התדר של הרעש וכן יחס האות לרעש של כל רכיב תדר. בעית השערוך של יחס האות לרעש נתגלתה כבעית מפתח ונמצאה קשה הרבה יותר לפתרון מזו של שערוך ווריאנס הרעש. בעיה זו טופלה בהרחבה בעבודה הנוכחית. בחנו כאן שלושה משערכים שנגזרו תחת הקריטריונים הבאים: maximum likelihood (ML), decision-directed (DD), ו-maximum a-posteriori - (MAP). מצאנו ששני המשערכים האחרונים הם מוצלחים במיוחד בבעיה הנוכחית ונותנים תוצאות דומות מאוד כאשר הם פועלים יחד עם משערך האמפליטודה הספקטרלית האופטימלי. במקרה זה מתקבלת הנחתה רציבית של רעש הרקע ואילו הרעש הנותר נשמע כרעש רחב סרט. זוהי תוצאה חשובה כאשר זוכרים שבמערכות ההדגשה הקיימות מתקבל רעש נותר מוסיקלי המפריע מאוד לתפישה של אות הדבור. מבחינה מעשית ניתן לראות את המשערך הפשוט של ה-DD כתחליף למשערך ה-MAP המורכב הרבה יותר אולם גם המבוסס יותר. תאור מפורט של משערך ה-MAP ומימושו ע"י האלגוריתם של ויטרבי ניתן למצוא בנספח ג' של עבודה זו.

הנושא האחרון שטופל במסגרת עבודה זו קשור בשיפור פעולתו של מקדד אותות דבור הפועל בשיטת ה-(ATC) Adaptive transform coding. כאן בחנו את האפשרות של הקטנת הספק רעש הקוונטיזציה תוך ביצוע פעולות הדגשה על יציאת ה-quantizer. כמו כן בחנו את המקדד בפעולתו על אותות דבור רועשים. בשני המקרים אנו מציעים לשערך באופן אופטימלי את האמפליטודה הספקטרלית של רכיבי התדר, כאשר כאן

ההתמרה היא מסוג cosine-transform. ביצוע פעולות ההדגשה גרר שיפור באיכות האות המקודד רק עבור המקרה בו אות הכניסה למקדד היה רועש. נושא זה נדון בפרוט בפרק 3 ובנספח ד' של העבודה.

1.5 תרומת המחקר

תרומת העבודה הנוכחית לפתרון בעית הדגשת אותות דבור הטבולים ברעש מתבטאת בשני מישורים, התאורטי והמעשי. במישור התאורטי עסקנו בשערוך אופטימלי של האמפליטודה הספקטרלית לזמן קצר של אות הדבור שהיא בעלת חשיבות מרכזית בתהליך התפישה שלו ע"י מערכת השמע. כאן בחנו את המשערך האופטימלי במובן השגיאה הריבועית הממוצעת המינימלית, הן כאשר השגיאה היתה בין האמפליטודות הספקטרליות והן כאשר היא היתה בין הלוגריתמים שלהם. התוצאות שקיבלנו הכלילו תוצאות אחרות שהתקבלו בספרות כגון את המשערך המקובל של וינר. בנוסף פתרנו את השאלה שהיתה פתוחה עד כה ביחס לשערוך הפאזה לזמן קצר של אות הדבור. לבסוף הצענו משערכים מוצלחים ליחס אות לרעש של רכיבי התדר של אות הדבור שהיו נחוצים לשם ישום משערכי האמפליטודה הנ"ל.

במישור המעשי אנו מציעים אלגוריתם להדגשת אותות דבור הטבולים ברעש אדיטיבי וחסר קורלציה המשפר באופן ניכר את איכות האות הרועש. השוני הבסיסי בינו לבין אלגוריתמים קיימים הוא בכך שבאלגוריתם המוצע הרעש הנותר הוא לבן ולא מוסיקלי ולכן מפריע הרבה פחות למערכת השמע. הסיבוכיות של האלגוריתם המוצע אינה גדולה מזו של האלגוריתמים המקובלים האחרים. כמו כן הצענו דרך לשפר את פעולתו של מקדד אותות דבור הפועל בשיטת ה-ATC בתנאי רעש.

© Technion - Israel Institute of Technology, Elyachar Central Library

References

[1] J.S. Lim, ed., Speech Enhancement, Prentice-Hall Signal Processing Series, 1983.

[2] B. Widrow et al., "Adaptive Noise Cancelling: Principles and Applications", Proc. IEEE, Vol. 63, pp. 1692-1716, December 1975.

[3] M.R. Sambar, N.S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantization Noise or Additions White Noise", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-24, pp. 488-494, Dec. 1976.

[4] S.M. Kay, "The Effects of Noise on the Autoregressive Spectral Estimator", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-27, pp. 478-485, Oct. 1979.

[5] J. Tierney, "A Study of LPC Analysis of Speech in Additive Noise", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-28, pp. 389-397, Aug. 1980.

[6] J.S. Lim, A.V. Oppenheim, "All-Pole Modeling of Degraded Speech", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-26, pp. 197-210, June 1978.

[7] S.M. Kay, "Noise Compensation for Autoregressive Spectral Estimators", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-28, pp. 292-303, June 1980.

[8] W.J. Done and C.K. Rushforth, "Estimating the Parameters of a Noisy All-Pole Process using Pole-Zero Modeling", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 228-231, 1979.

[9] B.R. Musicus and J.S. Lim, "Maximum Likelihood Parameter Estimation of Noisy Data", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 224-227, April 1979.

[10] Y. Grenier, K. Bry, J. LeRoux and M. Sulpis, "Autoregressive Models for Noisy Speech Signals", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1093-1096, 1981.

[11] R.P. Preuss, "A Frequency Domain Noise Cancelling Preprocessor for Narrowband Speech Communication Systems", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 212-215, 1979.

[12] S. Maitra, "Reducing the Effect of Background Noise for Low-Bit-Rate Voice Digitizers", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 696-698, 1980.

[13] C.K. Un, K.Y. Choi, "Improved LPC Analysis of Noisy Speech by Autocorrelation Subtraction Method", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1082-1085, 1981.

[14] O.M.M. Mitchell, C.A. Ross and G.H. Yates, "Signal Processing for a Cocktail Party Effect", J. Acoust. Soc. Amer., Vol. 50, No. 2 (part 2), pp. 656-660, August 1971.

[15] S.G. Knorr, "A Hearing Aid for Subjects with Extreme High-Frequency Losses", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-24, No. 6, pp. 473-480, Dec. 1976.

[16] P. Yanick and H. Drucker, "Signal Processing to Improve Intelligibility in the Presence of Noise for Persons with a Ski-Slope Hearing Impairment", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-24, No. 6, pp. 507-512, Dec. 1976.

© Technion - Israel Institute of Technology, Elyachar Central Library

[17] M.A. Richards, "Helium Speech Enhancement Using the Short-Time Fourier Transform", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-30, No. 6, pp. 841-853, Dec. 1982.

[18] M.R. Schroeder, "Models of Hearing", Proc. IEEE, Vol. 63, pp. 1332-1350, Sept. 1975.

[19] J.L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd ed., New York, Springer-Verlag, 1972.

[20] G. Fant, Acoustic Theory of Speech Production. The Hague, The Netherlands: Lexington, MA, Mouton, 1970.

[21] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals. Englewood Cliffs, N.J.: Prentice Hall, 1978.

[22] A. Agrawal and W.C. Lin, "Effect of Voiced Speech Parameters on the Intelligibility of PB Words", J. Acoust. Soc. Amer., Vol. 57, pp. 217-222, Jan. 1975.

[23] A. Feit, "Intelligibility Enhancement of Noisy Speech Signals", M.Sc. Thesis, Technion, Haifa, July 1973.

[24] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", Proc. IEEE, Vol. 67, pp. 1586-1604, Dec. 1979.

[25] R.J. McAulay and M.L. Malpass, "Speech Enhancement using a Soft-Decision Noise Suppression Filter", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-28, pp. 137-145, April 1980.

[26] S.F. Boll, "Suppression of Acoustics Noise in Speech using Spectral Subtraction", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-27, pp. 113-120, April 1979.

[27] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 208-211, 1979.

[28] R.H. Frazier, S. Samsam, L.D. Braida, A.V. Oppenheim, "Enhancement of Speech by Adaptive Filtering", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 251-253, 1976.

[29] Y.M. Perlmutter, L.D. Braida, R.H. Frazier and A.V. Oppenheim, "Evaluation of a Speech Enhancement System", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 212-215, 1977.

[30] J.S. Lim, A.V. Oppenheim and L.D. Braida, "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-26, pp. 354-358, Aug. 1978.

[31] M.R. Sambar, "Adaptive Noise Cancelling for Speech Signals", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-26, pp. 419-423, Oct. 1978.

[32] T.W.. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection", J. Acoust. Soc. Amer., Vol. 60, pp. 911-918, Oct. 1976.

[33] A. Nehorai, Adaptive Filtering of Speech Signals from Noise, M.Sc. Thesis, Technion, Haifa, Aug. 1979.

[34] Y. Ephraim and D. Malah, "Adaptive Speech Signals Dereverberation", in the 11th Convention of Electrical and Electronics Engineers, Tel-Aviv, Oct. 1979.

[35] N.J. Miller, "Recovery of Singing Voice from Noise by Synthesis", Thesis Tech. Rep. ID VTEC-CSC-74-013, May 1973, Univ. Utah, Computer Science Library, Salt Lake City, UT.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

[36] J.D. Wise, J.R. Caprio and T.W. Parks, "Maximum Likelihood Pitch Estimation", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-24, pp. 418-423, Oct. 1976.

[37] R.J. McAulay, "Optimum Speech Classification and its Application to Adaptive Noise Cancelling", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 425-428, April 1977.

[38] J.S. Lim, "Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive White Noise", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-26, pp. 471-472, Oct. 1978.

[39] H. Drucker, "Speech Processing in a High Ambient Noise Environment", IEEE Trans. Audio Electroacoust., Vol. AU-16, pp. 165-168, June 1968.

[40] R.J. Niederjohn and J.H. Grotelueschen, "The Enhancement of Speech Intelligibility in High Noise Levels by High-Pass Filtering Followed by Rapid Amplitude Compression", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-24, pp. 277-282, Aug. 1976.

[41] I.B. Thomas and A. Ravindran, "Intelligibility Enhancement of Already Noisy Speech Signals", J. Audio Eng. Soc. Vol. 22, pp. 234-236, May 1974.

[42] D.L. Wang and J.S. Lim, "The Unimportance of Phase in Speech Enhancement", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-30, No. 4, pp. 679-681, Aug. 1982.

[43] R.M. Gray, A. Buzo, A.H. Gray, Y. Matsuyama, "Distortion Measures for Speech Processing", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-28, pp. 367-376, Aug. 1980.

פרק 2

הדגשת אותות דבור תוך שימוש במשערך השגיאה הריבועית הממוצעת המינימלית של האמפליטודה הספקטרלית לזמן קצר

בפרק זה נבסח את בעית השערוך של האמפליטודה הספקטרלית לזמן קצר של אות הדבור, נפרט את ההנחות עבורן אנו פותרים בעיה זו ונציג את המשערכים השונים שהתקבלו בעבודה. משערכים אלה כוללים את משערכי השגיאה הריבועית הממוצעת המינימלית של האמפליטודה הספקטרלית ושל הלוגריתם שלה. כמו כן נביא את ההרחבות של משערכים אלו כאשר אי הוודאות בקיום האות ברכיבי התדר הרועשים נלקחת בחשבון. נדון גם בבעית השערוך האופטימלי של האקספוננט הקומפלקסי של הפאזה לזמן קצר ובצרוף המשערך המתקבל למשערך האופטימלי של האמפליטודה הספקטרלית לזמן קצר. בהמשך נדון בשערוך יחס האות לרעש של כל רכיב תדר המהווה פרמטר של המשערכים הנ"ל. לבסוף נתאר את המערכת להדגשת דבור בה יושמו משערכי האמפליטודה הספקטרלית וכן את ביצועיה. כל הנושאים הנ"ל נדונים בהרחבה בנספחים א'-ג' של עבודה זו ולכן נביא כאן את עיקרי הדברים בלבד.

בעית השערוך של האמפליטודה הספקטרלית לזמן קצר של אות הדבור, מנוסחת כאן כבעית השערוך של האמפליטודה של כ"א ממקדמי פרוק פוריה (הקומפלקסים) של אות הדבור $\{x(t)\ ,\ 0 \le t \le T\}$, בהנתן האות הרועש $\{y(t)\ ,\ 0 \le t \le T\}$. אנו מניחים שמקדמי פרוק פוריה של כ"א מהתהליכים (אות הדבור והרעש) ניתנים ליצוג כמשתנים אקראיים גאוסיים בלתי תלויים סטטיסטית. מודל זה מנצל תכונות אסימפטוטיות (עבור $T \to \infty$) של מקדמי פרוק פוריה. במיוחד אנו מנצלים את העובדה שכל מקדם פרוק הוא סכום (או אינטגרל) משוקלל של משתנים אקראיים הנובעים מדגמי התהליך המתאים. לכן בהסתמך על משפט הגבול המרכזי ובתנאי ש-T מספיק גדול, ניתן לייצג כל מקדם פוריה כמשתנה אקראי בעל פילוג סגולי גאוסי. העובדה שמשפט זה תקף תחת תנאים די כלליים [1,2] עבור משתנים אקראיים התלויים סטטיסטית, אשר יכולים להיות בעלי פילוגים סגוליים שונים, מעודדת את השימוש במודל זה בבעיה הנדונה. כמו כן אנו מנצלים את העובדה שעבור T מספיק גדול מקדמי פרוק פוריה הם חסרי קורלציה [3]. תחת המודל הגאוסי חוסר קורלציה זו שקולה להנחת האי תלות הסטטיסטית.

© Technion - Israel Institute of Technology, Elyachar Central Library

נסמן ב- $X_k \triangleq A_k \exp(j\alpha_k)$, D_k , $Y_k \triangleq R_k \exp(j\theta_k)$ את הרכיב הספקטרלי ה-k של אות הדבור, הרעש והאות הרועש, בהתאמה, בקטע $[0,T]$. תחת המודל הנ"ל אנו מקבלים שהמשערך האופטימלי במובן השגיאה הריבועית הממוצעת המינימלית של האמפליטודה A_k נתון ע"י:

$$
\begin{aligned}
\hat{A}_k &= E\{ A_k \mid y(t) \, , \, 0 \le t \le T \} \\
&= E\{ A_k \mid Y_o, Y_1, \dots \} \\
&= E\{ A_k \mid Y_k \} \\
&= \Gamma(1.5) \frac{\sqrt{v_k}}{\gamma_k} M(-0.5 \, ; \, 1 \, ; \, -v_k) R_k
\end{aligned}
\tag{2.1}
$$

כאשר המעבר מההתניה ב- $\{y(t) \, , \, 0 \le t \le T\}$ להתניה ברכיבי התדר $\{Y_o, Y_1, \dots\}$ תקף עבור המודל הסטטיסטי שהנחנו, כמוסבר ב-[4]. $\Gamma(1.5) = \sqrt{\pi}/2$ היא פונקצית גאמה ו- $M(a \, ; \, c \, ; \, x)$ היא ה- confluent hypergeometric function.

v_k מוגדר ע"י:

$$v_k \triangleq \frac{\xi_k}{1 + \xi_k} \gamma_k \tag{2.2}$$

כאשר ξ_k ו- γ_k מוגדרים ע"י:

$$\xi_k \triangleq \frac{E\{|X_k|^2\}}{\lambda_d(k)} \tag{2.3}$$

$$\gamma_k \triangleq \frac{R_k^{\,2}}{\lambda_d(k)} \tag{2.4}$$

ו- $\lambda_d(k)$ הוא ווארינס הרכיב ה-k של הרעש,

$$\lambda_d(k) \triangleq E\{|D_k|^2\} \tag{2.5}$$

ξ_k ו- γ_k בקראים יחס האות לרעש הא-פריורי והא-פוסטריורי בהתאמה.

© Technion - Israel Institute of Technology, Elyachar Central Library

עבור יחס אות לרעש גבוה המאופין ע"י $\xi_k >> 1$, ניתן להראות ש-:

$$\hat{A}_k \cong \frac{\xi_k}{1+\xi_k} R_k \qquad (2.6)$$

משערך זה נקרא משערך אמפליטודה ויגרי, כיוון שהוא מהווה את הערך המוחלט של המשערך הוינרי של הרכיב הספקטרלי X_k.

בנספח א' (ציור 1) אנו מתארים את המשערכים (2.1) ו-(2.6) ע"י עקומי הגבר פרמטרים, כאשר ההגבר מוגדר ע"י:

$$G(\xi_k, \gamma_k) \triangleq \frac{\hat{A}_k}{R_k} \qquad (2.7)$$

כמו כן התנהגות פונקצית הגבר זו מתוארת בפרוט בנספח זה. שני המשערכים (2.1) ו-(2.6) מושווים בעבודה זו על בסיס השגיאה הריבועית הממוצעת וההטיה של כ"א מהם. כמו כן נבחנת רגישותם לאי-דיוק בשערוך יחס האות לרעש ξ_k הדרוש לשם מימושם. השוואה זו נעשית עבור יחסי אות לרעש נמוכים בלבד, כיוון שכפי שראינו שני המשערכים מתלכדים ביחסי אות לרעש גבוהים. אנליזה שגיאה זו מורה על עדיפות ברורה למשערך האופטימלי הנותן שגיאה ריבועית ממוצעת והטיה הקטנים באופן ניכר מאלו המתקבלים במשערך הוינרי. תוצאות אנליזה זו מתוארים בציורים 2 ו-3 של נספח א'.

לצורך שחזור אות הדבור תוך ניצול משערך האמפליטודה הספקטרלית שגזרנו לעיל, נבחן כעת את המשערך האופטימלי של האקספוננט הקומפלקסי $\exp(j\alpha_k)$ ונדון בצרופו למשערך האמפליטודה (2.1). משערך זה מתקבל בדומה ל-(2.1) והוא נתון ע"י:

$$\begin{aligned} e^{j\tilde{\alpha}_k} &= E\{ e^{j\alpha_k} \mid y(t)\ ,\ 0 \le t \le T \} \\ &= E\{ e^{j\alpha_k} \mid Y_o, Y_1, \ldots \} \\ &= E\{ e^{j\alpha_k} \mid Y_k \} \\ &= E\{ e^{-j\phi_k} \mid Y_k\}\, e^{j\theta_k} \\ &\triangleq (\widetilde{\cos}\,\phi_k - \widetilde{\sin}\,\phi_k)\, e^{j\theta_k} \\ &= \Gamma(1.5)\sqrt{v_k}\, M(0.5\ ;\ 2\ ;\ -v_k)\, e^{j\theta_k} \end{aligned} \qquad (2.8)$$

© Technion - Israel Institute of Technology, Elyachar Central Library

כאשר $\phi_k \triangleq \theta_k - \alpha_k$. $\widetilde{\cos}\ \phi_k$ הוא המשערך האופטימלי של ϕ_k בהנתן Y_k. קל להראות שבמקרה הנדון $\widetilde{\sin}\ \phi_k$ (המוגדר באופן דומה) שווה לאפס. צרוף המשערך הנ"ל למשערך האמפליטודה $\hat{A}_k$ הנתון ב-(2.1), נותן את המשערך הבא עבור הרכיב הספקטרלי ה-k של אות הדבור.

$$\tilde{X}_k = \hat{A}_k\ \widetilde{\cos}\ \phi_k\ e^{j\theta_k} \qquad (2.9)$$

הערך המוחלט של המשערך הנ"ל מייצג כעת משערך אמפליטודה חדש שאינו אופטימלי כיוון ש-$\hat{A}_k$ הוא אופטימלי. לכן שימוש במשערך האקספוננט הקומפלקסי (2.8) אינו מתאים למטרותנו, כיוון שאנו מעוניינים בראש ובראשונה בשערוך אופטימלי של האמפליטודה.

ע"מ להמנע מהבעיה הנ"ל, בחנו את המשערך של האקספוננט הקומפלקסי $\exp(j\alpha_k)$ תחת האילוץ שהערך המוחלט שלו ישווה ליחידה. לשם כך פתרנו את בעית האופטי-מיזציה הבאה:

$$\min_{e^{j\hat{\alpha}_k}} E\{|e^{j\alpha_k} - e^{j\hat{\alpha}_k}|^2\} \qquad (2.10)$$
$$\text{subject to : } |e^{j\hat{\alpha}_k}| = 1$$

קבלנו שהמשערך האופטימלי המאולץ הנ"ל הוא האקספוננט הקומפלקסי של הפאזה הרועשת. כלומר,

$$e^{j\hat{\alpha}_k} = e^{j\theta_k} \qquad (2.11)$$

בנספח א' אנו מרחיבים את הדיון הנ"ל. בין היתר אנו מראים שמשערך הרכיב הספקטרלי (2.9) קרוב למשערך הוינרי, כאשר $\hat{A}_k$ ב-(2.9) הוא המשערך האופטימלי הנתון ב-(2.1). מכאן אנו מגיעים למסקנה ששיפור שערוך האקספוננט הקומפלקסי (בהשוואה לשימוש בפאזה הרועשת) משפר את שערוך צורת הגל אך יחד עם זאת פוגע בשערוך האמפליטודה הספקטרלית.

שתי הרחבות מוצלחות של משערך האמפליטודה הספקטרלית נדונות בעבודה זו. ראשית בחנו את המשערך האופטימלי של האמפליטודה הספקטרלית הלוקח בחשבון את אי הוודאות בקיום האות ברכיבי התדר הרועשים. תחת מודל סטטיסטי לפיו ההופעה של האות ברכיבי התדר הרועשים היא אקראית ובלתי תלויה עבור רכיבי התדר השונים, אנו מקבלים:

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

$$\hat{A}_k = \frac{\Lambda(Y_k,q_k)}{1+\Lambda(Y_k,q_k)} E\{ A_k \mid Y_k,H_k^1 \} \tag{2.12}$$

כאשר:

$$\Lambda(Y_k,q_k) \triangleq \frac{1-q_k}{q_k}\frac{p(Y_k|H_k^1)}{p(Y_k|H_k^o)} = \frac{1-q_k}{q_k}\frac{\exp(v_k)}{1+\xi_k} \tag{2.13}$$

והוא יחס הסבירות המוכלל. $p(Y_k|H_k^1)$ ו-$p(Y_k|H_k^o)$ הם בהתאמה הפילוגים הסגוליים של Y_k בהנתן ההשערה של קיום ואי-קיום האות ברכיב התדר Y_k. q_k מסמן את ההסתברות לאי קיום האות ברכיב התדר Y_k. $E\{A_k|Y_k,H_k^1\}$ הוא המשערך האופטימלי של A_k תחת ההשערה שהאות קיים באופן וודאי במדידה Y_k. למעשה זהו המשערך שקבלנו ב-(2.1). המשערך (2.12) מתואר בציור 4 של נספח א' ע"י משפחת עקומים פרמטריים. כמו כן התנהגות עקומים אלו מוסברת שם ומושווית עם זו של העקומים המתאימים למשערך (2.1). שים לב שבציור 4 הנ"ל η_k מסמן את יחס האות לרעש הא-פריורי ולא ξ_k. ההבדל בין השנים נובע מכך ש-ξ_k מתאים למצב בו מניחים שהאות קיים בוודאות במדידות. קל להראות שהקשר בין שני גדלים אלו נתון ע"פ:

$$\eta_k = (1-q_k)\ \xi_k \tag{2.14}$$

ההרחבה השניה של משערך האמפליטודה הספקטרלית נדונה בנספח ב' של עבודה זו. כאן אנו גוזרים משערך אופטימלי תחת הקריטריון הבא:

$$\min_{\hat{A}_k} E\{(\log A_k - \log \hat{A}_k)^2\} \tag{2.15}$$

המוטיבציה לניצול קריטריון זה נובעת מהצלחת השימוש בקריטריון דומה למטרות אנליזה וזיהוי של אותות דבור. בעבודה זו אנו בוחנים לראשונה קריטריון זה עבור בעית ההדגשה של אותות דבור הטבולים ברעש. ניתן להראות שהמשערך הנובע מ-(2.15) נתון ע"י:

$$\hat{A}_k = \exp\{ E(\ln A_k \mid Y(t),\ 0 \le t \le T)\}$$

$$= \exp\{ E(\ln A_k \mid Y_o, Y_1, \ldots \qquad)\}$$

$$= \exp\{ E(\ln A_k \mid Y_k)\}$$

$$= \frac{\xi_k}{1+\xi_k} \exp\{ \frac{1}{2} \int_{v_k}^{\infty} \frac{e^{-t}}{t}\, dt \} R_k \qquad (2.16)$$

האינטגרל המופיע ב-(2.16) ידוע כאינטגרל האקספוננציאלי של v_k והוא ניתן לחישוב יעיל ע"י שימוש בקרוב פולינומיאלי מתאים [5]. עקומי ההגבר הנובעים מהמשערך (2.16) מתוארים בציור 1 של נספח ב'. תכונה בולטת של עקומים אלו היא שהם תמיד נמוכים מעקומי ההגבר המתאימים למשערך (2.1). כלומר המשערך (2.16) נותן תמיד הגבר נמוך יותר מאשר המשערך (2.1). קל להוכיח תכונה זו באמצעות אי השוויון של ינסן.

ההרחבה של המשערך (2.16) למקרה בו אי הוודאות בקיום האות ברכיבים הספקטרליים הרועשים נלקחת בחשבון נבחנה בנספח ב'. היות והמשערך המתקבל במקרה זה (עם $q_k > 0$) לא היה מוצלח במיוחד בהדגשת אותות דבור, לא נפרטו כאן.

לצורך ישום משערכי האמפליטודה (2.1), (2.12) ו-(2.16) במערכת מעשית להדגשת דבור, יש צורך לדעת את יחס האות לרעש הא-פריורי של כל רכיב תדר וכן את ווארינס הרעש בכל רכיב תדר. כיוון שבבעיה הנוכחית אותה אנו תוקפים הנחנו שלרשותנו עומד האות הרועש בלבד, יש צורך בשערוך גדלים אלו. בעבודה זו עסקנו בעיקר בבעיה היותר קשה של שערוך יחס האות לרעש הא-פריורי ואילו את ווארינס הרעש בכל רכיב תדר שערכנו פעם אחת בלבד מתוך קטע רעש התחלתי.

מכיוון שאות הדבור אינו סטציונרי, יחס האות לרעש הא-פריורי של כל רכיב תדר משתנה בזמן ולכן עלינו לשערכו מחדש בכל מסגרת אנליזה. בעבודה זו נבחנו שלוש שיטות לשערוך פרמטר זה והן כוללות שיערוך בקריטריון הסבירות המירבית, קריטריון הסבירות הא-פוסטריורית המירבית (MAP) ובגישת "ההחלטה המכוונת". כפי שהסברנו בפרק 1 שני המשערכים האחרונים היו מוצלחים במיוחד בהקשר של הדגשת אותות דבור הטבולים ברעש, ונתנו תוצאות טובות יותר מאשר משערך הסבירות המירבית.

© Technion - Israel Institute of Technology, Elyachar Central Library

משערך "ההחלטה המכוונת" נתון ע"י (ראה נספח א'):

$$\hat{\xi}_{k,n} = \alpha \frac{\hat{A}_k^{\,2}(n-1)}{\hat{\lambda}_d(k,n-1)} + (1-\alpha)\ P\,[\gamma_{k,n} - 1] \tag{2.17}$$

כאשר $\xi_{k,n}$, $A_k(n)$, $\lambda_d(k,n)$ ו-$\gamma_{k,n}$ מסמנים את יחס האות לרעש הא-פריורי, האמפליטודה, ווארינס הרעש ויחס האות לרעש הא-פוסטריורי עבור רכיב התדר ה-k המתאים במסגרת האנליזה ה-n. α הוא פרמטר מיצוע המקבל ערכים בתחום $0 \le \alpha \le 1$.
$P[\cdot]$ הוא אופרטור המוגדר ע"י:

$$P[x] \triangleq \begin{cases} x & x \ge 0 \\ 0 & \text{אחרת} \end{cases} \tag{2.18}$$

ותפקידו למנוע מ-$\hat{\xi}_{k,n}$ להיות שלילי במידה ו- $(\gamma_{k,n} - 1)$ הוא שלילי. את משערך ווארינס הרעש $\hat{\lambda}_d(k,n)$ ניתן לקבל מקטעי השקט של אות הדבור הסמוכים ביותר לקטע בו מתבצעת ההדגשה. האינטרפרטציה של משערך זה כמשערך "ההחלטה-המכוונת" נובעת מכך שהשערוך ברגע ה-n מסתמך על שערוך האמפליטודה ברגע ה- n-1.

משערך ה-MAP של יחס האות לרעש של רכיב התדר ה-k מתואר בפרוט בנספח ג'. משערך זה נגזר בהנחה שהרעש הוא סטציונרי (כלומר $\lambda_d(k) = \lambda_d(k,n)$) ושערכו של $\lambda_d(k)$ ידוע. באופן מעשי נשתמש במשערך של $\lambda_d(k)$ ונציבו למשערך ה-MAP שיתקבל. משערך ה-MAP של יחס האות לרעש מתקבל מפתרון בעית המכסימיזציה הבאה:

$$\begin{aligned} \underline{\xi}_{k,N} &= \arg\max_{\underline{\xi}_{k,N}} p(\underline{\xi}_{k,N} | \underline{\gamma}_{k,N}) \\ &= \arg\max_{\underline{\xi}_{k,N}} p(\underline{\gamma}_{k,N} | \underline{\xi}_{k,N})\ p(\underline{\xi}_{k,N}) \end{aligned} \tag{2.19}$$

כאשר $\underline{\xi}_{k,N} \triangleq (\xi_{k,1}, \xi_{k,2}, \ldots, \xi_{k,N})$ ובאופן דומה $\underline{\gamma}_{k,N} \triangleq (\gamma_{k,1}, \gamma_{k,2}, \ldots, \gamma_{k,N})$. לחישוב $p(\underline{\gamma}_{k,N} | \underline{\xi}_{k,N})$ אנו מניחים שמרכיבי הוקטור $\underline{\gamma}_{k,N}$ אינם תלויים סטטיסטית בהינתן הוקטור $\underline{\xi}_{k,N}$. לחישוב $p(\underline{\xi}_{k,N})$ אנו מניחים מודל מרקובי מתאים ליצירת הווארינסים של רכיב התדר ה-k של אות הדבור. תחת הנחות אלו אנו מסגננים את בעית השערוך כבעיה רקורסיבית דיסקרטית ופותרים אותה באמצעות האלגוריתם של ויטרבי. כיוון שאנו משתמשים בהחלטה מושהית לגבי השערוך של $\xi_{k,n}$, המשערך המתקבל קרוב למשערך ה-MAP האמיתי.

© Technion - Israel Institute of Technology, Elyachar Central Library

בישום מעשי של משערך ה-MAP ושל משערך "ההחלטה-המכוונת", הפועלים ביחד עם משערך האמפליטודה הספקטרלית (2.1) לצורך הדגשת אותות דבור, התקבלו תוצאות דומות מאוד עם עדיפות מסוימת למשערך ה-MAP. בשני המקרים התקבל שיפור ניכר באיכות אות הדבור (ראה להלן). לתוצאה זו חשיבות מעשית רבה כיוון שהיא מורה שניתן להשתמש במשערך "ההחלטה-המכוונת" המבוסס פחות (מבחינה אנליטית) מאשר משערך ה-MAP, אך הנותן תוצאות מאוד דומות.

משערכי האמפליטודה הספקטרלית שפותחו בעבודה זו נבחנו במערכת להדגשת דבור שדומתה במחשב. אות הכניסה למערכת זו הוא אות דבור דגום ב-8kHz ובעל רוחב סרט של 0.2-3.2kHz, אשר התווסף לו רעש ברמה מתאימה. כל קטע מאות הכניסה הנ"ל שאורכו 32msec והחופף לקטע הקודם ב-24msec עובר אנליזה ספקטרלית. האמפליטודה של כל רכיב ספקטרלי משוערכת ומשמשת יחד עם האקספוננט הקומפלקסי של הפאזה הרועשת לשם שערוך הרכיבים הספקטרליים של אות הדבור. דגמי אות הדבור המודגש מתקבלים מדגמי התמרת פוריה ההפוכה של הרכיבים הספקטרליים המשוערכים בשיטת ה- overlap and add. בישום משערכי האמפליטודה הספקטרלית בחנו הן את הישום המדויק והן את השימוש בטבלאות המכילות מספיק דגמים של פונקציות ההגבר המתאימות. מסתבר שניתן להקטין באופן ניכר את סיבוכיות המערכת ע"י שימוש בטבלאות ההגבר, עם פגיעה מינימלית באיכות האות המודגש.

במסגרת עבודה זו בחנו אותות דבור בעלי יחס אות לרעש של 5dB, 0dB ו-5dB-. את התוצאות הטובות ביותר קבלנו כאשר שערוך האמפליטודה הספקטרלית בוצע ע"י המשערך (2.12) עם $q_k = 0.2$, הפועל ביחד עם משערך "ההחלטה-המכוונת" (עם $\alpha = 0.99$). איכות האות המודגש שהתקבל בדרך זו הינה טובה באופן ניכר מזו של האות הרועש. שיפור איכות זה מתבטא בהנחתה רצינית של רעש הרקע. כמו כן רעש הרקע הנותר נשמע אחיד. לאות המודגש נלווה עוות מסוים שהולך וגדל ככל שיחס האות לרעש בכניסה קטן. אולם האות נשמע די ברור גם ביחס אות לרעש נמוך של 0dB. תוצאות מאוד דומות התקבלו כאשר השתמשנו במשערך (2.16) שפעל ביחד עם משערך ההחלטה-המכוונת (עם $\alpha = 0.98$).

© Technion - Israel Institute of Technology, Elyachar Central Library

References

[1] D. Middelton, Introduction to Statistical Communication Theory, McGraw-Hill, N.Y. 1960. Chap.7, Appendix 1.

[2] S. Bernstein, Sur l'extension du theoreme limite du calcul des probabilites aux sommes des quantites dependantes, Math. Ann., 97:1, 1926.

[3] W.B. Davenport and W.L. Root, "An Introduction to the Theory of Random Signals and Noise, McGraw-Hill, New York., Chap.6.

[4] T.T. Kadota, "Optimal Reception of Binary Gaussian Signals", Bell Sys. Tech. J., Vol. 43, pp. 2767-2810, Nov. 1964.

[5] I.B.M. Application Program, System/360 Scientific Subroutine Package (360A-CM-03X) Version III, pp. 368-369, 1968.

© Technion - Israel Institute of Technology, Elyachar Central Library

פרק 3

שלוב הדגשה וקדוד מסתגל במשור התדר של אותות דבור רועשים

בפרק זה נדון בישום הטכניקות שפותחו בפרק 2 לשם שיפור האיכות של אות דבור משוחזר המתקבל ממקדד מסתגל הפועל במישור התדר בתנאי כניסה רועשים. מקדד זה קרוי בספרות adaptive transform coder (ATC). ההדגשה תעשה כאן במשולב עם הקדוד בתחום התדר, אך בטרם מתבצע הקדוד עצמו. נושא זה נדון בהרחבה בנספח ד' של עבודה זו ולכן יובאו כאן עיקר הדברים בלבד.

מקדד ה-ATC הינו יעיל ביותר לקדוד אות דבור בקצבים של 7.2-16 kb/s. בקצב של 16 kb/s או יותר הוא נותן אות דבור באיכות גבוהה הקרויה toll quality. בקצב הנמוך של 7.2 kb/s הוא נותן אות דבור באיכות המתאימה לתקשורת והקרויה communication quality. מקדד זה נחקר רבות בהקשר של קדוד אותות דבור הנקיים מרעש [1,2]. עקרון פעולתו מבוסס על התמרת כל מסגרת של אות הדבור (בד"כ בעלת משך של 32 msec) למישור התדר וכימוי (קוונטיזציה) כל רכיב תדר בנפרד, בהתאם להקצאת סיביות דינמית ולצעד כימוי מסתגל. ההתמרה בה משתמשים במקדד זה הינה התמרת ה- discrete cosine transform (DCT). הקצאת הסיביות לרכיבי התדר של אות הדבור במסגרת אנליזה נתונה וכן קביעת צעד הכימוי לכל רכיב תדר, נעשים על סמך ידיעת הווארינס של כל אחד מרכיבי התדר. ווארינסים אלו משוערכים בצורה פרמטרית בכל מסגרת ומשודרים למקלט כאינפורמצית צד. חוק חלוקת הסיביות והצעד המנורמל של הכימוי נקבעים על סמך ההנחה שרכיבי התדר של אות הדבור מפולגים גאוסית. הנחה זו אומתה גם בתצפיות מעשיות על אות הדבור [1,2]. לפיכך לשם קבלת חוק החלוקה האופטימלי מנצלים את פונקצית קצב-עוות של מקור גאוסי. כמו כן לקבלת צעד הכימוי המנורמל האופטימלי מנצלים את המכמה (quantizer) של Max [3].

מקור ה-ATC נמצא רגיש לנוכחות רעש באות הכניסה אליו וביצועיו ירודים במצב כזה. שלא כמו במקדדי צורות גל אחרים (לדוגמא PCM, DPCM וכו') בהם רעש הכניסה משוקף ליציאה, כאן לרעש אפקט מזיק נוסף המתבטא בקלקול השערוך הפרמטרי של הווארינסים של רכיבי התדר. כתוצאה מכך חלוקת הסיביות בין רכיבי התדר וכן צעד הכימוי האופטימלי לכל רכיב תדר משתבשים. שיבוש זה גורם לאי רציפויות באות המקודד אשר פוגמים במידה רבה באיכותו ובמובנותו.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

בעבודה זו בחנו את האפשרות לשיפור פעולתו של מקדד ה-ATC הפועל על אות דבור רועש, ע"י שערוך אופטימלי של האמפליטודה הספקטרלית לזמן קצר של אות הדבור ונצול הפאזה הרועשת בטרם מתבצע הקדוד. מכיוון שלרכיבי ה-DCT ולרכיבי ה-DFT אותה עוטפת ספקטרלית [2], שערוך האמפליטודה הספקטרלית במקרה הנוכחי מזדהה עם שערוך האמפליטודה של רכיבי ה-DCT. לכן גישה זו היא נוחה במיוחד בהקשר הנוכחי.

גזירת משערך האמפליטודה הספקטרלית במקרה הנדון מבוססת על מודל סטטיסטי דומה לזה שהובח בפרק 2, אשר מנצל תכונות אסימפטוטיות של רכיבי התדר. אנו מניחים שרכיבי ה-DCT של אות הדבור וכן של הרעש ניתנים ליצוג כמשתנים אקראיים גאוסיים בלתי תלויים סטטיסטית. כמו כן נניח שהרעש הוא סטציונרי, אדיטיבי לאות הדבור וחסר קורלציה אתו. מעניין לציין שהמודל הסטטיסטי הנ"ל עבור רכיבי התדר של אות הדבור נצפה גם באופן מעשי ע"י Zelinski and Noll [1,2] כפי שכבר ציינו קודם. בהנחות אלו נקבל את המשערך האופטימלי במובן השגיאה הריבועית הממוצעת המינימלית של האמפליטודה הספקטרלית באופן הבא: נסמן ב- X_k, D_k וב-Y_k את רכיב ה-DCT ה-k של אות הדבור, הרעש והאות הרועש בהתאמה. המשערך האופטימלי של $|X_k|$ בהנתן רכיבי התדר הרועשים $\{Y_o,Y_1,\dots, Y_{M-1}\}$ נתון ע"י:

$$
\begin{aligned}
|\hat{X}_k| &= E\{\, |X_k| \mid Y_o,Y_1,\dots,Y_{M-1}\} \\
&= E\{\, |X_k| \mid Y_k \,\} \\
&= \int_{-\infty}^{\infty} |x_k|\, p(x_k|Y_k)\, dx_k
\end{aligned}
\tag{3.1}
$$

כאשר במעבר לשורה השניה של (3.1) נצלנו את האי-תלות הסטטיסטית של רכיבי התדר. M הוא מספר רכיבי התדר בהתמרת ה-DCT של מסגרת אנליזה נתונה. בהסתמך על המודל הגאוסי, קל להראות תוך ניצול הנוסחאות [4:3.546.2,3.562.4] ש-:

$$
|\hat{X}_k| = \frac{\xi_k}{1+\xi_k}\left[\, \Phi\left(\sqrt{\frac{v_k}{2}}\right) + \sqrt{\frac{2}{\pi}\,\frac{1}{v_k}}\, \exp\left(-\frac{v_k}{2}\right)\right]|Y_k| \tag{3.2}
$$

כאשר:

$$
\Phi(x) \triangleq \frac{2}{\sqrt{\pi}} \int_{o}^{x} e^{-t^2}\, dt \tag{3.3}
$$

© Technion - Israel Institute of Technology, Elyachar Central Library

v_k ו-ξ_k מוגדרים ע"י (2.2) ו-(2.3) בהתאמה. פונקצית ההגבר המתקבלת מ-(3.2) מתוארת בציור 2 של נספח ד'. בישום המשערך הנ"ל במערכת ה-ATC השתמשנו במשערך יחס האות לרעש ξ_k הנתון ב-(2.17), כאשר α נקבע על סמך ניסויי שמיעה לא פורמליים. ווארינס רכיב התדר של הרעש שוערך כרגיל מקטע התחלתי שהכיל רעש בלבד ושאורכו במקרה הנוכחי הוא 640 msec.

במסגרת עבודה זו בחנו את הגירסה של ה-ACT המכונה Speech-specific ACT והמתוארת בפרוט ב-[2]. הפעלנו מקדד זה בקצבים של 12 kb/s ו-16 kb/s עבור יחסי אות לרעש בכניסה של 5dB ו-10dB, כאשר הרעש הוא רחב סרט. הערך של α בו השתמשנו בישום משערך יחס האות לרעש של רכיב התדר (2.17) הוא: $\alpha = 0.94$ עבור 16 kb/s ו- $\alpha = 0.85$ עבור 12 kb/s. שאר פרמטרי המערכת נקבעו כמומלץ ב-[2]. הגישה המוצעת בעבודה זו שפרה באופן ניכר את איכות האות המשוחזר על אף שהוא איבד במקצת מחדותו. רמת הרעש באות המשוחזר היתה נמוכה בהרבה מזו שבאות הכניסה וכן אי הרציפויות שאפיינו את האות המשוחזר הרועש נעלמו כמעט.

נציין שבמסגרת עבודה זו בחנו גם את האפשרות של שיפור באיכות האות המשוחזר כאשר אות הכניסה למקדד הוא נקי מרעש. שלוש גישות נבחנו כאן. בראשונה בצענו פעולות הדגשה על יציאת המכמה של כ"א מרכיבי התדר, במטרה להקטין את רמת רעש הכימוי. כאן בצענו שערוך אופטימלי במובן השגיאה הריבועית הממוצעת המינימלית של רכיב ה-DCT של אות הדבור מתוך הרכיב המכיל רעש כימוי. בגישה השניה השתמשנו בכימוי עם dither [5] במטרה לבצע דה-קורלציה של האות ורעש הכימוי. הגישה השלישית משלבת את שתי הגישות הנ"ל ובה מבצעים פעולות הדגשה על יציאת המכמה המנצל dither. ברור שהגישה הראשונה והשלישית הגיונית במידה והמכמה איננו אופטימלי במובן של Max [3]. אולם זהו המקרה המעשי היות ובד"כ משתמשים במכמה אחיד. מענין לציין שבשתי גישות אלו משערך רכיב ה-DCT של אות הדבור הוא ה-centroid של השטח תחת הפילוג הסגולי של הרכיב הנ"ל באינטרוול הכימוי. זוהי תוצאה מעניינת המתלכדת עם זו של Max המתקבלת בתכנון המכמה הלא אחיד האופטימלי. אולם אצל Max האופטימיזציה מתבצעת גם על צעד הכימוי. נציין לבסוף שביישום הנוכחי שערכנו את רכיב התדר של אות הדבור ולאו דוקא את האמפליטודה הספקטרלית שלו, כיוון שכאן סימן רכיב ה-DCT ידוע במדויק ולכן שני השערוכים הנ"ל נותנים אותה תוצאה.

למרבה הצער אף גישה מהשלוש הנ"ל לא הובילה לשיפור באיכות האות המקודד. אחד ההסברים לכך הוא שההדגשה הנ"ל מקרבת אותנו למעשה למכמה האופטימלי הלא אחיד של Max. אולם כפי שניתן לראות מציור 5 שב-[3], השיפור (במונחים של שגיאה ריבועית ממוצעת) המתקבל משימוש במכמה אופטימלי לא אחיד בהשוואה לשימוש במכמה אופטימלי אחיד איננו גדול. במקרה שלנו בו מספר הסיביות המוקצות לכל רכיב תדר איננו עולה על 4 [2], השיפור המכסימלי הצפוי הוא בירידה של כ-20% בשגיאה הריבועית הממוצעת.

References

(1) R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, No. 4, pp. 299-309, Aug. 1977.

(2) J.M. Tribolet and R.E. Crochiere, "Frequency Domain Coding of Speech", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, No. 5, pp. 512-530, Oct. 1979.

(3) J. Max, "Quantizing for Minimal Distortion", IRE Trans. Inform. Theory, Vol. IT-6, pp. 7-12, March 1960.

(4) I.S. Gradshteyn and I.M. Ryzhik, Table of Integrals, Series, and Products, Academic Press Inc., 1980.

(5) L. Schuchman, "Dither Signals and their Effect on Quantization Noise", IEEE Trans. Commun. Tech., Vol. COM-12, pp. 162-165, Dec. 1964.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

פרק 4 - דיון ומסקנות

בעבודה זו טפלנו בבעית ההדגשה של אותות דבור הטבולים ברעש, ע"י שערוך אופטימלי של האמפליטודה הספקטרלית לזמן קצר של אות הדבור שהיא בעלת חשיבות מרכזית בתהליך התפישה שלו ע"י מערכת השמע. תוך ניצול תכונות סטטיסטיות אסימפטוטיות של רכיבי התדר, גזרנו משערך שגיאה ריבועית ממוצעת מינימלית של האמפליטודה הספקטרלית לזמן קצר של אות הדבור. כמו כן גזרנו תחת אותו קריטריון ומודל סטטיסטי את משערך לוגריתם האמפליטודה הספקטרלית והרחבנו את המשערכים הנ"ל כך שיתאימו למצב הריאלי בו האות איננו נמצא כל העת במדידות הרועשות.

לשם שחזור אות הדבור תוך ניצול המשערך האופטימלי של האמפליטודה הספקטרלית, בחנו כאן את השערוך האופטימלי במובן השגיאה הריבועית הממוצעת המינימלית של האקספוננט הקומפלקסי של רכיב תדר של אות הדבור. הראינו שעל מנת לא לפגוע בשערוך האופטימלי של האמפליטודה הספקטרלית, יש להשתמש באקספוננט הקומפלקסי של הפאזה הרועשת.

הגישה שננקטה בעבודה זו הובילה לשיפור ניכר באיכות האות המודגש. בפרט, שיפור זה התקבל כאשר השתמשנו במשערך האמפליטודה הלוקח בחשבון את אי-הוודאות בקיום האות ברכיבי התדר הרועשים והפועל ביחד עם משערך ההחלטה-המכוונת של יחס האות לרעש של רכיב התדר. שיפור האיכות הנ"ל התבטא בירידה משמעותית של רעש הרקע ובכך שהרעש הנותר נשמע כרעש רחב סרט בעל גוון אחיד. העוות הנלווה לאות הדבור המודגש איננו גדול והאות נשמע ברור למדי גם במקרה הקשה בו יחס האות לרעש בכניסה הוא 0dB. אין בידנו תשובה לשאלה האם השיטה הנ"ל משפרת גם את מובנות האות הרועש, כיוון שבמסגרת עבודה זו לא בוצעו מבחני מובנות. הסיבה לכך נעוצה בקשיים אוביקטיביים של ביצוע מבחני מובנות מהימנים. בעיה זו חריפה אף יותר במקום בו שפת האם איננה אנגלית. יתכן ומבחנים כאלו יבוצעו בעתיד הקרוב בארה"ב ע"י חברה המתמחה בנושא. לפי שעה נוכל לקבוע בוודאות שתרומת העבודה היא בשיפור האיכות של האות הרועש ולעובדה זו חשיבות בזכות עצמה כפי שהובהר בפרק המבוא.

לדעתנו קיימות מספר דרכים להמשך העבודה. ראשית ניתן לפתח את הרעיון של שערוך בתנאי אי-וודאות של קיום האות במדידות הרועשות ולהתאימו יותר לבעיה הנוכחית. כך למשל ניתן להגדיר שלוש השערות המאפיינות את אות הדבור: האחת מציינת אות דבור קולי, השניה אות דבור אל-קולי והשלישית מציינת אי-קיום אות או במלים אחרות את מצב השקט. במקרה זה קל להראות שהמשערך האופטימלי במובן השגיאה הריבועית הממוצעת המינימלית של האמפליטודה A_k, בהנתן הרכיב הספקטרלי הרועש Y_k, הוא:

$$\hat{A}_k = \frac{\Lambda_{vs}(Y_k,q_{vs})}{1+\Lambda_{vs}(Y_k,q_{vs})+\Lambda_{us}(Y_k,q_{us})} E\{ A_k \mid Y_k, H_v \}$$

$$+ \frac{\Lambda_{us}(Y_k,q_{us})}{1+\Lambda_{vs}(Y_k,q_{vs})+\Lambda_{us}(Y_k,q_{us})} E\{ A_k \mid Y_k, H_u \} \qquad (4.1)$$

כאשר $\Lambda(\cdot)$ הוא מעין יחס סבירות מוכלל המוגדר ע"י:

$$\Lambda_{ij}(Y_k,q_{ij}) \triangleq \frac{q_i \; p(Y_k|H_i)}{q_j \; p(Y_k|H_j)} \qquad i,j\epsilon\{ v,u,s\} \qquad (4.2)$$

H_v, H_u ו-H_s מסמנים את ההשערות המתאימות לאות דבור קולי, אל-קולי ושקט בהתאמה. q_v, q_u ו-q_s הן ההסתברויות של שלוש ההשערות H_v, H_u ו-H_s בהתאמה. $E\{A_k|Y_k,H_v\}$ ו-$E\{A_k|Y_k,H_u\}$ הן המשערכים האופטימליים של A_k בהנתן שאות הדבור הוא קולי או אל-קולי בהתאמה. לשערוך המתבצע בדרך הנ"ל עשוי להיות יתרון ביחס לשערוך שבצענו אנו, כיוון שהוא מאפשר לנצל משערכים המתאימים במיוחד לאותות דבור קוליים או אל-קוליים ולשלבם במסגרת אחת. כך למשל ניתן יהיה במסגרת המשערך $E\{A_k|Y_k,H_v\}$ להוסיף אינפורמציה אודות מחזור ה-pitch של אות הדבור.

אפשרות אחרת להרחבת העבודה הנוכחית קשורה בשערוך ריבוע האמפליטודה הספקטרלית שהוא פשוט הרבה יותר מזה של שערוך האמפליטודה הספקטרלית עצמה. למשל בהנחות שפורטו בפרק 2 ניתן להראות בקלות שהמשערך האופטימלי של ריבוע האמפליטודה הספקטרלית נתון ע"פ:

$$\hat{A}_k^2 = \frac{v_k}{\gamma_k^2}(1+v_k)\, R_k^2 \qquad (4.3)$$

מ-(4.3) ניתן לקבל משערך תת אופטימלי לאמפליטודה הספקטרלית ע"י הוצאת שורש ריבועי.

הרחבה אחרת הקשורה לנושא הנ"ל אך העשויה להקל בהנחות לפיהם בוצע השערוך (4.3), היא הבאה [1]: נרשום את הרכיב הספקטרלי ה-k של אות הדבור:

$$X_k = \frac{1}{T} \int_0^T x(\tau) \exp(-j\frac{2\pi}{T} k\tau)\ d\tau \qquad (4.4)$$

משערך השגיאה הריבועית הממוצעת המינימלית של $A_k^2 = |X_k|^2$ בהנתן $\{ y(t) , 0 \leq t \leq T\}$, נתון ע"פ:

$$\hat{A}_k^2 = E\{ A_k^2 \mid y(t) , 0 \leq t \leq T \}$$

$$= \frac{1}{T^2} \iint_0^T E\{x(\tau)x(s) | y(t), 0 \leq t \leq T\} \exp[-j \frac{2\pi}{T} k(\tau-s)]\ d\tau ds \qquad (4.5)$$

מ-(4.5) נובע שבעית השערוך של A_k^2 הופכת להיות בעית השערוך של $E\{x(\tau)x(s) | y(t) , 0 \leq t \leq T\}$ עבור $0 \leq \tau , s \leq T$. אין זו בעיה קלה לפתרון ואולם יתכן שצורת הצגה כזו עשויה להוביל לקרובים נאותים.

אפשרות שלישית להרחבת העבודה הנוכחית, היא שערוך בקריטריון הסבירות הא-פוסטריורית המירבית של האמפליטודה הספקטרלית. כאן ניתן לנצל רעיונות דומים לאלו שפותחו בשערוך ווארינס רכיב התדר של אות הדבור וכן לנצל את האלגוריתם של ויטרבי.

מסתבר שרעיונות להדגשת אותות דבור אינם חסרים ואולם עקב אכילס של הבעיה טמון בהעדר קריטריונים אוביקטיביים הנמצאים בקורלציה עם התפישה של אות הדבור ע"י מערכת השמע, לפיהם ניתן להעריך את המערכות השונות. אומנם קיימים היום מספר קריטריונים שפותחו בעיקר למטרות קידוד של אות הדבור [2], אולם קריטריונים אלו אינם ישימים לבעיות ההדגשה של אות דבור רועש. הסיבה לכך נעוצה בעובדה שקריטריונים אלו מנצלים למשל את המודל האוטו-רגרסיבי של אות הדבור שכמובן איננו תקף עבור אות רועש וקרוב לוודאי גם לא עבור האות המודגש. נראה לנו שפיתוח קריטריונים מעין אלו עשוי לתרום באופן משמעותי ביותר לפתרון בעית ההדגשה של אותות דבור. נושא זה מהווה בעיה פתוחה ובוודאי ימשיך להעסיק חוקרים רבים בעתיד.

© Technion - Israel Institute of Technology, Elyachar Central Library

References

(1) S. Shitz, Private communication.

(2) R.M. Gray, A. Buzo, A.H. Gray, Y. Matsuyama, "Distortion Measures for Speech Processing", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, pp. 367-376, Aug. 1980.

© Technion - Israel Institute of Technology, Elyachar Central Library

<u>נספח א'</u> - הדגשת אותות דבור תוך שימוש במשערך השגיאה הריבועית הממוצעת המינימלית של האמפליטודה הספקטרלית לזמן קצר

© Technion - Israel Institute of Technology, Elyachar Central Library

SPEECH ENHANCEMENT USING A MINIMUM MEAN SQUARE ERROR SPECTRAL AMPLITUDE ESTIMATOR[1]

ABSTRACT

This paper focuses on the class of speech enhancement systems which capitalize on the major importance of the *short-time spectral amplitude (STSA)* in speech perception. A system which utilizes a minimum mean square error (MMSE) STSA estimator is proposed, and compared with other widely used systems, which are based on Wiener filtering and the 'spectral subtraction' algorithm. The derivation of the MMSE estimator is based on modeling speech and noise *spectral components* as statistically independent Gaussian random variables.

In this paper we derive the MMSE STSA estimator, analyze its performance, and compare it with the Wiener STSA estimator. We also examine the MMSE STSA estimator under uncertainty of signal presence in the noisy observations.

For constructing the enhanced signals, the MMSE STSA estimator is combined with the complex exponential of the noisy phase. It is shown here that the latter is the optimal MMSE complex exponential estimator, which does not affect the STSA estimation.

The proposed approach results in a significant reduction of the noise, and provides enhanced speech with *colorless* residual noise. The complexity

---

[1]The research was supported by Technion V.P.R. Fund - Natkin Fund for Electrical Engineering Research.

© Technion - Israel Institute of Technology, Elyachar Central Library

of the proposed algorithm is approximately as that of other systems in the discussed class.

I. INTRODUCTION

The problem of enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available, has recently received much attention. This is due to the many potential applications a successful speech enhancement system can have, and because of the available technology which enables the implementation of such intricate algorithms. A comprehensive review of the various speech enhancement systems which emerged in recent years, and their classification according to the aspects of speech production and perception they capitalize on, can be found in [1].

We focus here on the class of speech enhancement systems, which capitalize on the major importance of the short-time spectral amplitude (STSA) in speech perception [1,2]. In these systems the STSA of the speech signal is estimated, and combined with the short-time phase of the degraded speech, for constructing the enhanced signal. The 'spectral subtraction' algorithm and Wiener filtering are well known examples [1,3]. In the 'spectral subtraction' algorithm, the STSA is estimated as the square root of the maximum likelihood (ML) estimator, of each signal spectral component variance [3]. In systems which are based on Wiener filtering, the STSA estimator is obtained as the modulus of the optimal minimum mean square error (MMSE) estimator, of each signal spectral component [1,3]. These two STSA estimators were derived under Gaussian assumption.

Since the 'spectral subtraction' STSA estimator is derived from an optimal (in the ML sense) variance estimator, and the Wiener STSA estimator is derived from the optimal MMSE signal spectral estimator, both are

© Technion - Israel Institute of Technology, Elyachar Central Library

not optimal *spectral amplitude* estimators, under the assumed statistical model and criterion. This observation led us to look for an optimal STSA estimator, which is derived directly from the noisy observations. We concentrate here on the derivation of an optimal MMSE STSA estimator, and on its application in a speech enhancement system.

The STSA estimation problem is formulated here, as that of estimating the modulus of each complex Fourier expansion coefficient[2] of a given speech segment. This formulation is motivated by the fact that the Fourier expansion coefficients of a given signal segment are samples of its Fourier transform, and by the close relation between the Fourier series expansion and the discrete Fourier transform. The latter relation enables an efficient implementation of the resulting algorithm, by utilizing the FFT algorithm.

To derive the optimal STSA estimator, the a-priori probability distributions of the speech and noise Fourier expansion coefficients should be known. Since in practice they are unknown, one can measure each probability distribution, or alternatively, assume a reasonable statistical model.

In the discussed problem, the speech and possibly also the noise are non-stationary processes. Therefore, they are non-ergodic processes as well. This fact excludes the convenient possibility of obtaining the above probability distributions by examining the long-time behavior of each process. Hence, the only way which can be used, is to examine independent sample functions belonging to the ensemble of each process. e.g., for the speech process these sample functions can be obtained from different speakers. However, since the probability distributions we are dealing with are time-varying (due to the non-stationarity of the processes), their measurement and characterization by the above way is complicated, and

[2]The complex Fourier expansion coefficients are also referred here as spectral components.

© Technion - Israel Institute of Technology, Elyachar Central Library

the entire procedure seems to be impracticable.

For the above reasons, a statistical model is used here. This model is based on asymptotic ($T \to \infty$) statistical properties of the Fourier expansion coefficients. Specifically, we assume that the Fourier expansion coefficients of each process can be modeled as statistically independent Gaussian random variables. The mean of each coefficient is assumed to be zero, since the processes involved here are assumed to have zero mean. The variance of each speech Fourier expansion coefficient is time-varying, due to speech non-stationarity. This Gaussian statistical model is motivated by the central limit theorem, as each Fourier expansion coefficient is after all a weighted sum (or integral) of random variables resulting from the process samples. The fact that the central limit theorem is valid (under general conditions) also for dependent random variables, which may have different distributions [4,5], encourages the use of the Gaussian model in the discussed problem.

The statistical independence assumption in the Gaussian model, is actually equivalent to the assumption that the Fourier expansion coefficients are uncorrelated. This latter assumption is commonly justified by the fact that the normalized correlation between different Fourier expansion coefficients approaches zero, as the analysis frame length approaches infinity [6].

In our problem, the analysis frame length T cannot be too large, due to the quasi-stationarity of the speech signal. Its typical value is 20-40msec. This may cause the Fourier expansion coefficients to be correlated to a certain degree. Nevertheless, we continue with this statistical independence assumption, in order to simplify the resulting algorithm. The case of statistically dependent expansion coefficients is now under investigation.

© Technion - Israel Institute of Technology, Elyachar Central Library

In practice, an appropriate window (e.g. Hanning) is applied to the noisy process, which reduces the correlation between widely separated spectral components, at the expense of increasing the correlation between adjacent spectral components. This is a consequence of the wider main lobe, but the lower side lobes of a window function, in comparison to the rectangular window.

In conclusion of the above discussion concerning the statistical model used here, we note that since the true statistical model is inaccessible, the validity of the proposed one can be judged a-posteriori on the basis of the results obtained here. In addition, the term "optimal" attributed to the estimator derived here, should be understood in connection with the assumed statistical model.

In this paper we derive the optimal MMSE STSA estimator based on the above statistical model, and compare its performance with that of the Wiener STSA estimator. This comparison is of interest, since the Wiener estimator is a widely used STSA estimator, which is also derived under the same statistical model.

The Gaussian statistical model assumed above, does not take into account the fact that the speech signal is not surely present in the noisy observation. This model results in a Rayleigh distribution for the amplitude of each signal spectral component, which assumes insignificant probability for low amplitude realizations. Therefore, this model can lead to less suppression of the noise, than other amplitude distribution models (e.g., Gamma) which assume high probability for low amplitude realizations. However, using a statistical model of the latter type, can lead to a worse amplitude estimation when the signal is present in the noisy observations.

© Technion - Israel Institute of Technology, Elyachar Central Library

One useful approach to the solution of this problem, is to derive an optimal MMSE estimator which takes into account the uncertainty of speech presence in the noisy observations [3,7,8]. Such an estimator can be derived on the basis of the above Gaussian statistical model, and by assuming that the signal is present in the noisy observations with probability $p<1$ only. The parameter p supplies a useful degree of freedom, which enables to compromise between noise suppression and signal distortion. This is of course an advantage in comparison to the use of a Gamma type statistical model.

The above approach is applied in this paper, and the resulting STSA estimator is compared with the McAulay and Malpass [3] estimator, in enhancing speech. The latter estimator is an appropriately modified ML STSA estimator, which assumes that the signal is present in the noisy spectral components with a probability of $p=0.5$.

In this paper we also examine the estimation of the complex exponential of the phase, of a given signal spectral component. The complex exponential estimator is used in conjunction with the optimal STSA estimator, for constructing the enhanced signal. We derive here the optimal MMSE complex exponential estimator, and discuss its effect on the STSA estimation. We show that the complex exponential of the noisy phase, is the optimal complex exponential estimator which does not affect the STSA estimation.

The paper is organized as follows: In Section II and Appendix A we derive the optimal STSA estimator, and compare its performance with that of the Wiener STSA estimator. In Section III and Appendix C we extend the optimal STSA estimator, and derive it under uncertainty of signal presence in the noisy spectral components. In Section IV and Appendix D we discuss

© Technion - Israel Institute of Technology, Elyachar Central Library

the optimal estimation of the complex exponential of the phase. In Section V we discuss the problem of estimating the variances of the speech and noise spectral components, which are the parameters of the statistical model. In Section VI we describe the proposed speech enhancement system, and compare it with the other widely used systems mentioned above. In Section VII we summarize the paper and draw conclusions.

II. OPTIMAL SHORT-TIME SPECTRAL AMPLITUDE ESTIMATOR

In this section we derive the optimal STSA estimator, under the statistical model assumed in Section I. We also analyze its performance, and examine its sensitivity to a key parameter of the statistical model. This performance and sensitivity analysis is also done for the Wiener STSA estimator, and the two estimators are compared on this basis.

Derivation of Amplitude Estimator

Let $x(t)$ and $d(t)$ denote the speech and the noise processes, respectively. The observed signal $y(t)$ is given by:

$$y(t) = x(t) + d(t), \qquad 0 \le t \le T \tag{1}$$

where without loss of generality we let the observation interval to be $[0,T]$. Let $X_k \triangleq A_k \exp(j\alpha_k)$, D_k, and $Y_k \triangleq R_k \exp(j\vartheta_k)$ denote the k-th spectral component of the signal $x(t)$, the noise $d(t)$, and the noisy observations $y(t)$, respectively, in the analysis interval $[0,T]$. Y_k (and similarly X_k and D_k) is given by:

$$Y_k = \frac{1}{T}\int_0^T y(t) exp(-j\frac{2\pi}{T}kt)dt \qquad k=0,\pm1,\pm2,\ldots \tag{2}$$

Based on the formulation of the estimation problem given in the previous section, our task is to optimally estimate the modulus A_k, from the degraded signal $\{y(t),\ 0 \le t \le T\}$.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

Toward this end, we note that the signal $\{y(t),\ 0 \le t \le T\}$ can be written in terms of its spectral components Y_k by [6]:

$$y(t) = \underset{K \to \infty}{l.i.m} \sum_{k=-K}^{K} Y_k exp(j \frac{2\pi}{T} kt) \qquad 0 \le t \le T \tag{3}$$

where $l.i.m$ means limit in the mean. Moreover, on the basis of the Gaussian statistical model for the spectral components assumed here, the series (3) converges almost surely to $y(t)$, for every $t \in [0,T]$. Therefore, it can be shown that $\{y(t),\ 0 \le t \le T\}$ and $\{Y_0, Y_1, \ldots\}$ bear the same information (up to events whose probability is zero) [9, Appendix D]. This means that the estimation problem can be reduced to be that of optimally estimating A_k from the infinite countable set of observations $\{Y_0, Y_1, \ldots\}$. In addition, since the spectral components are assumed to be statistically independent, the optimal MMSE amplitude estimator can be derived from Y_k only. In conclusion, the optimal MMSE estimator $\hat{A}_k$ of A_k is obtained as follows:

$$\begin{aligned} \hat{A}_k &= E\{A_k \mid y(t),\ 0 \le t \le T\} \\ &= E\{A_k \mid Y_0, Y_1, \cdots\} \\ &= E\{A_k \mid Y_k\} \\ &= \frac{\int_0^\infty \int_0^{2\pi} a_k p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k} \end{aligned} \tag{4}$$

where $E\{\cdot\}$ denotes the expectation operator, and $p(\cdot)$ denotes a probability density function (PDF).

Under the assumed statistical model, $p(Y_k \mid a_k, \alpha_k)$ and $p(a_k, \alpha_k)$ are given by

$$p(Y_k \mid a_k, \alpha_k) = \frac{1}{\pi \lambda_d(k)} \exp\{-\frac{1}{\lambda_d(k)} \mid Y_k - a_k e^{j\alpha_k} \mid^2\} \tag{5}$$

© Technion - Israel Institute of Technology, Elyachar Central Library

$$p(a_k, \alpha_k) = \frac{a_k}{\pi \lambda_x(k)} \exp\left\{ - \frac{a_k^2}{\lambda_x(k)} \right\} \tag{6}$$

where $\lambda_x(k) \triangleq E\{|X_k|^2\}$, and $\lambda_d(k) \triangleq E\{|D_k|^2\}$, are the variances of the k-th spectral component of the speech and the noise respectively. Substituting (5) and (6) into (4) gives (see Appendix A):

$$\hat{A}_k = \Gamma(1.5) \frac{\sqrt{v_k}}{\gamma_k} M(-0.5;1;-v_k) R_k \tag{7}$$

$$= \Gamma(1.5) \frac{\sqrt{v_k}}{\gamma_k} exp\left(-\frac{v_k}{2}\right)\left[(1+v_k) I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] R_k$$

$\Gamma(\cdot)$ denotes the gamma function, with $\Gamma(1.5)=\sqrt{\pi}/2$; $M(a;c;x)$ is the confluent hypergeometric function [4: A.1.14]; $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order respectively. v_k is defined by:

$$v_k \triangleq \frac{\xi_k}{1+\xi_k} \gamma_k \tag{8}$$

where ξ_k and γ_k are defined by:

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \tag{9}$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)} \tag{10}$$

ξ_k and γ_k are interpreted (after McAulay and Malpass [3]) as the a-priori and a-posteriori signal to noise ratio (SNR) respectively.

A similar expression to (7) was obtained in [11,12], when the amplitude of a Gaussian sinusoidal random process buried in Gaussian noise is optimally estimated.

It is of interest to examine the asymptotic behavior of $\hat{A}_k$ at high SNR, i.e., at $\xi_k >> 1$. By considering the exponential distribution of v_k (i.e., $p(v_k)=1/\xi_k \exp(-v_k/\xi_k)$), it is easy to see that $\xi_k \gg 1$ implies $v_k \gg 1$ with high probability. Therefore, to examine $\hat{A}_k$ at $\xi_k \gg 1$, we substitute the following approximation of the confluent hypergeometric function [4: A.1.16b] in (7).

$$M(-0.5;1;-v_k) \cong \frac{\sqrt{v_k}}{\Gamma(1.5)} \qquad v_k \gg 1 \tag{11}$$

we get:

$$\hat{A}_k \cong \frac{\xi_k}{1+\xi_k} R_k \qquad high \quad SNR \tag{12}$$

$$\triangleq A_k^w$$

Since we estimate the spectral component $X_k = A_k exp(j\alpha_k)$ by $\hat{X}_k = \hat{A}_k \exp(j\vartheta_k)$, where $\exp(j\vartheta_k)$ is the complex exponential of the noisy phase (see Section IV), we get from (12) the following approximation for the k-th spectral component estimator:

$$\hat{X}_k \cong \frac{\xi_k}{1+\xi_k} Y_k \qquad high \quad SNR \tag{13}$$

$$\triangleq X_k^w$$

This estimator is in fact the optimal MMSE estimator of the k-th spectral component, i.e., the Wiener estimator. For this reason, (12) is referred as a Wiener amplitude estimator.

It is useful to consider the optimal amplitude estimator $\hat{A}_k$ in (7), as being obtained from R_k, by a multiplicative non-linear optimal gain function which is defined by:

$$G_{opt}(\xi_k, \gamma_k) \triangleq \frac{\hat{A}_k}{R_k} \tag{14}$$

From (7) we see that this gain function depends only on the a-priori and the a-posteriori SNR, ξ_k and γ_k, respectively. Several gain curves which result from (7) and (14) are shown in Fig. 1. $\gamma_k - 1$ in this figure is interpreted as the 'instantaneous SNR', since $\gamma_k \triangleq R_k^2 / \lambda_d(k)$, and R_k is the modulus of the signal plus noise resultant spectral component.

The gain curves in Fig. 1 show an increase in gain as the instantaneous SNR $\gamma_k - 1$ decreases, while the a-priori SNR ξ_k is kept constant. This

© Technion - Israel Institute of Technology, Elyachar Central Library

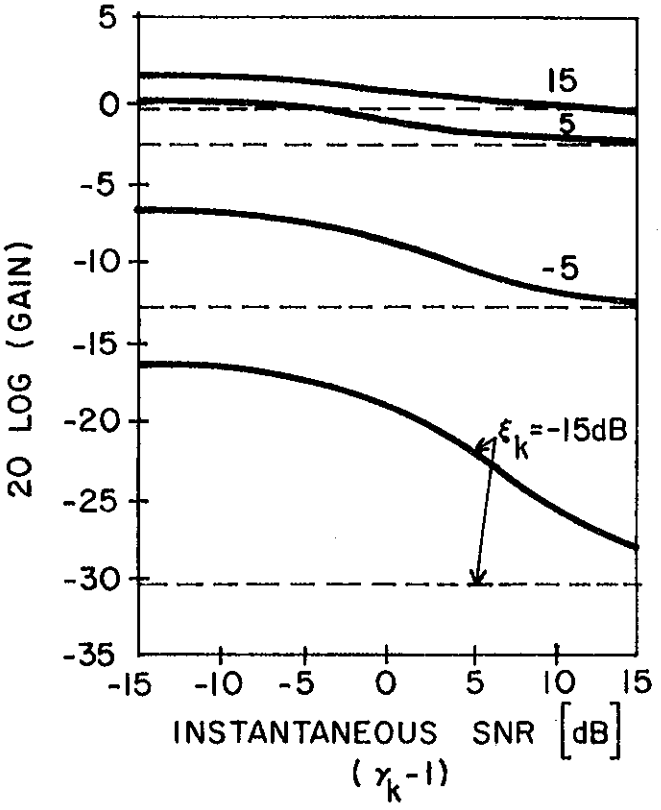


Fig. 1: Parametric gain curves describing:
(a) - Optimal gain function $G_{opt}(\xi_k, \gamma_k)$ defined by (7) and (14), (bold lines).
(b) - Wiener gain function $G_w(\xi_k, \gamma_k)$ defined by (15), (dashed line).

<u>ציור 1</u> : עקומי הגבר פרמטריים המתארים:
(a) - פונקצית ההגבר האופטימלי $G_{opt}(\xi_k,\gamma_k)$ המוגדרת ע"י (7) ו-(14) (קו מלא).
(b) - פונקצית ההגבר הוינרית $G_\omega(\xi_k,\gamma_k)$ המוגדרת ע"י (15) (קו מרוסק).

© Technion - Israel Institute of Technology, Elyachar Central Library

behavior is explained below, on the basis of the fact that the optimal estimator finds a compromise between what it knows from the a-priori information, and what it learns from the noisy data.

Let ξ_k result from some fixed values of $\lambda_x(k)$ and $\lambda_d(k)$. The fixed value of $\lambda_x(k)$ determines the most probable realizations of A_k, which are considered by the optimal estimator. This is due to the fact that $\lambda_x(k)$ is the only parameter of $p(a_k)$ (see (6)). On the other hand, the fixed value of $\lambda_d(k)$ makes γ_k to be proportional to R_k, since $\gamma_k \overset{\Delta}{=} R_k^2 / \lambda_d(k)$. Therefore, as γ_k decreases and ξ_k is fixed, the estimator should compromise between the most probable realizations of A_k, and the decreasing values of R_k. Since A_k is estimated by $\hat{A}_k = G_{opt}(\xi_k, \gamma_k) R_k$, this can be done by increasing $G_{opt}(\xi_k, \gamma_k)$.

Figure 1 shows also several gain curves corresponding to the Wiener gain function which results from the amplitude estimator (12). This gain function is given by:

$$G_w(\xi_k, \gamma_k) = \frac{\xi_k}{1+\xi_k} \tag{15}$$

and it is independent of γ_k. The convergence of the optimal and the Wiener amplitude estimators at high SNR, is clearly demonstrated in this figure.

It is interesting to note that the same gain curves as those belonging to the optimal gain function $G_{opt}(\xi_k, \gamma_k)$, were obtained by a 'vector spectral subtraction' amplitude estimation approach [13]. In this approach, the amplitude estimator is obtained from two mutually dependent optimal MMSE estimators, of the amplitude and the cosine of the phase error, (i.e., the phase $\vartheta_k - \alpha_n$). Since an estimator of the cosine of the phase error is used for estimating the amplitude, this approach is interpreted as a 'vector spectral subtraction' amplitude estimation. The derivation of the amplitude estimator by the above approach, is presented in Appendix B.

© Technion - Israel Institute of Technology, Elyachar Central Library

Error Analysis and Sensitivity

The optimal amplitude estimator (7) is derived under the implicit assumption that the a-priori SNR ξ_k and the noise variance $\lambda_d(k)$ are known. However, in the speech enhancement problem discussed here, these parameters are unknown in advance, as the noisy speech alone is available. Therefore they are replaced by their estimators in a practical system. For this reason it is of interest to examine the sensitivity of the optimal amplitude estimator to inaccuracy in these parameters.

We found that the a-priori SNR is a key parameter in the discussed problem, rather than the noise variance which is easier to estimate. Therefore, we examine here the sensitivity of the optimal amplitude estimator to the a-priori SNR ξ_k only. In addition, for similar reasons, we are interested here especially in the sensitivity at low a-priori SNR (i.e., $\xi_k \ll 1$).

We present here a sensitivity analysis which is based on the calculation of the mean square error (MSE) and the bias associated with the optimal estimator (7), when the a-priori SNR ξ_k is perturbated. This sensitivity analysis provides also an error analysis, since the latter turns out to be a particular case of the former.

A similar problem to the above one, arises in the Wiener amplitude estimator, which depends on the a-priori SNR parameter (see (12)). Since the Wiener estimator is widely used in speech enhancement systems, we give here a sensitivity analysis for this estimator as well, and compare it with the optimal amplitude estimator.

Let ξ_k^* denote the nominal a-priori SNR, and $\tilde{\xi}_k \triangleq \xi_k^* + \Delta\xi_k$ denote its perturbated version. The optimal amplitude estimator which uses the perturbated ξ_k is obtained from (7), and is given by:

© Technion - Israel Institute of Technology, Elyachar Central Library

$$\hat{A}_k = \Gamma(1.5)\frac{\sqrt{\tilde{v}_k}}{\gamma_k}M(-0.5;\,1;\,-\tilde{v}_k)R_k \tag{16}$$

where $\tilde{v}_k$ is defined by:

$$\tilde{v}_k = \frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\gamma_k, \tag{17}$$

Similarly, the Wiener amplitude estimator with the perturbated ξ_k is obtained from (12), and is given by:

$$A_k^w = \frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}R_k, \tag{18}$$

To calculate the residual MSE in the optimal amplitude estimation (16) for low a-priori SNR values, it is most convenient to expand $M(a;c;x)$ in (16) by the following series [4: A.1.14]:

$$M(a,c,x) = \sum_{r=0}^{\infty}\frac{(a)_r}{(c)_r}\frac{x^r}{r!} \tag{19}$$

$$= 1+\frac{a}{c}\frac{x}{1!}+\frac{a(a+1)}{c(c+1)}\frac{x^2}{2!}+\cdots$$

where $(a)_r \triangleq a(a+1)\cdots(a+r-1)$, $(a)_0 \triangleq 1$. By so doing, and using the fact that γ_k is exponentially distributed, i.e.,

$$p(\gamma_k) = \frac{1}{1+\xi_k^*}\exp\left(-\frac{\gamma_k}{1+\xi_k^*}\right) \qquad \gamma_k \ge 0 \tag{20}$$

we get the normalized residual MSE $\varepsilon_{opt}(\xi_k^*,\tilde{\xi}_k)$ by:

$$\varepsilon_{opt}(\xi_k^*,\tilde{\xi}_k) \triangleq E\{[A_k-\hat{A}_k]^2\}/E\{[A_k-E(A_k)]^2\} \tag{21}$$

$$= \frac{1}{1-\pi/4}\Bigg\{1+\frac{\pi}{4}\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\frac{1}{\xi_k^*}\sum_{r,l=0}^{\infty}\frac{(-0.5)_r(-0.5)_l}{(1)_r(1)_l}\frac{1}{r!l!}\left[\frac{1+\xi_k^*}{1+\tilde{\xi}_k}\right]^{r+l}(-\tilde{\xi}_k)^{r+l}\Gamma(r+l+1)$$

$$-2\frac{\pi}{4}\left[\frac{\xi_k^*}{1+\xi_k^*}\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\right]^{1/2}\frac{1}{\xi_k^*}\sum_{r,l=0}^{\infty}\frac{(-0.5)_r(-0.5)_l}{(1)_r(1)_l}\frac{1}{r!l!}\left[\frac{1+\xi_k^*}{1+\tilde{\xi}_k}\right]^{l}(-\xi_k^*)^r(-\tilde{\xi}_k)^l\Gamma(r+l+1)\Bigg\}$$

It can be shown by using Lebesgue monotonic convergence theorem, and Lebesgue dominated convergence theorem [14], that the commutation of the expectation and limit operations needed in the calculation of (21) are valid for $\xi^*<1$ and $\tilde{\xi}<(1-\xi^*)/2\xi^*$. Therefore, the resulting expression in (21) is also valid in that domain.

© Technion - Israel Institute of Technology, Elyachar Central Library

The normalized residual MSE $\varepsilon_w(\xi_k^*,\tilde{\xi}_k)$ resulting in the Wiener amplitude estimation (18), can be calculated similarly, and is given by:

$$\varepsilon_w(\xi_k^*,\tilde{\xi}_k) \triangleq E\{[A_k - A_k^w]^2\} / E\{[A_k - E(A_k)]^2\} \tag{22}$$

$$= \frac{1}{1-\pi/4}\left\{1 + \left[\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\right]^2 \left[\frac{1+\xi_k^*}{\xi_k^*}\right]\right.$$

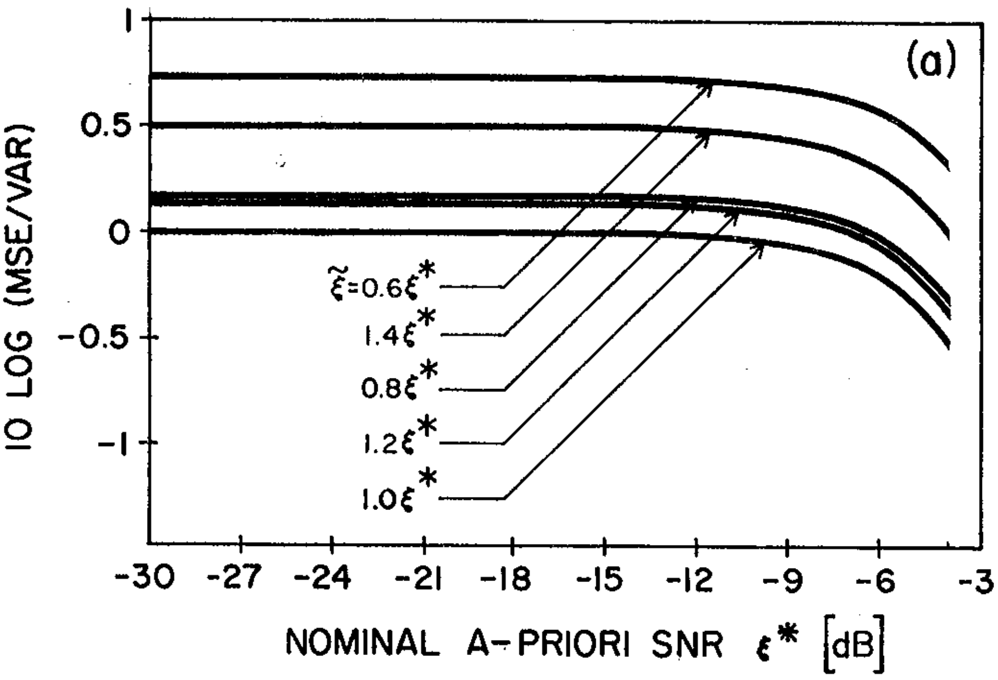
$$\left. - 2\Gamma(1.5)\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\,\frac{1}{(\xi_k^*)^{1/2}}\sum_{r=0}^{\infty}\frac{(-0.5)_r}{(1)_r}\,\frac{1}{r!}\,(-\xi_k^*)^r\,\Gamma(r+1.5)\right\}$$

which is valid for $\xi^* < 1$.

For low SNR the above expressions can also be used to calculate the nominal residual MSE, which corresponds to the MSE when the a-priori SNR is known exactly. This can be done by substituting $\tilde{\xi}_k = \xi_k^*$ in (21) and (22).

For very low SNR values, $\varepsilon_{opt}(\xi_k^*,\tilde{\xi}_k)$ and $\varepsilon_w(\xi_k^*,\xi_k)$ can be approximated by considering terms of up to third order only in the infinite sums of (21) and (22). Fig. 2 shows the residual MSE obtained in this way, as a function of the nominal a-priori SNR ξ_k^*, and for several values of $\Delta\xi_k / \xi_k^*$. A number of conclusions can be drawn now: First note from (21) and (22) that the nominal normalized MSE in the optimal estimation cannot be greater than unity, while in the Wiener amplitude estimation, it can be as high as $1/(1-\pi/4)$. Second, both estimators seem to be insensitive to small perturbations in the nominal a-priori SNR ξ_k^* value. Finally, it is interesting to note that both estimators are more sensitive to under-estimates of the a-priori SNR than to its over-estimates. In addition, by using an over-estimate of ξ_k^* in the Wiener amplitude estimation, the residual MSE decreases. This surprising fact can be explained by noting that the Wiener estimator is not an optimal amplitude estimator. Therefore, using an erroneous value of ξ_k^*, can either increase or decrease the MSE.

© Technion - Israel Institute of Technology, Elyachar Central Library



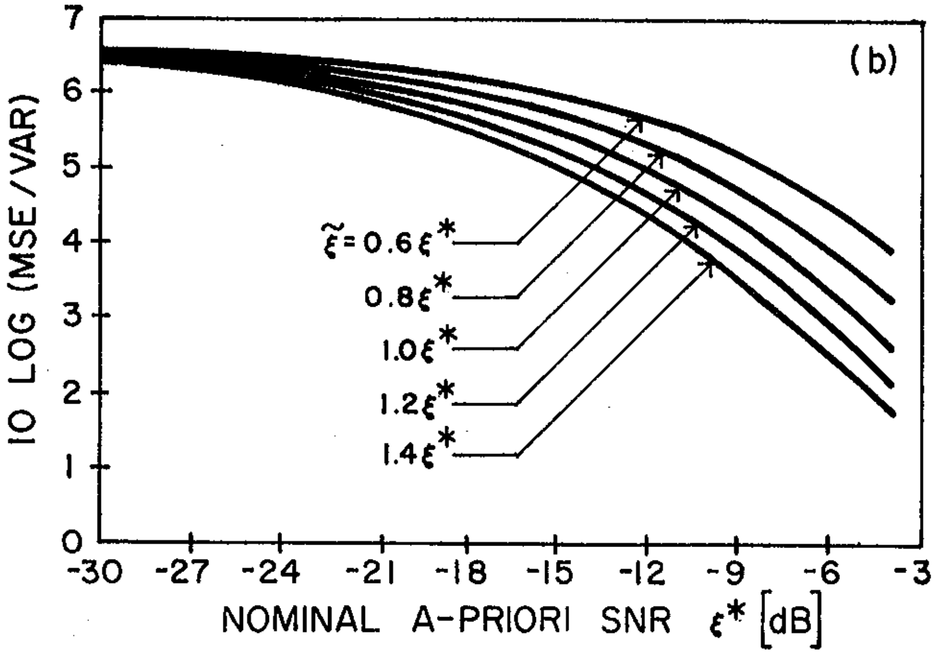


Fig. 2: Normalized MSE in amplitude estimations for perturbated values of the a-priori SNR.
(a) - Optimal estimator (Eq. (16))
(b) - Wiener estimator (Eq. (18))

<u>ציור 2</u> : שגיאה ריבועית ממוצעת מנורמלת בשערוכי האמפליטודה, עבור ערכים שונים של אי דיוק בערך יחס האות לרעש הא-פריורי.

(a) - משערך אופטימלי (16)

(b) - משערך וינר (18)

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

We turn now to the calculation of the normalized bias of each estimator, when the a-priori SNR is perturbated. The normalized bias is defined here as the ratio between the expected value of the amplitude estimation error, and the expected value of the amplitude.

The normalized bias $B_{opt}(\xi_k^{*},\tilde{\xi}_k)$ of the optimal estimator at low a-priori SNR, is obtained by using (16), (19), and (20). It is equal to:

$$B_{opt}(\xi_k^{*},\tilde{\xi}_k) \triangleq E\{A_k - \hat{A}_k\} / E\{A_k\} \tag{23}$$

$$= 1 - \left[\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\,\frac{1}{\xi_k^{*}}\right]^{1/2} \sum_{r=0}^{\infty} \frac{(-0.5)_r}{(1)_r} \left[\frac{1+\xi_k^{*}}{1+\tilde{\xi}_k}\right]^{r} (-\tilde{\xi}_k)^{r}$$

and is valid for $\xi\tilde{\xi}<1$. The normalized bias $B_w(\xi_k^{*},\tilde{\xi}_k)$ of the Wiener amplitude estimator, is easily obtained from (18), and is given by:

$$B_w(\xi_k^{*},\tilde{\xi}_k) \triangleq E\{A_k - A_k^{w}\} / E\{A_k\} \tag{24}$$

$$= 1 - \frac{\tilde{\xi}_k}{1+\tilde{\xi}_k} \left[\frac{1+\xi_k^{*}}{\xi_k^{*}}\right]^{1/2}$$

Fig. 3 shows the bias of the optimal and the Wiener amplitude estimators, as a function of ξ_k^{*}. $B_{opt}(\xi_k^{*},\tilde{\xi}_k)$ in this figure is calculated by using terms of up to the third order in the infinite sum in (23).

III. OPTIMAL AMPLITUDE ESTIMATOR UNDER UNCERTAINTY OF SIGNAL PRESENCE

In this section we derive the optimal MMSE amplitude estimator under the assumed Gaussian statistical model, and uncertainty of signal presence in the noisy observations. By so doing we extend the optimal amplitude estimator derived in Section II, as will be clarified later.

Signal absence in the noisy observations $\{y(t),\ 0 \le t \le T\}$ is frequent, since speech signals contain large portions of silence. This absence of signal implies its absence in the noisy spectral components as well. However, it is also possible that the signal is present in the noisy observations, but appears with insignificant energy in some noisy spectral components,

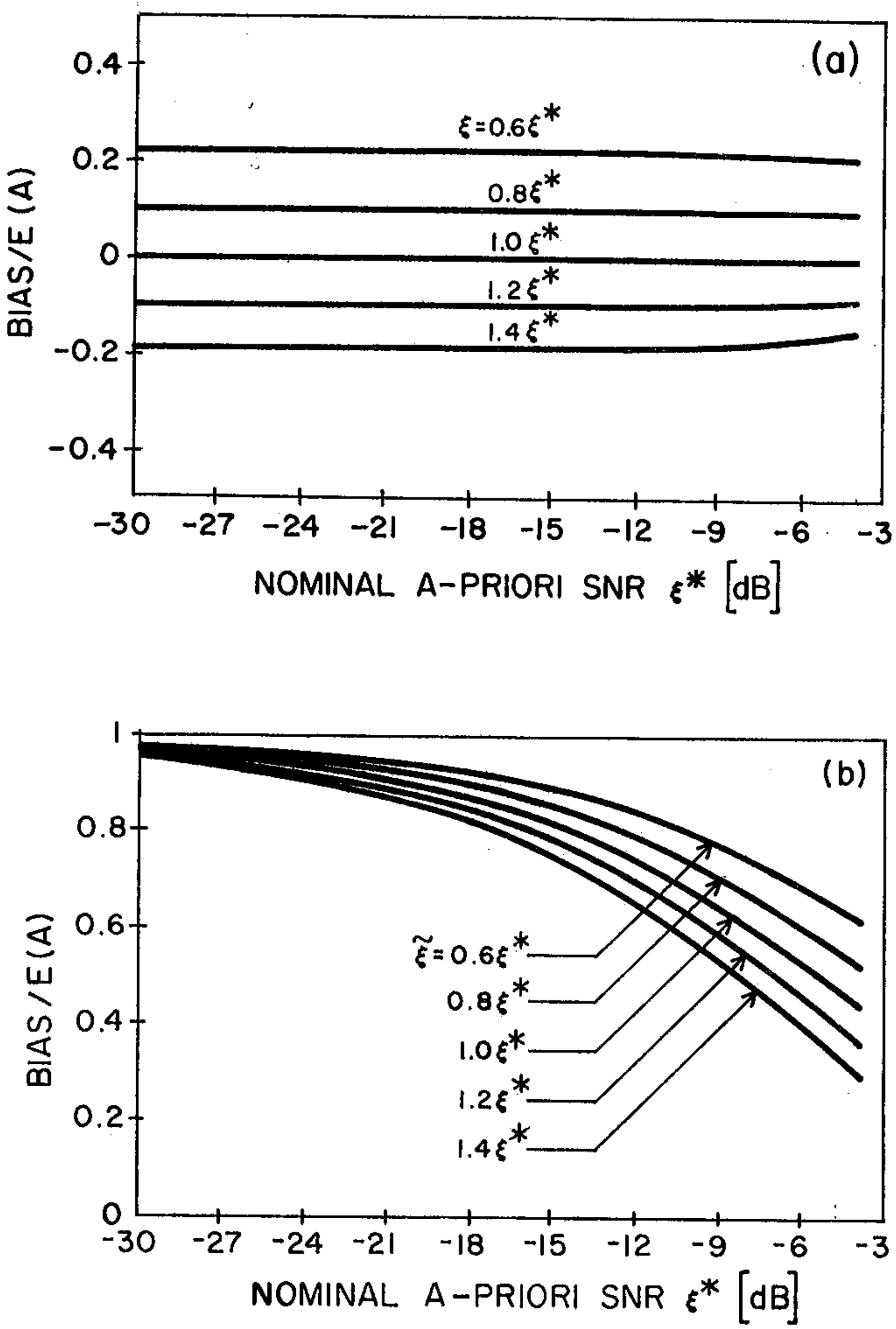


Fig. 3: Normalized bias of amplitude estimators for perturbated values of the a-priori SNR.
(a) - Optimal estimator (Eq. (16))
(b) - Wiener estimator (Eq. (18))

ציור 3 : הטיה מנורמלת של משערכי האמפליטודה עבור ערכים שונים של אי-דיוק בערך יחס האות לרעש הא-פריורי.

(a) - משערך אופטימלי (16)

(b) - משערך וינר (18)

© Technion - Israel Institute of Technology, Elyachar Central Library

which are randomly determined. This is a typical situation when the analyzed speech is of voiced type, and the analysis is not synchronized with the pitch period.

The above discussion suggests two statistical models for speech absence in the noisy spectral components. In the first one, speech is assumed to be either present or absent, with given probabilities, in all of the noisy spectral components. The reasoning behind this model is that signal presence or absence should be the same in all of the noisy spectral components, since the analysis is done on a finite interval. In the second model which represents the other extreme, a statistically independent random appearance of the signal in the noisy spectral components is assumed. As is implied by the above discussion, this model is more appropriate for voiced speech signals, when weak signal spectral components are considered as if they were absent.

These two models, and the resulting optimal amplitude estimators based upon them, are examined in details in [8]. We found that the estimator whose derivation is based on the second model, is especially successful in speech enhancement applications. Therefore, we present in this section its derivation, and leave for Appendix C the discussion concerning the optimal estimator which is based on the first model.

The idea of utilizing the uncertainty of signal presence in the noisy spectral components, for improving speech enhancement results, was first proposed by McAulay and Malpass [3]. In their work they actually capitalize on the above second model of signal absence, and modify appropriately a ML amplitude estimator. In Section VI we compare the McAulay and Malpass amplitude estimator, with the one which we derive here, in enhancing speech.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

Derivation of Amplitude Estimator Under Signal Presence Uncertainty

The optimal MMSE estimator which takes into account the uncertainty of signal presence in the noisy observations, was developed by Middleton and Esposito [7]. Based on our second model for signal absence, in which statistically independent random appearance of the signal in the noisy spectral components is assumed, and on the statistical independence of the spectral components assumed in Section I, this optimal estimator is given by [8] (see Appendix C):

$$\hat{A}_k = \frac{\Lambda(Y_k, q_k)}{1 + \Lambda(Y_k, q_k)} E\{A_k \mid Y_k, H_k^1\} \tag{25}$$

where $\Lambda(Y_k, q_k)$ is the generalized likelihood ratio defined by:

$$\Lambda(Y_k, q_k) = \mu_k \frac{p(Y_k \mid H_k^1)}{p(Y_k \mid H_k^0)} \tag{26}$$

with $\mu_k \triangleq (1-q_k)/q_k$, and q_k is the probability of signal absence in the k-th spectral component. H_k^0 and H_k^1 denote the two hypotheses of signal absence and presence, respectively, in the k-th spectral component. $E\{A_k \mid Y_k, H_k^1\}$ is the optimal MMSE amplitude estimator, when the signal is surely present in the k-th spectral component. This is in fact the optimal estimator (7). Therefore, in order to derive the new optimal estimator (25), we need to calculate the additional function $\Lambda(Y_k, q_k)$ only. This can be easily done by using the Gaussian statistical model assumed for the spectral components, or equivalently, by using (5) and (6). We get:

$$\Lambda(Y_k, q_k) = \mu_k \frac{\exp(v_k)}{1 + \xi_k} \tag{27}$$

where ξ_k in (27) is now defined by:

$$\xi_k \triangleq \frac{E\{A_k^2 \mid H_k^1\}}{\lambda_d(k)} \tag{28}$$

This definition agrees with its previous definition in (9), since there the signal is assumed to be surely present in the noisy spectral components.

© Technion - Israel Institute of Technology, Elyachar Central Library

It is more convenient to make $\Lambda(Y_k, q_k)$ and the resulting amplitude estimator, a function of $\eta_k \stackrel{\Delta}{=} E\{A_k^2\}/\lambda_d(k)$ which is easier to estimate than ξ_k. η_k is related to ξ_k by:

$$\eta_k \stackrel{\Delta}{=} \frac{E\{A_k^2\}}{\lambda_d(k)} \tag{29}$$

$$= (1-q_k)\,\frac{E\{A_k^2\,|\,H_k^1\}}{\lambda_d(k)}$$

$$= (1-q_k)\,\xi_k$$

Thus, by considering $\Lambda(Y_k, q_k)$ in (27) as $\Lambda(\xi_k, \gamma_k, q_k)$, and using $E\{A_k\,|\,y_k, H_k^1\} = G_{opt}(\xi_k, \gamma_k)R_k$, where $G_{opt}(\xi_k, \gamma_k)$ is the gain function defined by (7) and (14), the optimal amplitude estimator (25) can be written as:

$$\hat{A}_k = \frac{\Lambda(\xi_k, \gamma_k, q_k)}{1+\Lambda(\xi_k, \gamma_k, q_k)} G_{opt}(\xi_k, \gamma_k) R_k \Big|_{\xi_k = \frac{\eta_k}{1-q_k}} \tag{30}$$

$$\stackrel{\Delta}{=} G_{opt}^D(\eta_k, \gamma_k, q_k) R_k$$

Note that if $q_k=0$, then $\Lambda/(1+\Lambda)$ in (30) equals unity, and also $\eta_k = \xi_k$. In this case $G_{opt}^D(\eta_k, \gamma_k, q_k)$ turns out to be equal to $G_{opt}(\xi_k, \gamma_k)$. Thus the optimal amplitude estimator (7) can be considered as a particular case of the optimal amplitude estimator (30).

Several gain curves which result from $G_{opt}^D(\eta_k, \gamma_k, q_k)$ in (30), are described in Fig. 4 for $q_k=0.2$. It is interesting to compare these gain curves with those of $G_{opt}(\xi_k, \gamma_k)$ which are depicted in Fig. 1. Especially it is intersting to see how the trend of the gain curves in each pair, corresponding to the same value of the a-priori SNR, changes, as this value increases. The decrease in gain as γ_k decreases and η_k is high, for the case in which $q_k > 0$, is in contrast to the increase in gain for the case in which $q_k=0$ (i.e., for $G_{opt}(\xi_k, \gamma_k)$). This is probably a result of favoring the hypothesis of signal absence by the amplitude estimator (30) in such a situation.

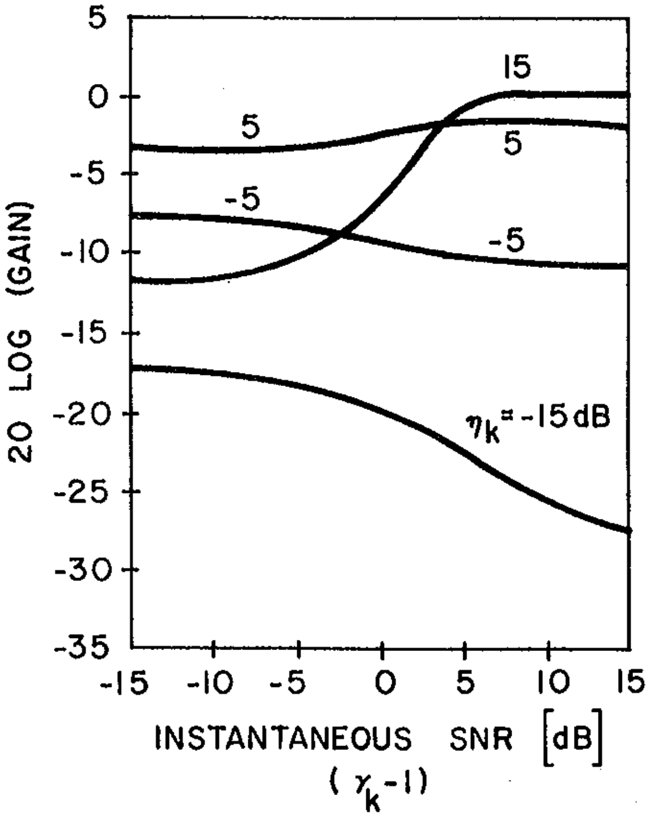


Fig. 4: Parametric gain curves describing the optimal gain function $G_{opt}^{D}(\eta_k, \gamma_k, q_k)$ defined by (30) for $q_k = 0.2$.

<u>ציור 4</u> : עקומי הגבר פרמטריים המתארים את פונקצית ההגבר האופטימלית $G_{opt}^{D}(\eta_k, \gamma_k, q_k)$ המוגדרת ע"י (30) עבור $q_k = 0.2$.

© Technion - Israel Institute of Technology, Elyachar Central Library

We conclude this section by noting that the estimator (25) which takes into account the signal presence uncertainty, could be obtained from the estimator (4) which assumes that the signal is surely present, if $p(a_k,\alpha_k)$ in (4) is chosen appropriately. This can be done by using

$$p(a_k,\alpha_k) = (1-q_k)p(a_k,\alpha_k \mid H_k^1)+q_k\delta(a_k,\alpha_k), \tag{31}$$

where $p(a_k,\alpha_k \mid H_k^1)$ is the common PDF of A_k and α_k when the signal is surely present and $\delta(a_k,\alpha_k)$ is a Dirac function. Under the Gaussian assumption used here, $p(a_k,\alpha_k \mid H_k^1)$ is given by (6). This is an interesting interpretation of the estimator (25), which was originally derived in [7] by minimizing the mean square estimation error (see Appendix C). It also indicates that the estimator derived by using (4) with the above $p(a_k,\alpha_k)$, (or equivalently (25)), is optimal for a class of a-priori PDFs which differ in the probability assumed for signal absence.

IV. OPTIMAL MMSE COMPLEX EXPONENTIAL ESTIMATOR

In the previous sections we gave the motivation for using the optimal STSA estimator of the speech signal, and derived it. In this section we concentrate on the derivation of an optimal MMSE estimator of the complex exponential of the phase. This estimator is combined with the optimal STSA estimator for constructing the enhanced signal. We base the estimation on the same statistical model assumed in Section I, which was used in the derivation of the STSA estimator.

We show that the optimal complex exponential estimator has a non-unity modulus. Therefore, combining it with an optimal amplitude estimator, results in a new amplitude estimator which is no longer optimal. On the other hand, the optimal complex exponential estimator whose modulus is constrained to be unity, is the complex exponential of the noisy phase.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

We also show in this section that the optimal estimator of the principle value of the phase, is the noisy phase itself. This result is of interest, although it does not provide another estimator for the complex exponential than the above constrained one. Its importance follows from the fact that it is unknown which one, the phase or its complex exponential, is more important in speech perception. Therefore, the optimal estimators of both of them should be examined.

Derivation of Optimal Complex Exponential Estimator

Based on the statistical model assumed in Section I, the optimal MMSE estimator of the complex exponential $e^{j\alpha_k}$, given the noisy observations $\{y(t),\ 0 \le t \le T\}$, is given by:

$$\begin{aligned} e^{\widetilde{j\alpha_k}} &= E\{e^{j\alpha_k} \mid y(t),\ 0 \le t \le T\} \qquad (32) \\ &= E\{e^{j\alpha_k} \mid Y_0, Y_1, \cdots\} \\ &= E\{e^{j\alpha_k} \mid Y_k\} \\ &= E\{e^{-j\varphi_k} \mid Y_k\} e^{j\vartheta_k} \\ &= [E\{\cos\varphi_k \mid Y_k\} - jE\{\sin\varphi_k \mid Y_k\}] e^{j\vartheta_k} \end{aligned}$$

where φ_k is the phase error which is defined by $\varphi_k \overset{\Delta}{=} \vartheta_k - \alpha_k$, and ϑ_k is the noisy phase. $E\{\sin\varphi_k \mid Y_k\}$ and $E\{\cos\varphi_k \mid Y_k\}$ were calculated for the Gaussian statistical model assumed here. By using (5) and (6) we obtain (see Appendix D):

$$E\{\sin\varphi_k \mid Y_k\} = 0 \qquad (33)$$

and

$$\begin{aligned} e^{\widetilde{j\alpha_k}} &= E\{\cos\varphi_k \mid Y_k\} e^{j\vartheta_k} \qquad (34) \\ &= \Gamma(1.5)\sqrt{v_k}\ \ M(0.5;\ 2;\ -v_k)\, e^{j\vartheta_k} \\ &= \Gamma(1.5)\sqrt{v_k}\, exp\,(-v_k/2)[I_0(v_k/2) + I_1(v_k/2)] e^{j\vartheta_k} \end{aligned}$$

$\triangleq \widetilde{\cos\varphi_k}\, e^{j\vartheta_k}$

The combination of the optimal estimator $e^{\widetilde{j\alpha_k}}$ with an independently derived amplitude estimator $\hat{A}_k$, results in the following estimator $\tilde{X}_k$ for the k-th spectral component:

$$\tilde{X}_k = \hat{A}_k\, \widetilde{\cos\varphi_k}\; e^{j\vartheta_k} \tag{35}$$

The modulus of the spectral estimator $\tilde{X}_k$ represents now a new amplitude estimator which equals to $\hat{A}_k \widetilde{\cos\varphi_k}$. If $\hat{A}_k$ is an optimal estimator, (e.g., (7)), then the resulting new amplitude estimator is no longer optimal.

It is worthwhile to further investigate the estimator (35), when $\hat{A}_k$ is the optimal estimator from (7). We show now that this estimator is nearly equivalent to the Wiener spectral estimator X_k^w, which is given by (13). On the one hand, this fact implies that $\tilde{X}_k$ is a nearly optimal MMSE spectral estimator, since the Wiener spectral estimator is an optimal one. On the other hand, this fact enables to estimate the degradation in the amplitude estimation, by using the error analysis of the previous section.

To show that $\tilde{X}_k$ in (35) and X_k^w in (13) are nearly equivalent, we compare their gain curves for the SNR values which are of interest here. Several of these gain curves are shown in Fig. 5. The closeness of the gain curves which correspond to the same value of ξ_k, implies that the two estimators $\tilde{X}_k$ and X_k^w are nearly equivalent.

Due to the major importance of the STSA in speech perception, it is of interest to derive an optimal estimator of the complex exponential of the phase, which does not affect the amplitude estimation.

To derive the above estimator, which we denote by $e^{\widehat{j\alpha_k}}$, we solve the following constrained optimization problem

© Technion - Israel Institute of Technology, Elyachar Central Library

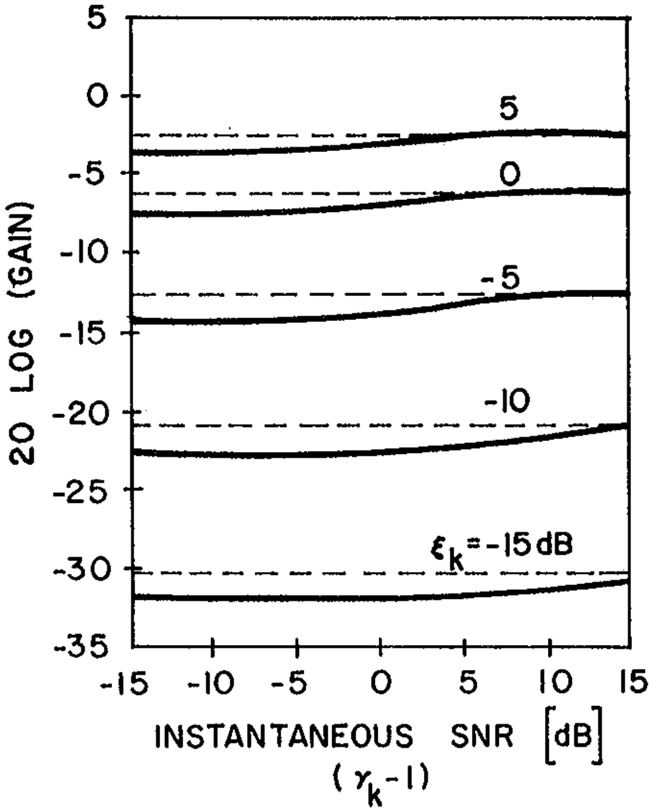


Fig. 5: Parametric gain curves resulting from:
(a) - Optimal amplitude estimator combined with optimal complex exponential estimator (Eq. (35)).
(b) - Wiener gain function (Eq. (15)).

ציור 5 : עקומי הגבר המתקבלים מתוך:
(a) - צרוף משערך האמפליטודה האופטימלי והמשערך האופטימלי של האקספוננט הקומפלקסי (35).
(b) - פונקצית ההגבר הוינרית (15).

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

$$\min_{e^{j\hat{\alpha}_k}} E\{|e^{j\alpha_k} - e^{j\hat{\alpha}_k}|^2\} \tag{36}$$

$$\textit{subject to } |e^{j\hat{\alpha}_k}| = 1$$

Using the Lagrange multipliers method, we get (see Appendix D):

$$e^{j\hat{\alpha}_k} = e^{j\vartheta_k} \tag{37}$$

That is, the complex exponential of the noisy phase, is the best MMSE complex exponential estimator which does not affect the amplitude estimation.

Optimal Phase Estimator

The optimal estimator of the principle value of the phase is derived here by minimizing the following distortion measure [15]:

$$E\{1 - \cos(\alpha_k - \hat{\alpha}_k)\} \tag{38}$$

This measure is invariant under modulo 2π transformation of the phase α_k, the estimated phase $\hat{\alpha}_k$, and the estimation error $\alpha_k - \hat{\alpha}_k$. For small estimation errors, (38) is a type of least square criterion, since $1-\cos\beta \cong \beta^2/2$ for $\beta \ll 1$.

The optimal estimator $\hat{\alpha}_k$ which minimizes (38), is easily shown to satisfy:

$$tg\ \hat{\alpha}_k = \frac{E\{\sin\alpha_k \mid Y_k\}}{E\{\cos\alpha_k \mid Y_k\}} \tag{39}$$

By using $\alpha_k = \vartheta_k - \varphi_k$, and $E\{\sin\varphi_k \mid Y_k\} = 0$ (see (33)), it is easy to see that:

$$E\{\sin\alpha_k \mid Y_k\} = \sin\vartheta_k \widetilde{\cos\varphi_k} \tag{40}$$

$$E\{\cos\alpha_k \mid Y_k\} = \cos\vartheta_k \widetilde{\cos\varphi_k} \tag{41}$$

On substituting (40) and (41) into (39), we get:

$$tg\ \hat{\alpha}_k = tg\ \vartheta_k \tag{42}$$

or alternatively $\hat{\alpha}_k = \vartheta_k$.

V. VARIANCE ESTIMATION OF SPECTRAL COMPONENTS

In this section we address the problem of estimating the variance of a signal spectral component, and of a noise spectral component. The estimators of these variances are used for estimating the a-priori SNR, which is a parameter of the optimal and Wiener STSA estimators. Due to non-stationarity of the speech signal and possibly also of the noise process, these variances are time varying, and therefore they should be re-estimated for each analysis frame.

We examine here the estimation of the variance of a signal spectral component only. The easier problem of estimating the variance of a noise spectral component is well treated in the literature (e.g., [16,17]), and therefore it is not discussed here. The general approach is to estimate the noise spectral components variances from non-speech intervals, which are most adjacent in time to the analyzed frame. This approach is of course suitable only if the noise is 'stationary' over a sufficiently long time interval.

Two approaches for estimating a signal spectral component variance are considered here. The first is based on a ML estimation approach, and the second is based on a 'decision-directed' estimation method. We present here the derivation of each estimator, and leave for the next section the discussion concerning its application and performance in the proposed speech enhancement system.

Maximum Likelihood Estimation Approach

The ML estimation approach is most commonly used for estimating an unknown parameter of a given PDF (e.g., $\lambda_x(k)$ in (6)), when no a-priori information about it is available. We derive now the ML estimator of the k-th signal spectral component variance in the n-th analysis frame. We base

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

the estimation on L consecutive observations $\underline{Y}_k(n) \overset{\Delta}{=} \{Y_k(n), Y_k(n-1), \ldots, Y_k(n-L+1)\}$, which are assumed to be statistically independent. This assumption is reasonable when the analysis is done on non-overlapping frames. However, in the system used here overlapping is done (see Section VI). Nevertheless, we continue with this assumption, since the statistical dependence is difficult to be modeled and handled. We also assume that the signal and noise k-th spectral component variances, $\lambda_x(k)$ and $\lambda_d(k)$ respectively, are slowly varying parameters, so that they can be considered constant during the above L observations. Finally, we assume that the k-th noise spectral component variance is known.

The ML estimator $\hat{\lambda}_x(k)$ of $\lambda_x(k)$, which is constrained to be non-negative, is the non-negative argument which maximizes the joint conditional PDF of $\underline{Y}_k(n)$ given $\lambda_x(k)$ and $\lambda_d(k)$. Based on the Gaussian statistical model and the statistical independence assumed for the spectral components, this PDF is given by:

$$p(\underline{Y}_k(n) \mid \lambda_x(k), \lambda_d(k)) = \prod_{l=0}^{L-1} \frac{1}{\pi(\lambda_x(k)+\lambda_d(k))} \exp\left(-\frac{R_k^2(n-l)}{\lambda_x(k)+\lambda_d(k)}\right) \tag{43}$$

where $R_k(l) \overset{\Delta}{=} |Y_k(l)|$. $\hat{\lambda}_x(k)$ is easily obtained from (43), and equals to:

$$\hat{\lambda}_x(k) = \begin{cases} \frac{1}{L}\sum_{l=0}^{L-1} R_k^2(n-l) - \lambda_d(k) & \textit{if non-negative} \\ 0 & \textit{otherwise} \end{cases} \tag{44}$$

This estimator suggests the following estimator for the a-priori SNR ξ_k.

$$\hat{\xi}_k = \begin{cases} \frac{1}{L}\sum_{l=0}^{L-1} \gamma_k(n-l) - 1 & \textit{if non-negative} \\ 0 & \textit{otherwise} \end{cases} \tag{45}$$

where $\gamma_k(l) \overset{\Delta}{=} |Y_k(l)|^2 / \lambda_d(k)$ is the a-posteriori SNR in the l-th analysis frame. Note that the estimator (45) assumes the knowledge of $\lambda_d(k)$, which is needed to calculate $\gamma_k(l)$. In practice $\lambda_d(k)$ is estimated independently (as discussed above), and substituted in (45).

It is interesting to consider the ML estimator (44) when $L=1$. In this case we get the 'power spectral subtraction" estimator derived in [3]. The application of the corresponding ξ_k estimator (45) (with $L=1$) to the optimal estimator (7), results in a gain function which depends on γ_k only. Surprisingly, this gain function is almost identical to the 'spectral subtraction' gain function for a wide range of SNR values. The 'spectral subtraction' gain function is given below by (46) [1], and the above nearly equivalence occurs when $\beta=1$.

$$G_{SP}(\gamma_k) \triangleq \frac{A_{SP}(\gamma_k)}{R_k} \tag{46}$$

$$= \sqrt{1-\frac{\beta}{\gamma_k}}, \qquad \beta\geq 1$$

This fact is demonstrated in Fig. 6. For comparison purposes, the same figure shows also the gain curve for the Wiener amplitude estimator, which results from (12) and the same a-priori SNR estimator (i.e., (45) with $L=1$).

In practice, the running average needed in (45), is replaced by a recursive averaging with a time constant comparable to the correlation time of γ_k. That is, the estimator of ξ_k in the n-th analysis frame is obtained by:

$$\bar{\gamma}_k(n) = \alpha\bar{\gamma}_k(n-1)+(1-\alpha)\frac{\gamma_k(n)}{\beta}, \qquad 0\leq\alpha<1, \quad ,\beta\geq 1. \tag{47}$$

$$\hat{\xi}_k(n) = \begin{cases} \bar{\gamma}_k(n)-1 & \bar{\gamma}_k(n)-1\geq 0 \\ 0 & otherwise \end{cases}$$

β is a correction factor, and it plays here the same role as in the 'spectral subtraction' estimator (46). The values of α and β are determined by informal listening, as is explained in Section VI.

'Decision-Directed' Estimation Approach

We consider now the estimation of a signal spectral component variance by a 'decision-directed' method. This estimator is found to be very

© Technion - Israel Institute of Technology, Elyachar Central Library

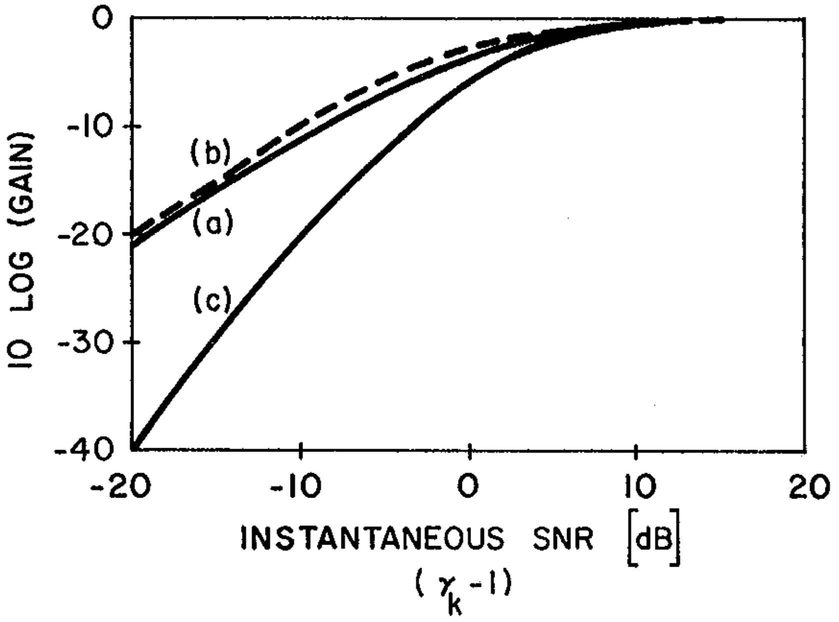


Fig. 6: Gain curves describing:
(a) - Optimal gain function $G_{opt}(\xi_k, \gamma_k)$ defined by (7) and (14), with $\xi_k = \gamma_k - 1$.
(b) - 'Spectral subtraction' gain function (46) with $\beta = 1$.
(c) - Wiener gain function $G_w(\xi_k, \gamma_k)$ (Eq. (15)) with $\xi_k = \gamma_k - 1$.

ציור 6 : עקומי הגבר המתארים:

(a) - פונקצית הגבר אופטימלית $G_{opt}(\xi_k,\gamma_k)$ המוגדרת ע"י (7) ו-(14), עם $\xi_k = \gamma_k - 1$.

(b) - פונקצית ההגבר בשיטת ההחסרה הספקטרלית (46) עם $\beta = 1$.

(c) - פונקצית הגבר וינרית $G_\omega(\xi_k,\gamma_k)$ (15) עם $\xi_k = \gamma_k - 1$.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

useful when it is combined with either the optimal or the Wiener amplitude estimator.

As done previously, we assume that the variance of each noise spectral component is known, and therefore we discuss the equivalent problem of estimating the a-priori SNR of a spectral component. In practice, an estimator of the noise variance is used in the resulting estimator.

Let $\xi_k(n)$, $A_k(n)$, $\lambda_d(k,n)$, and $\gamma_k(n)$, denote the a-priori SNR, the amplitude, the noise variance, and the a-posteriori SNR, respectively, of the corresponding k-th spectral component in the n-th analysis fame. The derivation of the a-priori SNR estimator is based here on the definition of $\xi_k(n)$, and its relation to the a-posteriori SNR $\gamma_k(n)$, as given below:

$$\xi_k(n) = \frac{E\{A_k^2(n)\}}{\lambda_d(k,n)} \tag{48}$$

$$\xi_k(n) = E\{\gamma_k(n)-1\} \tag{49}$$

Using (48) and (49) we can write:

$$\xi_k(n) = E\{\frac{1}{2}\,\frac{A_k^2(n)}{\lambda_d(k,n)}+\frac{1}{2}\,[\gamma_k(n)-1]\} \tag{50}$$

The proposed estimator $\hat{\xi}_k(n)$ of $\xi_k(n)$ is deduced from (50), and is given by:

$$\hat{\xi}_k(n) = \alpha\frac{\hat{A}_k^2(n-1)}{\lambda_d(k,n-1)}+(1-\alpha)P[\gamma_k(n)-1], \quad 0\le\alpha<1 \tag{51}$$

where $\hat{A}_k(n-1)$ is the amplitude estimator of the k-th signal spectral component in the (n-1)-th analysis frame, and $P[\cdot]$ is an operator which is defined by:

$$P[x] = \begin{cases} x & if\ x\ge 0 \\ 0 & otherwise \end{cases} \tag{52}$$

By comparing (50) and (51), we see that $\hat{\xi}_k(n)$ is obtained from (50) by dropping the expectation operator, using the amplitude estimator of the (n-1)-th frame instead of the amplitude itself in the n-th fame, introducing

a weighting factor between the two terms of $\xi_k(n)$, and using the operator $P[\cdot]$ defined in (52). $P[\cdot]$ is used to ensure the positiveness of the proposed estimator in case $\gamma_k(n)-1$ is negative. It is also possible to apply the operator P on the right hand side of (51) rather than on $\gamma_k(n)-1$ only. However, from our experience both alternatives give very similar results.

The proposed estimator for $\xi_k(n)$ is a 'decision-directed' type estimator, since $\hat{\xi}_k(n)$ is updated on the basis of a previous amplitude estimate.

By using $\hat{A}_k(n)=G(\hat{\xi}_k(n),\gamma_k(n))R_k(n)$, where $G(\cdot,\cdot)$ is a gain function which results from either the optimal or the Wiener amplitude estimator, (51) can be written in a way which emphasizes its recursive nature. We get from (51):

$$\hat{\xi}_k(n) = \alpha G^2(\hat{\xi}_k(n-1),\gamma_k(n-1))\gamma_k(n-1)+(1-\alpha)P[\gamma_k(n)-1] \tag{53}$$

Several initial conditions were examined by simulations. We found that using $\xi_k(0)=\alpha+(1-\alpha)P(\gamma_k(0)-1)$ is appropriate, since it minimizes initial transition effects in the enhanced speech.

The theoretical investigation of the recursive estimator (53) is very complicate due to its highly non-linear nature. Even for the simple gain function of the Wiener amplitude estimator it was difficult to analyze. Therefore, we examined it by simulation only, and determined in this way the 'best' value of α.

VI. SYSTEM DESCRIPTION AND PERFORMANCE EVALUATION

In this section we first describe the proposed speech enhancement system, which was implemented on a general purpose computer (Eclipse S-250). Then we describe the performance of this system, based on informal listening, when each of the STSA estimators discussed in this paper is applied.

© Technion - Israel Institute of Technology, Elyachar Central Library

System Description

The input to the proposed system is an 8kHz sampled speech of 0.2-3.2kHz bandwidth, which was degraded by uncorrelated additive noise. Each analysis frame which consists of 256 samples of the degraded speech, and overlaps the previous analysis frame by 192 samples, is spectrally decomposed by means of a discrete short-time Fourier transform (DSTFT) analysis [18,19], using a Hanning window. The STSA of the speech signal is then estimated, and combined with the complex exponential of the noisy phase. The estimated DSTFT samples in each analysis frame are used for synthesizing the enhanced speech signal, by using the well known weighted overlap and add method [19].

In applying the optimal amplitude estimators (7) and (30) in the proposed system, we examine their implementation through exact calculation as well as by using look-up tables. Each look-up table contains a finite number of samples of the corresponding gain function in a prescribe region of (ξ,γ). We found, for example, that when the input SNR is in the range [-5, 5]dB, and the "decision-directed" a-priori SNR is utilized, it suffices to use 961 samples of each gain function, which are obtained by uniformly sampling the range:
$-15 \leq [(\xi,\gamma-1) \text{ or } (\eta,\gamma-1)] \leq 15dB$. As judged by informal listening, this sampling of the gain functions results in a negligible additional residual noise to the enhanced signal. Therefore, the proposed system operating with the optimal amplitude estimator, can be implemented with a similar complexity to that of other commonly used systems, although a more sophisticated amplitude estimator is used here.

The proposed system is examined here for enhancing speech degraded by stationary noise. Therefore the variances of the noise spectral com-

© Technion - Israel Institute of Technology, Elyachar Central Library

ponents are estimated only once, from an initial noise segment having a duration of 320 msec. The estimated variances are used in the estimation of γ_k, and ξ_k by either (47) or (51).

Performance Evaluation

In this section we describe the performance of the above speech enhancement systems, when each of the STSA estimators considered in this paper is applied. Both a-priori SNR estimators (i.e., the ML and the "decision-directed") are examined. The values used here for the parameters α and β in (47) and (51), are the best ones found by simulations. Fig. 7 describes a chart of the comparison tests made here.

In each test, speech signals which were degraded by stationary uncorrelated additive wideband noise with SNR values of 5, 0, and -5dB, were enhanced. The speech material used includes the following sentences, each spoken by a female and a male:

Joe brought a young girl
An icy wind raked the beach
A lathe is a big tool

In addition the sentence "we were away a year ago" spoken by another male is examined. Four listeners participated in the comparison tests. In each test a pair of the enhanced speech signals were presented to the listeners, (through earphones), and they were asked to compare them on the following basis: amount of noise reduction, the nature of the residual noise (e.g., musical vis uniform), and distortion in the speech signal itself.

Let us consider first the tests in which STSA estimators whose derivation is based on the assumption that the speech is surely present in the noisy observaions are used.

© Technion - Israel Institute of Technology, Elyachar Central Library

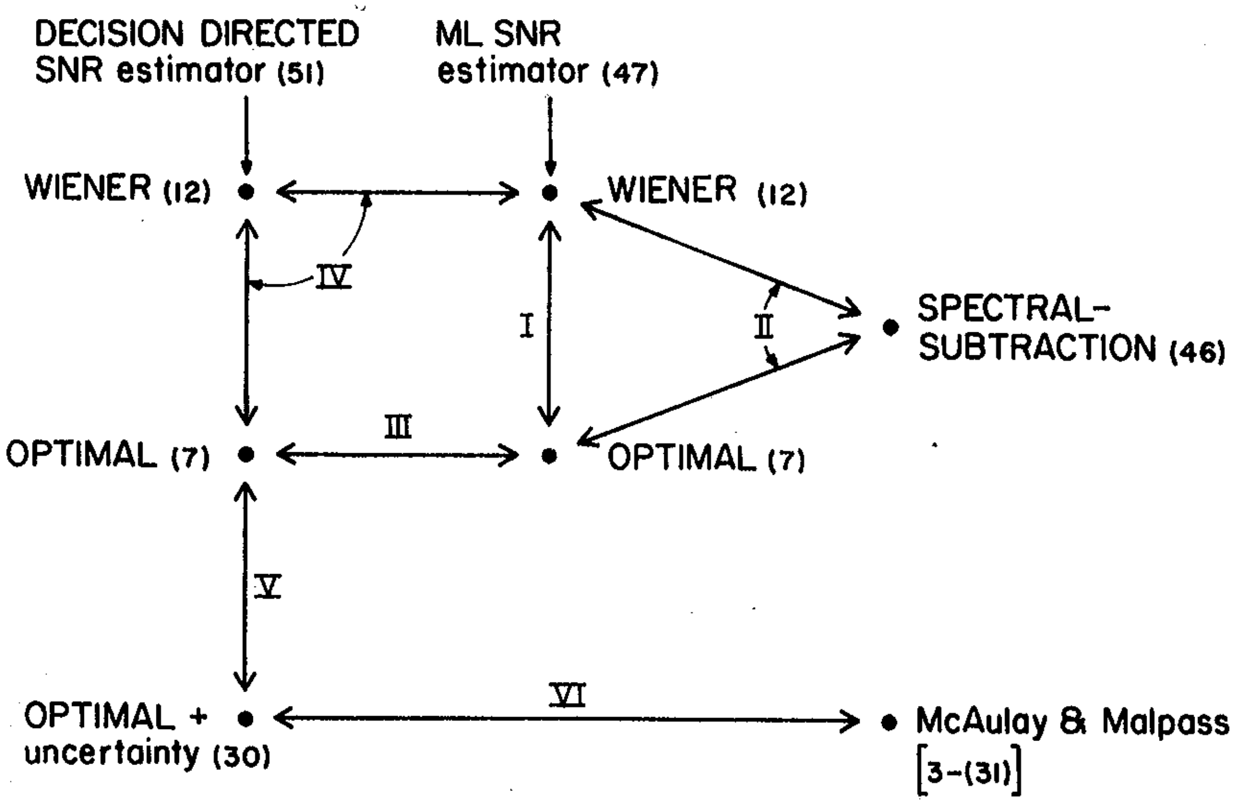


Fig. 7: Comparison listening tests chart.

ציור 7 : תרשים למבחני השמיעה שבוצעו.

© Technion - Israel Institute of Technology, Elyachar Central Library

Case I: Using either the optimal estimator (7) or the Wiener estimator (12), when the a-priori SNR is estimated by the ML estimator (47) with $\alpha=0.725$, $\beta=2$, gives a very similar enhanced speech quality. A significant reduction of the noise is perceived, but a 'musical noise' is introduced. The power of this 'musical noise' is very low at the 5dB SNR value, and it increases as the input SNR decreases. The distortions caused to the speech signal seem to be very small at the high SNR value of 5dB, and increase as the input SNR decreases. Nevertheless, at the SNR value of -5dB, the enhanced speech is still very intelligible.

Case II: The enhanced speech obtained by using the 'spectral subtraction' amplitude estimator (46) with $\beta=2$, suffers from a strong 'musical noise'. This 'musical noise' is of higher power level and wider band, than the 'musical noise' obtained in the above optimal and Wiener estimations (Case I). This is especially prominent at the low input SNR values of 0dB and -5dB. For this reason, the quality of the enhanced speech obtained by using either the optimal or the Wiener estimator, is much better than that obtained by using the 'spectral subtraction' estimator.

Case III: Using the optimal estimator (7), when the a-priori SNR is estimated by the 'decision-directed' estimator (51) with $\alpha=0.98$, results in a great reduction of the noise, and provides enhanced speech with *colorless* residual noise. This colorless residual noise was found to be much less annoying and disturbing, than the 'musical noise' obtained when the a-priori SNR is estimated by the ML estimator (47). As could be judged by informal listening, the distortions in the enhanced speech obtained by using the optimal estimator with either the ML or the 'decision-directed' a-priori SNR estimator, are very similar.

© Technion - Israel Institute of Technology, Elyachar Central Library

CASE IV: Using the Wiener amplitude estimator with the 'decision-directed' a-priori SNR estimator and α=0.98, results in a more distorted speech than that obtained by using the recently described optimal amplitude estimator. However, the residual noise level in the Wiener estimation is lower than that in the optimal estimation. Lowering the value of α, reduces the distortions of the enhanced speech, but introduces a residual 'musical noise' as well. This 'musical noise' is probably contributed by the second term of the 'decision-directed' estimator (i.e., $P[\gamma_k(n)-1]$ in (51)), whose relative weight increases as the value of α decreases. We found that using α=0.97 results in an enhanced speech whose distortion is similar to that obtained by using the optimal estimator. In addition, the level of the residual 'musical noise' obtained then, is lower than that obtained by using the ML a-priori SNR estimator.

CASE V: The optimal amplitude estimator (30), which takes into account the uncertainty of signal presence in the observed signal, results in a better enhanced speech quality than that obtained by using the optimal estimator (7). Specifically, by using (30) with q_k=0.2, and 'the decision-directed' a-priori SNR estimator (51) (when $\hat{\xi}_k(n)$ is replaced by $\hat{\eta}_k(n)$) with α=0.99, we get a further reduction of the residual noise, with negligible additional distortions in the enhanced speech signal.

CASE VI: The enhanced speech obtained by using the above optimal estimator ((30) with q_k=0.2, and (51) with α=0.99), was compared with the enhanced speech obtained by using the McAulay and Malpass amplitude estimator [3] (see Section III). The latter estimator was operated with the 'best' value (as judged by informal listening) of the a-priori SNR parameter, which was found to be 12dB in our experiment. It was found that the main difference between the two enhanced speech signals is in the nature of the

© Technion - Israel Institute of Technology, Elyachar Central Library

residual noise. When the optimal estimator is used the residual noise is colorless, while when the McAulay and Malpass estimator is used, musical residual noise results.

VII. SUMMARY AND DISCUSSION

We present in this paper an algorithm for enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available. The basic approach which was taken here, is to optimally estimate the short-time spectral amplitude (STSA) and complex exponential of the phase, of the speech signal. We use this approach of optimally estimating these two components of the short-time Fourier transform (STFT) separately, rather than optimally estimating the STFT itself, since the STSA of a speech signal rather than its waveform, is of major importance in speech perception. We showed that the STSA and the complex exponential cannot be estimated simultaneously in an optimal way. Therefore, we use an optimal MMSE STSA estimator, and combine it with an optimal MMSE estimator of the complex exponential of the phase which does not affect the STSA estimation. The latter constrained complex exponential estimator is found to be the complex exponential of the noisy phase.

In this paper we derive the optimal STSA estimator and analyze its performance. We showed that the optimal STSA estimator, and the Wiener STSA estimator which results from the optimal MMSE STFT estimator, are nearly equivalent at high SNR. On the other hand, the optimal STSA estimator results in significantly less MSE and bias when the SNR is low. This fact supports our approach to optimally estimate the perceptually important STSA.

An optimal MMSE STSA estimator which takes into account the uncertainty of signal presence in the noisy spectral components is also derived

© Technion - Israel Institute of Technology, Elyachar Central Library

in this paper, and examined in enhancing speech.

The optimal STSA estimator depends on the parameters of the statistical model it is based on. In the proposed algorithm these are the a-priori SNR of each spectral component, and the variance of each noise spectral component. The a-priori SNR was found to be a key parameter of the STSA estimator. It is demonstrated here that by using different estimators for the a-priori SNR, different STSA estimations result. For example, using the 'power spectral subtraction' method for estimating the a-priori SNR, results in an STSA estimator which is nearly equivalent to the 'spectral subtraction' STSA estimator.

We proposed here a 'decision-directed' method for estimating the a-priori SNR. This method was found to be useful when it is applied to either the optimal or the Wiener STSA estimator. By combining this estimator with the optimal STSA estimator which takes into account the uncertainty of signal presence in the noisy observations, we obtained the best speech enhancement results. Specifically, a significant reduction of the input noise is obtained, and the residual noise sounds colorless.

We believe that the full potential of the proposed approach is not yet exploited, although very encouraging results were obtained. Better results may be obtained if the a-priori SNR estimation could be improved. This issue is now being investigated.

Acknowledgement

The authors wish to thank Prof. I. Bar-David, Prof. M. Zakai and Dr. M. Sidi for fruitful and helpful discussions in the course of this work. The authors also wish to thank Mr. S. Shitz for critical reading of the manuscript and for his helpful comments.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

Appendix A

In this Appendix we derive the optimal amplitude estimator (7). On substituting (5) and (6) into (4), and using the integral representation of the modified Bessel function of zero order [10: 8.431.5],

$$I_0(z) = \frac{1}{2\pi}\int_0^{2\pi} \exp(z\cos\beta)d\beta \qquad (A.1)$$

we obtain:

$$\hat{A}_k = \frac{\int_0^\infty a_k^2 \exp(-\frac{a_k^2}{\lambda(k)}) I_0(2a_k\sqrt{\frac{v_k}{\lambda(k)}})da_k}{\int_0^\infty a_k \exp(-\frac{a_k^2}{\lambda(k)}) I_0(2a_k\sqrt{\frac{v_k}{\lambda(k)}})da_k} \qquad (A.2)$$

where v_k is defined by (8), and $\lambda(k)$ satisfying:

$$\frac{1}{\lambda(k)} = \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)} \qquad (A.3)$$

By using [10: 6.631.1, 8.406.3, 9.212.1] we get from (A.2):

$$\hat{A}_k = \lambda_k^{1/2}\Gamma(1.5)M(-0.5;\ 1;\ -v_k) \qquad (A.4)$$

$\hat{A}_k$ as given by (7) is obtained from (A.4) by using (A.3) and (8-10). The equivalent form of $\hat{A}_k$ as given in (7), is obtained by using [4: A.1.31a].

Appendix B VECTOR SPECTRAL SUBTRACTION

This Appendix deals with the estimation of the short-time spectral amplitude (STSA) of the speech signal, by the "Vector Spectral Subtraction" approach. The motivation for deriving this estimator was to extend the popular "spectral subtraction" STSA estimator [1]. We use here the same notation of section II, and consider the estimation of A_k from Y_k, which is given by:

$$Y_k = X_k + D_k \qquad (B.1)$$

In explaining the basic approach, it is useful to consider the "phasor representation" of (B.1), as shown in Fig. B.1. The "vector spectral subtraction" amplitude estimator results from two mutually dependent estimators, of the amplitude A_k and the cosine of the phase error φ_k. Specifically, it is derived in three steps. First, we derive an optimal MMSE estimator $\tilde{A}_k$ of A_k, from the observation (R_k, ϑ_k), assuming that φ_k is known. We get an estimator which depends on R_k and $\cos\varphi_k$. Then, we derive an optimal MMSE estimator $\tilde{\cos}\varphi_k$ of $\cos\varphi_k$, from the observation (R_k, ϑ_k), assuming that A_k is known. We get an estimator which depends on R_k and A_k. The "vector spectral subtraction" amplitude estimator $\hat{A}_k$, is finally obtained from the solution of the two estimation equations for $\tilde{A}_k$ and $\tilde{\cos}\varphi_k$, when each assumed known random variable ($\cos\varphi_k$ or A_k) is replaced by its estimator.

According to the above approach, $\tilde{A}_k$ is given by:

$$\begin{aligned}\tilde{A}_k &= E\{A_k \mid R_k, \vartheta_k, \varphi_k\} \\ &= E\{A_k \mid R_k, \vartheta_k, \alpha_k\} \\ &= \frac{\int_0^\infty a_k p(r_k, \vartheta_k \mid a_k, \alpha_k) p(a_k, \alpha_k) da_k}{\int_0^\infty p(r_k, \vartheta_k \mid a_k, \alpha_k) p(a_k, \alpha_k) da_k}\end{aligned} \tag{B.2}$$

Substituting (5) and (6) into (B.2), gives:

$$\tilde{A}_k = \frac{\xi_k}{1+\xi_k}\left[1+\frac{1}{\nu_k^2}\Lambda(\nu_k)\right] R_k \cos\varphi_k \tag{B.3}$$

where,

$$\nu_k \triangleq \sqrt{2\,\frac{\xi_k}{1+\xi_k}\gamma_k \cos\varphi_k} \tag{B.4}$$

$$\Lambda(\nu_k) \triangleq \frac{\sqrt{2\pi}\nu_k(0.5+erf(\nu_k))\exp(\nu_k^2/2)}{1+\sqrt{2\pi}\nu_k(0.5+erf(\nu_k))\exp(\nu_k^2/2)} \tag{B.5}$$

$$erf(\nu_k) \triangleq \frac{1}{\sqrt{2\pi}}\int_0^{\nu_k} \exp(-t^2/2)dt \tag{B.6}$$

$\tilde{A}_k$ depends on the cosine of the phase error which will be estimated next,

© Technion - Israel Institute of Technology, Elyachar Central Library

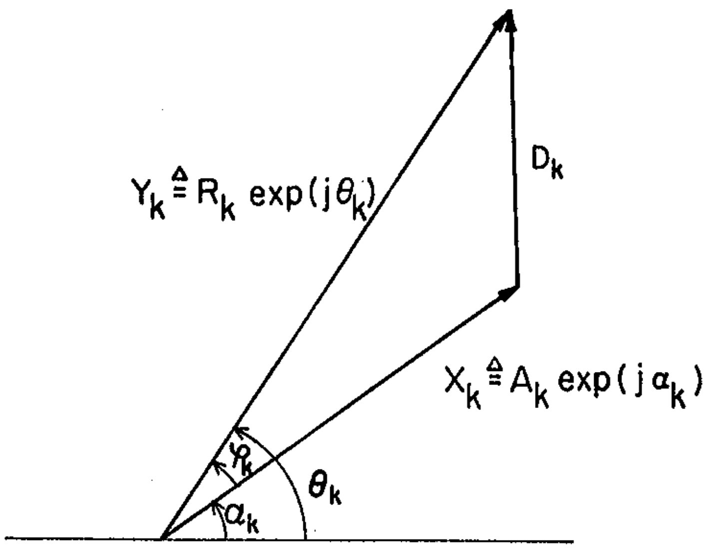


Fig. B.1: Phasor representation of (B.1).

<u>ציור B.1</u> : יצוג פאזורי של (B.1)

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

and on ξ_k and γ_k which are defined in (9) and (10) respectively. $\Lambda(\nu_k)$ is a monotonically decreasing function, approaching zero as the SNR value increases (i.e., as $\nu_k \to \infty$).

Assuming that the observed phase ϑ_k is given, φ_k is a function of α_k only. Therefore, the optimal MMSE estimator of $\cos\varphi_k$, given (R_k, ϑ_k), and assuming that A_k is known, is:

$$\tilde{\cos}\varphi_k = E\{\cos\varphi_k \mid R_k, \vartheta_k, A_k\} \tag{B.7}$$

$$= \frac{\int_0^{2\pi} \cos\varphi_k \, p(R_k, \vartheta_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k}{\int_0^{2\pi} p(R_k, \vartheta_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k}$$

Substituting (5) and (6) into (B.7), and performing the integration, gives:

$$\tilde{\cos}\varphi_k = \frac{I_1(\rho_k)}{I_0(\rho_k)} \tag{B.8}$$

where ρ_k is defined by:

$$\rho_k \triangleq 2\gamma_k \frac{A_k}{R_k} \tag{B.9}$$

The desired estimator $\hat{A}_k$ of A_k, results from the solution of the two non-linear equations (B.3) and (B.8), when $\cos\varphi_k$ in (B.3) and A_k in (B.8), are replaced by $\tilde{\cos}\varphi_k$ and $\tilde{A}_k$, respectively. Except for high SNR values, it is difficult to obtain or to prove the existence and uniqueness of a closed form mathematical solution. This problem is still open. However, a numerical approach resulted in a single solution for the gain function $G(\xi_k, \gamma_k) \triangleq \hat{A}_k / R_k$, and its gain curves were found to coincide with those of the optimal gain function $G_{opt}(\xi_k, \gamma_k)$. We conjecture that this coinciding is a consequence of the statistical independence assumption of the real and imaginary parts of each Fourier expansion coefficient, which results in the statistical independence of the amplitude and the phase. This probably enables to obtain the optimal amplitude estimator, by cross-coupling the

© Technion - Israel Institute of Technology, Elyachar Central Library

two partial optimal estimators, of the amplitude and the cosine of the phase error.

The interpretation of the proposed amplitude estimator as a "vector spectral subtraction" amplitude estimator, follows from the fact that A_k in the "triangle" shown in Fig. B.1, is eventually estimated from the observation R_k, the variance of the noise D_k, and by utilizing an estimate of $\cos\varphi_k$ (see (B.3)). This is in contrast with the "spectral subtraction" amplitude estimator (i.e. $\hat{A}_k=\sqrt{R_k^2-\lambda_d(k)}$), which is obtained from the observation R_k and the variance of the noise D_k, by applying the Pythagorean relation to the triangle in figure B.1, which is considered to be a right triangle. Since the optimal gain curves coincide with those of the "vector spectral subtraction" amplitude estimator, the above interpretation holds for the optimal amplitude estimator (7) as well.

We turn now to examine the asymptotic behavior of the "vector spectral subtraction" amplitude estimator at high SNR. Substituting $\Lambda(\nu_k)=0$ in (B.3) and paralleling the procedure described above for obtaining $\hat{A}_k$, we arrive at the following single equation:

$$\hat{A}_k \cong \frac{\xi_k}{1+\xi_k}\,\frac{I_1(2\gamma_k\hat{A}_k/R_k)}{I_o(2\gamma_k\hat{A}_k/R_k)}R_k \qquad \textit{high SNR} \qquad \text{(B.10)}$$

The solution of (B.10) for $\xi_k\to\infty$, gives the exact ML estimator of A_k [11,20]. The corresponding gain curve which results from a numerical solution, is shown in Fig. B.2. This figure demonstrates also the close relation between the ML and "spectral subtraction" amplitude estimators.

By letting $\xi_k\to\infty$ in (B.10), using the relation $I_1(\rho_k)=\partial I_o(\rho_k)/\partial\rho_k$, and the following asymptotic approximation for $I_o(\rho_k)$ [21],

$$I_o(\rho_k) \cong \frac{\exp(\rho_k)}{\sqrt{2\pi\rho_k}}, \qquad \rho_k>>1 \qquad \text{(B.11)}$$

we get the McAulay and Malpass [3] approximation for the ML amplitude

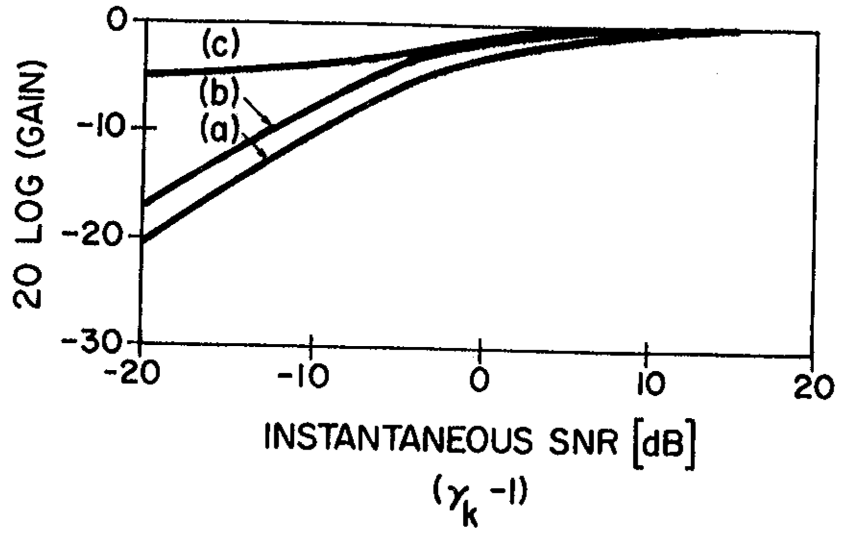


Fig. B.2: Gain curves describing:
(a) "Spectral Subtraction" gain function (Eq. (46) with β=1).
(b) ML amplitude estimator (Eq. (B.10) with $\xi\to\infty$).
(c) Approximated ML amplitude estimator (Eq. (B.12)).

ציור B.2 : עקומי הגבר המתארים:
(a) - פונקציית ההגבר בשיטת ההחסרה הספקטרלית (46) עם $\beta = 1$.
(b) - משערך הסבירות המירבית (B.10) עם $\xi_k \to \infty$.
(c) - קרוב למשערך הסבירות המירבית (B.12).

© Technion - Israel Institute of Technology, Elyachar Central Library

estimator:

$$\hat{A}_k \cong \frac{1}{2}[1+\sqrt{1-\frac{1}{\gamma_k}}]R_k \qquad \textit{high SNR} \tag{B.12}$$

The gain function which results from this estimator is also shown in Fig. B.2.

Note that although the gain curves of the optimal and the "vector spectral subtraction" amplitude estimators coincide, we get mathematically that they converge to different estimators (i.e., Wiener and ML respectively) at high SNR. This is of course a consequence of the different approximation made in each case, in order to get the asymptotic behavior. However, as is well known, both estimators result in similar performance at high SNR.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

Appendix C

In this Appendix we derive the optimal MMSE STSA estimators, under the two statistical models of signal absence in the noisy observations. This derivation is based on [7].

We begin with the first model in which signal is assumed to be either present with probability $1-q$, or absent with probability q, in all of the noisy spectral components. We base the estimation on a finite number of spectral components $\underline{Y} \triangleq (Y_0, Y_1, \ldots, Y_K)$. K is chosen to satisfy $K \geq BT+1$, where B denotes the signal bandwidth, and T is the analysis interval length.

The risk to be minimized is given by:

$$J = E\{C(\hat{A}_k, A_k, H)\} \qquad \text{(C.1)}$$

where $C(\cdot)$ is the cost function, and H is a binary random variable representing the two hypotheses. The PDF of H is given by:

$$p(h) = q\,\delta(h-H^0)+(1-q)\delta(h-H^1) \qquad \text{(C.2)}$$

where $\delta(\cdot)$ is a Dirac function, and H^0 and H^1 denote the hypotheses of signal absence and presence respectively. When a quadratic cost function is used, $C(\cdot)$ is given by:

$$C(\hat{A}_k, A_k, H) = \begin{cases} (A_k - \hat{A}_k)^2 & H = H^1 \\ \hat{A}_k^2 & H = H^0 \end{cases} \qquad \text{(C.3)}$$

By using (C.2), we can write (C.1) as:

$$\begin{aligned} J &= \int\int\int p(\underline{Y}\,|\,\underline{a}, h)p(\underline{a}\,|\,h)p(h)C(\hat{a}_k, a_k, h)\,dh\,d\underline{a}\,d\underline{Y} \qquad \text{(C.4)} \\ &= q\int\int p(\underline{Y}\,|\,\underline{a}, H^0)p(\underline{a}\,|\,H^0)C(\hat{a}_k, a_k, H^0)\,d\underline{a}\,d\underline{Y} \\ &+ (1-q)\int\int p(\underline{Y}\,|\,\underline{a}, H^1)p(\underline{a}\,|\,H^1)C(\hat{a}_k, a_k, H^1)\,d\underline{a}\,d\underline{Y} \end{aligned}$$

where $\underline{a} \triangleq (a_0, a_1, \ldots, a_K)$ is a realization of the vector $(A_0, A_1, \ldots, A_k)$. By using the fact that $p(\underline{a}\,|\,H^0) = \prod_{l=1}^{K} \delta(a_l)$, and (C.3), we obtain from (C.4):

$$J = \int\{qp(\underline{Y}\,|\,H^0)\hat{a}_k^2 + (1-q)\int p(\underline{Y}\,|\,\underline{a}, H^1)p(\underline{a}\,|\,H^1)(a_k-\hat{a}_k)^2 d\underline{a}\}\,d\underline{Y} \qquad \text{(C.5)}$$

The minimization of (C.5) with respect to $\hat{a}_k$ is done in the usual way, and

© Technion - Israel Institute of Technology, Elyachar Central Library

results in:

$$\hat{A}_k = \frac{\Lambda(\underline{Y},q)}{1+\Lambda(\underline{Y},q)} E\{A_k \mid \underline{Y}, H^1\} \tag{C.6}$$

where $\Lambda(\underline{Y},q)$ is the generalized likelihood ratio which is defined by:

$$\Lambda(\underline{Y},q) = \frac{1-q}{q} \frac{p(\underline{Y} \mid H^1)}{p(\underline{Y} \mid H^o)} \tag{C.7}$$

$E\{A_k \mid \underline{Y}, H^1\}$ is the optimal MMSE amplitude estimator when the signal is surely present in the noisy observation.

Based on the statistical model used in this paper, in which the spectral components of each process are assumed to be statistically independent, $\Lambda(\underline{Y},q)$ equals to:

$$\Lambda(\underline{Y},q) = \frac{1-q_k}{q_k} \prod_{l=0}^{K} \frac{p(y_l \mid H_l^1)}{p(y_l \mid H_l^o)} \tag{C.8}$$

and $E\{A_k \mid \underline{Y}, H^1\} = E\{A_k \mid Y_k, H_k^1\}$. In (C.8) H_k^o and H_k^1 denote the two hypotheses of signal absence and presence respectively in the k-th spectral component, and $q_k = q$ is the probability of signal absence in the k-th spectral component.

On the basis of the Gaussian statistical model assumed here, it is easy to see that:

$$\frac{p(Y_k \mid H_k^1)}{p(Y_k \mid H_k^o)} = \frac{\exp(v_k)}{1+\xi_k} \tag{C.9}$$

as explained in Section III. Combining (C.6), (C.8), and (C.9), we obtain:

$$\hat{A}_k = \frac{\mu_k \prod_{l=0}^{K} \frac{\exp(v_l)}{1+\xi_l}}{1+\mu_k \prod_{l=0}^{K} \frac{\exp(v_l)}{1+\xi_l}} E\{A_k \mid Y_k, H_k^1\} \tag{C.10}$$

$E\{A_k \mid Y_k, H_k^1\}$ is the estimator derived in (7).

The amplitude estimator (C.10) operating with the "decision-directed" a-priori SNR estimator (51), was not found to be useful while applied in the speech enhancement system of Section VI. Specifically, we found that the additional gain function $\Lambda/(1+\Lambda)$ which appears in (C.10) affects mainly the non-

© Technion - Israel Institute of Technology, Elyachar Central Library

speech intervals. As a consequence, a switching effect of the enhanced speech, and an emphasis of the residual noise in the speech intervals, is perceived. Fig. C.1 demonstrates the above effect. It shows the gain function $\Lambda/(1+\Lambda)$ (for $q=0.5$) which is superimposed on the normalized energy contour of a noisy speech signal. Each point in this figure corresponds to a specific analysis frame.

Fig. C.1 suggests however a potential application of the gain function $\Lambda/(1+\Lambda)$ (or equivalently of the likelihood ratio Λ). As can be seen, it has a good potential in post-detecting non-speech intervals in the observed signal. This advantage can be utilized, for example, when data is to be transmitted during non-speech intervals of a noisy speech signal.

We turn now to the derivation of the optimal MMSE STSA estimator (25), which is based on the second model of signal absence. According to this model, statistically independent random appearance of the signal in the noisy spectral components is assumed. The risk to be minimized is now given by:

$$J = E\{C(\hat{A}_k, A_k, H_k)\} \tag{C.11}$$
$$= \int\int\int\int p(\underline{Y}\,|\,\underline{a}, \underline{\alpha}, \underline{h})\,p(\underline{a}, \underline{\alpha}\,|\,\underline{h})\,p(\underline{h})\,C(\hat{a}_k, a_k, h_k)\,d\underline{h}\,d\underline{a}\,d\underline{\alpha}\,d\underline{Y}$$

where $\underline{H} \overset{\Delta}{=} (H_0, H_1, \ldots, H_k)$ is a vector of binary random variables, representing the two hypotheses of signal absence and presence in each of the noisy spectral components. The vector $\underline{h}$ which appears in (C.11) is a realization of $\underline{H}$. The vector $\underline{\alpha}$ is defined similarly to the vector $\underline{a}$. Now by using the statistical independence assumption of the spectral components of each process (i.e., speech and noise), and the above model of signal absence, we can write:

$$p(\underline{Y}\,|\,\underline{a}, \underline{\alpha}, \underline{h}) = \prod_{l=1}^{k} p(Y_l\,|\,a_l, \alpha_l, h_l) \tag{C.12}$$

$$p(\underline{a}, \underline{\alpha}\,|\,\underline{h}) = \prod_{l=1}^{k} p(a_l, \alpha_l\,|\,h_l)$$

$$p(\underline{h}) = \prod_{l=1}^{k} p(h_l)$$

On substituting (C.12) into (C.11) we get:

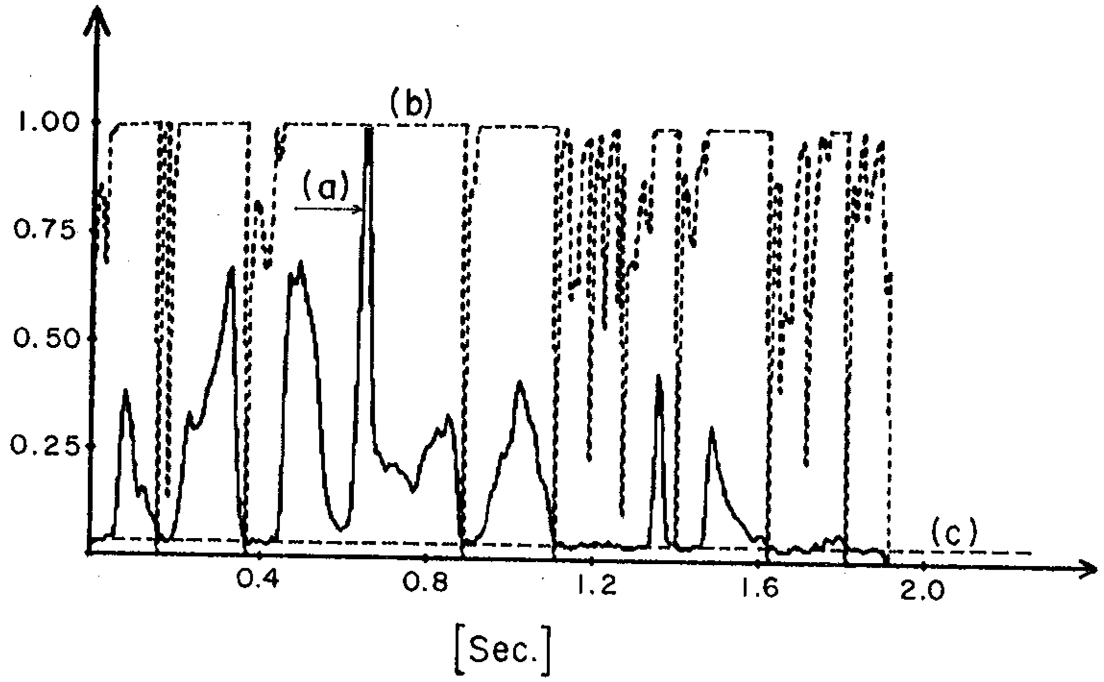


Fig. C.1: Switching effect demonstration:
(a) Normalized energy contour of noisy speech (5dB) in analyzed frames.
(b) Gain function Λ/ (1+Λ) in analyzed frames (see (C.10)) for q = 0.5.
(c) White noise spectral density.

ציור C.1 : הדגמת אפקט המיתוג:
(a) - עקום האנרגיה המנורמלת במסגרות האנליזה השונות של אות הדבור הרועש.
(b) - פונקצית ההגבר Λ/(1+Λ) במסגרות האנליזה השונות (ראה (C.10)) עבור q = 0.5.
(c) - הצפיפות הספקטרלית של הרעש הלבן.

© Technion - Israel Institute of Technology, Elyachar Central Library

$$J=\int\int\int\int p(Y_k \mid a_k, \alpha_k, h_k)\, p(a_k, \alpha_k \mid h_k)\, p(h_k)\, C(\hat{a}_k, a_k, h_k)\, dh_k\, da_k\, d\alpha_k\, dY_k \quad \text{(C.13)}$$

By using a similar cost function to (C.3), and a similar PDF for H_k to (C.2), we easily obtain the estimator (25) from (C.13). Note that due to the statistical independence assumptions specified by (C.12), the resulting estimator for A_k $(k \leq K)$ does not depend on K. Therefore, in this case the estimator can be considered as if it is obtained from an infinite number $(K \to \infty)$ of spectral components.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

Appendix D

In this Appendix we derive the optimal MMSE complex exponential estimators (34) and (37), under the assumed Gaussian statistical model.

To derive (34), we have to calculate $E\{\cos\varphi_k \mid Y_k\}$ and $E\{\sin\varphi_k \mid Y_k\}$ only, which appear in (32).

$$E\{\cos\varphi_k \mid Y_k\} = \int_0^{2\pi} \cos(\vartheta_k - \alpha_k) p(\alpha_k \mid Y_k) d\alpha_k \tag{D.1}$$

$$= \frac{\int_0^\infty \int_0^{2\pi} \cos(\vartheta_k - \alpha_k) p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k}$$

On substituting (5) and (6) into (D.1), and using the integral representation of the modified Bessel function of n-th order [10: 8.431.5],

$$I_n(z) = \frac{1}{2\pi} \int_0^{2\pi} \cos\beta n \exp(z\cos\beta) d\beta \tag{D.2}$$

we obtain:

$$E\{\cos\varphi_k \mid Y_k\} = \frac{\int_0^\infty a_k \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_1\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k}{\int_0^\infty a_k \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_0\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k} \tag{D.3}$$

where v_k is defined by (8), and $\lambda(k)$ satisfies:

$$\frac{1}{\lambda(k)} = \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)} \tag{D.4}$$

By using [10: 6.631.1, 8.406.3, 9.212.1], we get from (D.3):

$$E\{\cos\varphi_k \mid Y_k\} = \Gamma(1.5)\sqrt{v_k}\, M(0.5;2;-v_k) \tag{D.5}$$

The equivalent form of $E\{\cos\varphi_k \mid Y_k\}$ as given in (34), is obtained by using [4: A.1.31d].

To show that $E\{\sin\varphi_k \mid Y_k\}=0$ we substitute (5) and (6) into:

$$E\{\sin\varphi_k \mid Y_k\} = \frac{1}{p(Y_k)} \int_0^\infty \int_0^{2\pi} \sin(\vartheta_k - \alpha_k) p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k \tag{D.6}$$

© Technion - Israel Institute of Technology, Elyachar Central Library

We obtain:

$$E\{\sin\varphi_k \mid Y_k\} \sim \int_0^\infty a_k \exp(-\frac{a_k^2}{\lambda(k)})\int_0^{2\pi} \sin(\vartheta_k - \alpha_k)\exp(\frac{2a_k R_k}{\lambda_d(k)}\cos(\vartheta_k - \alpha_k))d\,\alpha_k\,da_k \tag{D.7}$$

where ~ denotes proportionality. Now it is easy to see that the inner integral in (D.7) equals zero.

To derive the estimator (37), we solve the following problem:

$$\min_{g(Y_k)} E\{|e^{j\alpha_k} - g(Y_k)|^2\} \tag{D.8}$$

subject to: $|g(Y_k)| = 1$

where $g(Y_k)$ is the constrained complex exponential estimator. Equivalently, we have to solve the following problem:

$$\min_{g(Y_k)} E\{|e^{j\alpha_k} - g(Y_k)|^2 \mid Y_k\} \tag{D.9}$$

subject to: $|g(Y_k)| = 1$

By using the Lagrange multiplier method, our problem turns out to be:

$$\min_{g(Y_k),\rho_k} E\{|e^{j\alpha_k} - g(Y_k)|^2 \mid Y_k\} + \rho_k(|g(Y_k)| - 1) \tag{D.10}$$

By denoting $g(Y_k) \triangleq g_R(Y_k) + jg_I(Y_k)$, and equating the derivatives of (D.10) with respect to $g_R(\cdot), g_I(\cdot)$, and ρ_k to zero, we obtain:

$$\frac{\partial(\cdot)}{\partial g_R(Y_k)} = 0 \Rightarrow g_R(Y_k)(2+\rho_k) = 2E\{\cos\alpha_k \mid Y_k\} \tag{D.11}$$

$$\frac{\partial(\cdot)}{\partial g_I(Y_k)} = 0 \Rightarrow g_I(Y_k)(2+\rho_k) = 2E\{\sin\alpha_k \mid Y_k\} \tag{D.12}$$

$$\frac{\partial(\cdot)}{\partial \rho_k} = 0 \Rightarrow (g_R^2(Y_k) + g_I^2(Y_k))^{1/2} = 1 \tag{D.13}$$

The solution of the above three equations (D.11-D.13) gives:

$$g(Y_k) = \frac{E\{e^{j\alpha_k} \mid Y_k\}}{|E\{e^{j\alpha_k} \mid Y_k\}|} \tag{D.14}$$

Now by using (40) and (41) we obtain from (D.14) the estimator (37).

We note that since (D.14) is the solution of (D.8) without referring to a specific statistical model, we can draw sufficient conditions for which $g(Y_k)=\exp(j\vartheta_k)$. It can be easily shown that if $E\{\sin\varphi_k \mid y_k\}=0$ then $g(Y_k)=\exp(j\vartheta_k)$. This can happen for example if the noise is Gaussian (as assumed here), and α_k is uniformly distributed on $[0, 2\pi]$ and statistically independent of a_k. In this case it is easy to see that $E\{\sin\varphi_k \mid Y_k\}$ is proportional to:

$$E\{\sin\varphi_k \mid Y_k\} \sim \int_0^{\infty} p(a_k)\exp(-\frac{a_k^2}{\lambda_d(k)})\int_0^{2\pi}\sin(\vartheta_k-\alpha_k)\exp(\frac{2a_k R_k}{\lambda_d(k)}\cos(\vartheta_k-\alpha_k))d\,\alpha_k\,da_k \quad \text{(D.15)}$$

and the inner integral in (D.15) equals zero.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

References

[1] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", Proc. IEEE, Vol. 67, pp. 1586-1604, Dec. 1979.

[2] J.L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd ed., New York, Springer-Verlag, 1972, p.210.

[3] R.J. McAulay and M.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 137-145, Apr. 1980.

[4] D. Middelton, Introduction to Statistical Communication Theory, McGraw-Hill, N.Y. 1960. Chap.7, Appendix 1.

[5] S. Bernstein, Sur l'extension du theoreme limite du calcul des probabilites aux sommes des quantites dependantes, Math. Ann., 97:1, 1926.

[6] W.B. Davenport and W.L. Root, "An Introduction to the Theory of Random Signals and Noise, McGraw-Hill, New York., Chap.6.

[7] D. Middleton and R. Esposito, "Simultaneous Optimum Detection and Estimation of Signals in Noise", IEEE Trans. Inform. Theory, Vol. IT-14, No. 3, pp. 434-444, May 1968.

[8] Y. Ephraim and D. Malah, "Speech Enhancement Under Uncertainty of Signal Presence in the Observed Signal", EE Publication No. 543, Dept. of Electrical Engineering, Technion, Haifa, Israel, July 1983.

[9] T.T. Kadota, "Optimal Reception of Binary Gaussian Signals", Bell Sys. Tech. J., Vol. 43, pp. 2767-2810, Nov. 1964.

[10] I.S. Gradshteyn and I.M. Ryzhik, Table of Integrals, Series, and Products, Academic Press, Inc., 1980.

[11] D. Middelton, "The Incoherent Estimation of Signal Amplitude in Normal Noise Backgrounds", in M. Rosenblatt (ed.): "Time Series Analysis", chap. 24, John Wiley & Sons, Inc., New York, 1963.

[12] Y. Ephraim and D. Malah, "Speech Enhancement Using Optimal Non-Linear Spectral Amplitude Estimation", in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 1118-1121, Apr. 1983.

[13] Y. Ephraim and D. Malah, "Speech Enhancement Using Vector Spectral Subtraction Amplitude Estimation", in Proc. IEEE 13th Convention of Elec. Electron. Eng. in Israel, Tel-Aviv, Mar. 1983.

[14] M. Loeve, Probability Theory, 3rd ed., Van Nostrand, Princeton, N.J., 1963.

[15] A.S. Wilsky, "Fourier Series and Estimation on the Circle with Applications to Synchronous Communication - Part I: Analysis", IEEE Trans. Inform. Theory, Vol. IT-20, No. 5, pp. 577-583, Sept. 1974.

[16] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-27, pp. 113-120, Apr. 1979.

[17] D.B. Paul, "The Spectral Envelope Estimation Vocoder", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-29, pp. 786-794, Aug. 1981.

[18] M.R. Portnoff, "Time Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis", IEEE Trans., Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 55-69, Feb. 1980.

[19] R.E. Crochiere, "A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 99-102, Feb. 1980.

[20] T.W. Eddy, "Maximum Likelihood Detection and Estimation for Harmonic Sets", J. Acoust. Soc. Amer., Vol. 68, No. 1, pp. 149-155, July 1980.

[21] H.L. Van Trees, Detection, Estimation and Modulation Theory, Part I., John Wiley & Sons, Inc., New York, 1968, p. 339.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

נספח ב' - הדגשת דבור תוך שימוש במשערך השגיאה הריבועית הממוצעת המינימלית עבור לוגריתם האמפליטודה הספקטרלית

SPEECH ENHANCEMENT USING A MINIMUM MEAN SQUARE ERROR LOG-SPECTRAL AMPLITUDE ESTIMATOR

ABSTRACT

In this paper we derive a short-time spectral amplitude (STSA) estimator which minimizes the mean square error of the log-spectra (i.e., the original STSA and its estimator), and examine it in enhancing speech. This estimator is also compared with the corresponding minimum mean square error (MMSE) STSA estimator derived previously.

It is found that the new estimator is very effective in enhancing speech, and it results in a better enhanced speech quality than that obtained by using the MMSE STSA estimator.

© Technion - Israel Institute of Technology, Elyachar Central Library

I. Introduction

Recently [1,2] we proposed an algorithm for enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available. This algorithm capitalizes on the major importance of the short-time spectral amplitude (STSA) in speech perception, and utilizes an optimal minimum mean square error (MMSE) STSA estimator for enhancing the speech signal.

While the distortion measure of mean square error of the spectra (i.e., the original STSA and its estimator) used in [1,2] is mathematically tractable, and leads also to good results, it is not the most subjectively meaningful one. For example, it is well known that the mean square error of the log-spectra is a more suitable distortion measure for speech processing, and it is extensively used for speech analysis and recognition [3]. For this reason, it is of great interest to examine the optimal STSA estimator, which minimizes the mean square error of the log-spectra, in enhancing speech signals. The derivation of the above optimal STSA estimator, and its comparison with the MMSE STSA derived in [1,2], are the subjects of this paper.

We use here the same formulation of the estimation problem, and the same statistical model, as in [2]. Specifically, the estimation problem of the STSA is formulated as that of estimating the amplitude of each Fourier expansion coefficient of the speech signal $\{x(t), 0 \le t \le T\}$, given the noisy process $\{y(t), 0 \le t \le T\}$. The Fourier expansion coefficients of the speech process, as well as of the noise process, are modeled as statistically independent Gaussian random variables. The validity of this model for the discussed problem is given in detail in [2].

We also extend the above optimal estimator, and derive it under uncertainty of signal presence in the noisy observations. The resulting estimator is compared with the corresponding one from [2].

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

The paper is organized as follows: In Section II we derive the optimal STSA estimator. In Section III we derive the optimal STSA estimator under uncertainty of signal presence in the noisy observations. In Section IV we first briefly describe the application of the optimal STSA estimator in the speech enhancement system used in [2]. Then we compare its performance by informal listening with that obtained by using the MMSE STSA from [2]. In Section IV we summarize the paper, and draw conclusions.

II. Derivation of Optimal Estimator

Let $X_k = A_k e^{j\alpha_k}$, D_k, and $Y_k = R_k e^{j\vartheta_k}$, denote the k-th Fourier expansion coefficient of the speech signal, the noise process, and the noisy observations, respectively, in the analysis interval [0,T]. According to the formulation of the estimation problem given in the previous section, we are looking for the optimal estimator $\hat{A}_k$ which minimizes the following measure:

$$E(\log A_k - \log \hat{A}_k)^2 \tag{1}$$

given the noisy observation $\{y(t),\ 0 \le t \le T\}$. The estimator is easily shown to be:

$$\hat{A}_k = \exp\{E[\ln A_k \,|\, y(t),\ 0 \le t \le T]\} \tag{2}$$

and it is independent of the basis chosen for the log in (1). As noted in [2], under the assumed statistical model, the expected value of A_k given $\{y(t),\ 0 \le t \le T\}$ equals to the expected value of A_k given Y_k only. Since this statement remains true when A_k is replaced by $\ln A_k$, the optimal estimator (2) equals to:

$$\hat{A}_k = \exp\{E[\ln A_k \,|\, Y_k]\} \tag{3}$$

The evaluation of $E[\ln A_k \,|\, Y_k]$ for the Gaussian model assumed here, is conveniently done by utilizing the moment generating function of $\ln A_k$ given Y_k. Let $Z_k = \ln A_k$. Then the moment generating function $\Phi_{Z_k|Y_k}(\mu)$ of Z_k given Y_k equals to:

© Technion - Israel Institute of Technology, Elyachar Central Library

$$\Phi_{Z_k|Y_k}(\mu) = E\{\exp(\mu Z_k) | Y_k\} \qquad (4)$$

$$= E\{A_k^{\mu} | Y_k\}$$

$E\{\ln A_k | Y_k\}$ is obtained from $\Phi_{Z_k|Y_k}(\mu)$ by

$$E\{\ln A_k | Y_k\} = \frac{d}{d\mu} \Phi_{Z_k|Y_k}(\mu) \Big|_{\mu=0} \qquad (5)$$

Therefore, our task is now to calculate $\Phi_{Z_k|Y_k}(\mu)$ and then to obtain $E\{\ln A_k | Y_k\}$ by using (5). $\Phi_{Z_k|Y_k}(\mu)$ is given by:

$$\Phi_{Z_k|Y_k}(\mu) = E\{A_k^{\mu} | Y_k\} \qquad (6)$$

$$= \frac{\int_0^{\infty}\int_0^{2\pi} a_k^{\mu} p(Y_k | a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k}{\int_0^{\infty}\int_0^{2\pi} p(Y_k | a_k, \mu_k) p(a_k, \alpha_k) d\alpha_k da_k}$$

On the basis of the Gaussian model assumed here, $p(Y_k | a_k, \alpha_k)$ and $p(a_k, \alpha_k)$ are given by [2]:

$$p(Y_k | a_k, \alpha_k) = \frac{1}{\pi\lambda_d(k)} \exp\{-\frac{1}{\lambda_d(k)} | Y_k - a_k e^{j\alpha_k} |^2\} \qquad (7)$$

$$p(a_k, \alpha_k) = \frac{a_k}{\pi\lambda_x(k)} \exp\{-\frac{a_k^2}{\lambda_x(k)}\} \qquad (8)$$

where $\lambda_d(k) \triangleq E\{|D_k|^2\}$, and $\lambda_x(k) \triangleq E\{|X_k|^2\}$, are the variances of the noise and the signal k-th spectral component. On substituting (7) and (8) into (6), using the integral representation of the modified Bessel function of zero order $I_o(\cdot)$ [4: 8.406.3, 8.411.1], and using [4: 6.631.1, 8.406.3, 9.212.1], we obtain:

$$\Phi_{Z_k|Y_k}(\mu) = \lambda_k^{\mu/2}\Gamma(\frac{\mu}{2}+1) M(-\frac{\mu}{2}; 1; -v_k) \qquad (9)$$

where $\Gamma(\cdot)$ is the Gamma function; $M(a;c;x)$ is the confluent hypergeometric function [4: 9.210.1]; λ_k satisfying the following equation:

$$\frac{1}{\lambda_k} = \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)} \qquad (10)$$

and v_k is defined by:

© Technion - Israel Institute of Technology, Elyachar Central Library

$$v_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k \tag{11}$$

where ξ_k and γ_k are defined by:

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \tag{12}$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)} \tag{13}$$

ξ_k and γ_k are interpretated as the a-priori and a-posteriori signal to noise ratio (SNR) [2]. Note that $\Phi_{Z_k|Y_k}(\mu)$ is the formula of the μ-th moment of a Rician random variable.

The derivative of $\Phi_{Z_k|Y_k}(\mu)$ with respect to μ is obtained as follows: First we note that $M(a;c;x)$ is defined by [4: 9.210.1]:

$$M(a;c;x) = \sum_{r=0}^{\infty} \frac{(a)_r}{(c)_r} \frac{x^r}{r!} \tag{14}$$

where $(a)_r \triangleq 1.a.(a+1),\ldots,(a+r-1)$, and $(a)_0 \triangleq 1$. $M(-\mu/2;1;-v_k)$ which appears in (9) can be differentiated term by term for $|\mu|<2$, since the series of the derivatives converges uniformly on that interval. The derivative of $M(-\mu/2;1;-v_k)$ at $\mu=0$ is obtained by using (14) and it equals to:

$$\frac{\partial}{\partial\mu} M(-\mu/2\,;\,1\,;\,-v_k)\,\Big|_{\mu=0} = -\frac{1}{2}\sum_{r=1}^{\infty} \frac{(-v)^r}{r!}\,\frac{1}{r} \tag{15}$$

The derivative of $\Gamma(\mu/2+1)$ is conveniently calculated by using the following series expansion of $\ln\Gamma(\mu/2+1)$ [4: 8.342.1] for $|\mu|<2$:

$$\ln\Gamma(\mu/2+1) = -c\frac{\mu}{2}+\sum_{r=2}^{\infty}\frac{(-\mu)^r}{2^r r}\alpha_r \tag{16}$$

where

$$\alpha_r \triangleq \sum_{n=1}^{\infty}\frac{1}{n^r} \tag{17}$$

and $c=0.577\,215\,664\,90$ is the Euler constant. Differentiation of (16) term by term, gives:

$$\frac{d}{d\mu}\Gamma(\frac{\mu}{2}+1)\,|_{\mu=0} = \Gamma(\frac{\mu}{2}+1)\frac{d}{d\mu}\ln\Gamma(\frac{\mu}{2}+1)\,|_{\mu=0} = -\frac{c}{2} \quad (18)$$

Now, by using (15), and (18), we obtain from (9):

$$\frac{d}{d\mu}\Phi_{Z_k|Y_k}(\mu)\,|_{\mu=0} = \frac{1}{2}\ln\lambda_k - \frac{1}{2}(c + \sum_{r=1}^{\infty}\frac{(-v_k)^r}{r!}\frac{1}{r}) = \frac{1}{2}\ln\lambda_k + \frac{1}{2}(\ln v_k + \int_{v_k}^{\infty}\frac{e^{-t}}{t}dt) \quad (19)$$

where the last equation is obtained from [4: 8.211.1, 8.214.1]. The integral in (19) is known as the exponential integral of v_k, and can be efficiently calculated [5]. On substituting (19) into (5) and using (10)-(12), we get from (3) the optimal amplitude estimator:

$$\hat{A}_k = \frac{\xi_k}{1+\xi_k}\exp\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\}R_k \quad (20)$$

It is useful to consider $\hat{A}_k$ as being obtained from R_k, by a multiplicative non-linear gain function which depends only on the a-priori and the a-posteriori SNR ξ_k and γ_n respectively. This gain function is defined by:

$$G(\xi_k,\gamma_k) \triangleq \frac{\hat{A}_k}{R_k} \quad (21)$$

and it is described by parametric gain curves in Fig. 1. This figure shows also the corresponding gain curves which result from the MMSE estimator for A_k derived in [2]. The behavior of these gain curves is explained in detail in [2], and this explanation holds as well for the new gain curves. It is interesting to note that the new gain function (which results from (20)) always gives lower gain than the one which results from the estimator of [2]. This is easy to prove by using Jensen's inequality:

$$\hat{A}_k = \exp\{E[\ln A_k\,|\,Y_k]\} \le \exp\{\ln E[A_k\,|\,Y_k]\} = E[A_k\,|\,Y_k] \quad (22)$$

© Technion - Israel Institute of Technology, Elyachar Central Library

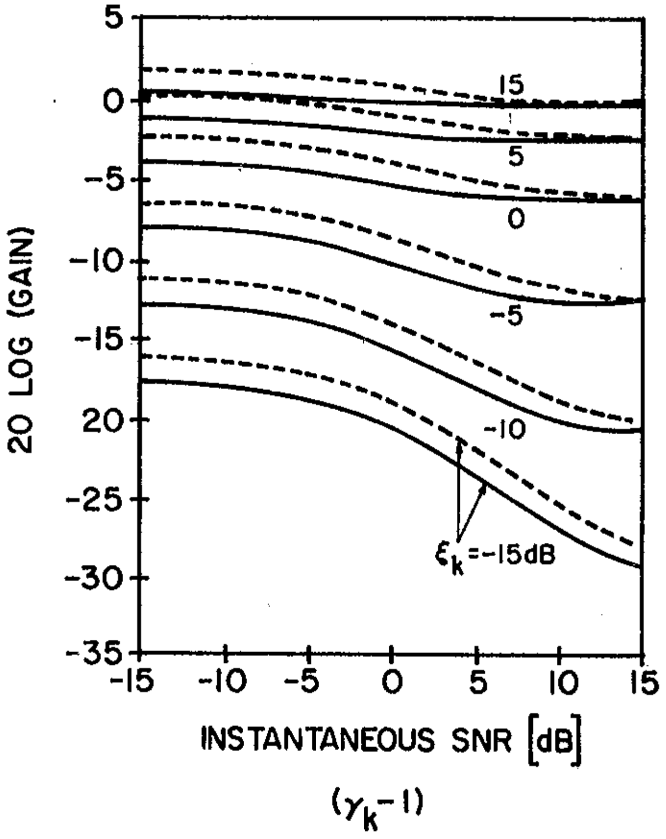


Fig. 1: Parametric gain curves describing:

(a) STSA estimator (20) (bold lines)

(b) MMSE-STSA estimator (7) in [2] (dashed lines)

ציור 1 : עקומי הגבר פרמטריים המתארים:

(a) - משערך האמפליטודה הספקטרלית (20).

(b) - משערך האמפליטודה הספקטרלית (7) ב-[2].

© Technion - Israel Institute of Technology, Elyachar Central Library

III. Derivation of Optimal Estimator Under Signal Presence Uncertainty

In this section we derived the optimal amplitude estimator which minimizes (1) under the assumed Gaussian statistical model, and the additional assumption that the signal is not surely present in the noisy observations. The resulting estimator is an extension of the estimators (20) and both estimators coincide as the probability of signal absence approaches zero.

By utilizing the statistical model for signal absence used in [2], which assumes statistically independent random appearance of the signal in the noisy spectral components, it can be shown that the optimal estimator is given by:

$$\hat{A}_k = \exp\left\{ \frac{\Lambda(Y_k, q_k)}{1+\Lambda(Y_k, q_k)} E[\ln A_k \mid Y_k, H_k^1] \right\} \tag{23}$$

where $\Lambda(Y_k, q_k)$ is the generalized likelihood ratio which is defined by:

$$\Lambda(Y_k, q_k) = \frac{1-q_k}{q_k} \frac{p(Y_k \mid H_k^1)}{p(Y_k \mid H_k^0)} \tag{24}$$

with q_k being the probability of signal absence in the k-th noisy spectral component, and H_k^0 and H_k^1 denote respectively the hypotheses of signal absence and presence. $E\{\ln A_k \mid Y_k, H_k^1\}$ is the optimal estimator of $\ln A_k$ given Y_k, assuming that the signal is surely present in the noisy spectral component Y_k. This estimator was already derived and is given by (20).

The generalized likelihood ratio $\Lambda(Y_k, q_k)$ was calculated for the assumed Gaussian model in [2], and equals to:

$$\Lambda(Y_k, q_k) = \frac{1-q_k}{q_k} \frac{\exp(v_k)}{1+\xi_k} \tag{25}$$

where now ξ_k is defined by:

$$\xi_k \triangleq \frac{E\{|X_k|^2 \mid H_k^1\}}{\lambda_d(k)} \tag{26}$$

$$= (1-q_k) \frac{E\{|X_k|^2\}}{\lambda_d(k)}$$

We note that the estimator (23) cannot be described by a gain function simi-

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

larly to the description of the estimator (20), since now the resulting gain function does not depend only on ξ_k and γ_k, but rather on the additional variable R_k.

IV. System Description and Performance Evaluation

Each of the two spectral amplitude estimators (20) and (23) was implemented in the speech enhancement system described in [2]. In this section we first briefly describe the above system, and then its performance.

System Description

The input to the speech enhancement system used here, is an 8kHz sampled speech of 0.2-3.2kHz bandwidth, which were degraded by uncorrelated additive noise. Each analysis frame of that input signal, which consists of 256 samples (32 msec) and overlaps the previous analysis frame by 192 samples, is spectrally decomposed by means of a discrete short-time Fourier transform (DSTFT) analysis [6], using a Hanning window. The amplitude of each speech DSTFT sample is then estimated, and the complex exponential of the noisy phase is used to produce the estimate of that DSTFT sample. The estimated DSTFT samples in each analysis frame are used for synthesizing the enhanced signal, by using the well known weighted overlap and add method [6].

In order to apply the amplitude estimators derived here, the variance $\lambda_d(k)$ of the k-th spectral component of the noise, and the a-priori SNR ξ_k, should be estimated. In the described system, $\lambda_d(k)$ is estimated once only (for a stationary noise process) from an initial non-speech segment of 320 msec of duration. This estimator of $\lambda_d(k)$ is used for calculating $\hat{\gamma}_k \triangleq R_k^2 / \hat{\lambda}_d(k)$ which is the estimator of γ_k (see (13)). The a-priori SNR ξ_k is estimated by the "decision-directed" estimator proposed in [2]. It is given by,

$$\hat{\xi}_k(n) = \alpha \frac{\hat{A}_k^2(n-1)}{\hat{\lambda}_d(k-1,n)} + (1-\alpha)P[\hat{\gamma}_k(n)-1] \tag{27}$$

where, $\hat{\xi}_k(n)$,$\hat{A}_k(n)$,$\hat{\lambda}_d(k,n)$ and $\hat{\gamma}_k(n)$, are the estimators of ξ_k,A_k,$\lambda_d(k)$ and γ_k, respectively, in the n-th analysis frame. $P[x]$ is defined by:

$$P[x] \triangleq \begin{cases} x & x \geq 0 \\ 0 & otherwise \end{cases} \tag{28}$$

Its function is to prevent (27) from being negative in case $\hat{\gamma}_k(n)-1$ is negative. The value of α was determined by informal listening, and its recommended value is α=0.98.

The calculation of the exponential integral which appears in the optimal estimator (20), is done efficiently by using appropriate polynomial approximation as explained in [5].

Performance Evaluation

The speech enhancement system described above was examined by informal listening in enhancing speech degraded by stationary uncorrelated additive white noise, with SNR values of 5dB, 0dB, and -5dB. The resulting enhanced speech was compared with that obtained in [2]. Fig. 2 shows a chart of the comparison tests considered here.

Case I: In this case we compare the STSA estimator (20) with the MMSE STSA estimator derived in [2, formula (7)]. The enhanced speech obtained by using (20) suffers much less residual noise, while no difference in the speech itself was noticed. The nature of the residual noise obtained with (20), sounds a little less uniform than when the MMSE STSA estimator is used. However, because of the lower residual noise level, this effect appears insignificant.

Case II: Here we compare the STSA estimator (20), with the estimator (23) which takes into account the signal presence uncertainty. With the latter estimator, a further reduction of the residual noise is obtained, but some effect of low pass

© Technion - Israel Institute of Technology, Elyachar Central Library

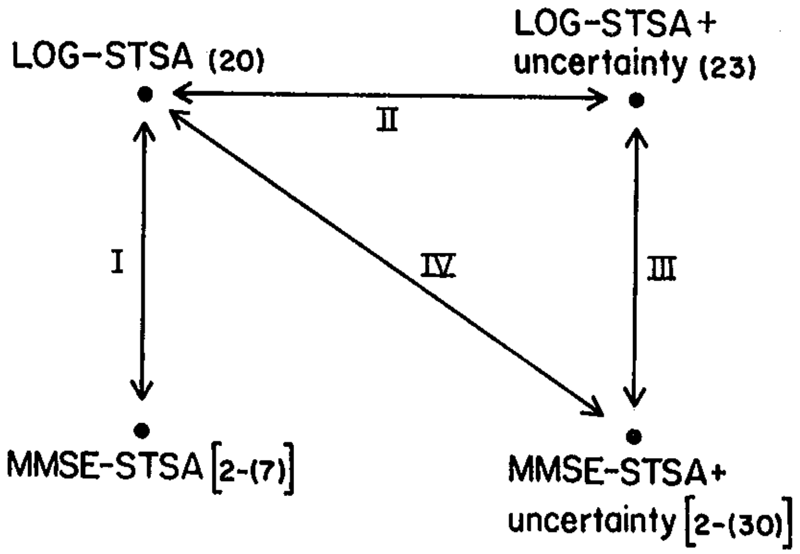


Fig. 2: Chart of comparison tests.

ציור 2 : תרשים למבחני השמיעה שבוצעו.

© Technion - Israel Institute of Technology, Elyachar Central Library

filtering of the speech signal was perceived. This effect is reduced as q_k is lowered, but then the amount of residual noise reduction gained by using this estimator ((23)), is also reduced. We found that using q_k=0.05 minimizes the above low-pass filtering effect, and still enables to reduce the residual noise level in comparison to that obtained by using the estimator (20).

Case III: In this case the two estimators (23) (with q_k=0.05) and (30) from [2], which take into account the speech presence uncertainty, are compared. We found that the residual noise level obtained when (23) is used, is lower than that obtained when the estimator from [2] is used. However, with the former estimator the above described low pass filtering effect was perceived.

Case IV: Here we compare the STSA estimator (20), with the MMSE STSA (30) from [2] which takes into account the signal presence uncertainty. We found that the enhanced speech obtained by both estimators sounds very similar, with the exception that with the first estimator the residual noise sounds a little less uniform.

V. Summary and Conclusions

In this paper we derive a STSA estimator which minimizes the mean square error of the log-spectra (i.e., the original STSA and its estimator), and examine it in enhancing speech. We found that this estimator is superior to the MMSE STSA estimator derived in [2], since it results in a much lower residual noise level, without further affecting the speech itself. In fact, the new estimator results in a very similar enhanced speech quality, as that obtained with the MMSE STSA estimator of [2], which takes into account the signal presence uncertainty. Since the new estimator can be implemented in a simpler way, it appears to be preferable.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

References

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using Optimal Non-Linear Spectral Amplitude Estimation", in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 1118-1121, Apr. 1983.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using an Optimal Non-Linear Spectral Amplitude Estimator", submitted for publication in IEEE Trans. Acoust., Speech, Signal Proc., May 1983.

[3] R.M. Gray, A. Buzo, A.H. Gray, Y. Matsuyama, "Distortion Measures for Speech Processing", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 367-376, Aug. 1980.

[4] I.S. Gradshteyn and Z.M. Ryzhik, Table of Integrals, Series, and Products, Academic Press, Inc., 1980.

[5] I.B.M Application Program, System/360 Scientific Subroutine Package (360A-CM-03X) Version III, pp. 368-369, 1968.

[6] R.E. Crochiere, "A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 99-102, Feb. 1980.

נספח ג' - שערוך יחס האות לרעש למטרות הדגשת דבור תוך שימוש באלגוריתם של ויטרבי

© Technion - Israel Institute of Technology, Elyachar Central Library

SIGNAL TO NOISE RATIO ESTIMATION FOR ENHANCING SPEECH USING THE VITERBI ALGORITHM

ABSTRACT

This paper deals with the problem of estimating the signal to noise ratio (SNR) of a noisy speech spectral component. Such an estimator is needed in the application of minimum mean square error estimators (e.g., Wiener). In this paper a maximum a-posteriori (MAP) estimator is proposed and compared with a maximum likelihood (ML) and a "decision-directed" SNR estimators. The proposed MAP estimator is implemented by using the Viterbi-algorithm. It is found that the MAP estimator results in a slightly better enhanced speech quality than the "decision-directed" estimator. Both estimators are superior in comparison to the ML one.

I. Introduction

Recently [1,2] we proposed an algorithm for enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available. This algorithm capitalizes on the major importance of the short-time spectral amplitude (STSA) in speech perception, and utilizes an optimal minimum mean square error (MMSE) STSA estimator for enhancing the speech signal.

The estimation problem of the MMSE STSA estimator as formulated in [2], is that of estimating the amplitude of each Fourier expansion coefficient of the speech signal $\{x(t), 0 \leq t \leq T\}$, given the noisy speech $\{y(t), 0 \leq t \leq T\}$. The derivation of the MMSE STSA estimator is based on the assumption that the spectral

© Technion - Israel Institute of Technology, Elyachar Central Library

components of the speech signal as well as of the noise process, can be modeled as statistically independent Gaussian random variables. The validity of this model for the discussed problem is given in detail in [2].

On the basis of the above statistical model, the estimation problem reduces to be that of estimating the amplitude of each Fourier expansion coefficient, given the corresponding noisy spectral component. The resulting estimator obtained in that way depends on two parameters, namely, the variances of the speech and noise spectral components. In the speech enhancement problem these parameters are unknown a-priori, as only the noisy speech is available. In addition, they are also time-varying due to the non-stationarity of speech signals, and possibly also of the noise process. Therefore, in order to apply the above STSA estimator, the variances of the speech and noise spectral components should be re-estimated in each analysis frame.

The estimate of the noise spectral component variance used in a given analysis frame, is commonly obtained from non-speech intervals which are most adjacent in time to that frame. This estimate is generally the average of periodograms belonging to these noise intervals [3,4]. Of course, this way of estimating the noise spectral component variance is suitable only for a stationary or at least a quasi-stationary noise process.

The estimation of the signal spectral component variance turns out to be a much more difficult problem, mainly due to the non-stationarity of the speech signal. In [2], the estimation of the signal spectral component variance, by a maximum likelihood (ML) method and a "decision-directed" approach, is considered. The "decision-directed" approach was found to be especially useful in the context of the discussed problem. In contrast with the ML estimation approach which results in musical residual noise in the enhanced speech, the "decision-directed" approach results in colorless residual noise, which is found

© Technion - Israel Institute of Technology, Elyachar Central Library

to be much less annoying and disturbing to the perception of the enhanced signal.

In this paper we deal with the problem of estimating a speech spectral component variance, and propose a maximum a-posteriori (MAP) estimator. The MAP estimation approach is an extension of the ML estimation method, as it incorporates a statistical model for generating each signal spectral component variance. Since the true model is inaccessible, the validity of the proposed one is judged a-posteriori, on the basis of the enhanced speech quality obtained when the MMSE STSA estimator is operated in conjunction with the MAP variance estimator. The MAP estimation approach results in a set of non-linear equations, which are solved here recursively by using the Viterbi-Algorithm (VA) [5,6].

Although the complexity of the above approach is high, its examination is of interest, since it provides a very flexible and systematic estimation approach for solving the difficult problem of estimating a speech spectral component variance. The MAP estimation approach is compared in this paper with the ML and the "decision-directed" estimation approaches.

The paper is organized as follows: In section II we formulate the MAP estimation problem of a speech spectral component variance, and describe its solution by the VA. In section III we compare the ML, "decision-directed", and the new MAP approaches. In section IV we summarize the paper and draw conclusions.

II. MAP Speech Spectral Component Variance Estimation

In this section we develop the MAP estimator of a speech spectral component variance, and describe its calculation by the VA. We fist present some preliminary material from [2], which is needed here.

© Technion - Israel Institute of Technology, Elyachar Central Library

Let $X_{k,n} \triangleq A_{k,n}\exp(j\alpha_{k,n})$, $D_{k,n}$ and $Y_{k,n} \triangleq R_{k,n}\exp(j\vartheta_{k,n})$, denote the k-th Fourier expansion coefficient of the speech signal, the noise process, and the noisy observation, respectively, in the n-th analysis frame. Under the Gaussian statistical model assumed in this paper, the optimal MMSE amplitude estimator $\hat{A}_{k,n}$ is given by [2]:

$$\hat{A}_{k,n} = G_{opt}(\xi_{k,n},\gamma_{k,n})R_{k,n} \tag{1}$$

where,

$$G_{opt}(\xi_{k,n},\gamma_{k,n}) \triangleq \Gamma(1.5)\frac{\sqrt{v_{k,n}}}{\gamma_{k,n}}M(-0.5;\ 1;\ -v_{k,n}) \tag{2}$$

is a multiplicative gain function which depends on $\xi_{k,n}$ and $\gamma_{k,n}$, which are defined by:

$$\xi_{k,n} \triangleq \frac{\lambda_x(k,n)}{\lambda_d(k,n)} \tag{3}$$

$$\gamma_{k,n} \triangleq \frac{R_{k,n}^2}{\lambda_d(k,n)} \tag{4}$$

$\lambda_x(k,n) \triangleq E\{|X_{k,n}|^2\}$, and $\lambda_d(k,n) \triangleq E\{|D_{k,n}|^2\}$ are respectively the variances of the k-th spectral component of the speech and noise in the n-th analysis frame. $\xi_{k,n}$ and $\gamma_{k,n}$ are interpreted as the a-priori and a-posteriori signal to noise ratio (SNR) respectively. $v_{k,n}$ is defined by:

$$v_{k,n} \triangleq \frac{\xi_{k,n}}{1+\xi_{k,n}}\gamma_{k,n} \tag{5}$$

$\Gamma(\cdot)$ is the Gamma function with $\Gamma(1.5)=\Gamma\pi/2$. $M(-0.5;\ 1;\ -v_{k,n})$ is the confluent hypergeometric function. For high SNR, it can be shown that $G_{opt}(\xi_{k,n},\gamma_{k,n})$ approaches the Wiener gain function $G_W(\xi_{k,n},\gamma_{k,n})$ given by [2]:

$$G_W(\xi_{k,n},\gamma_{k,n}) = \frac{\xi_{k,n}}{1+\xi_{k,n}} \tag{6}$$

The optimal gain function is described in [2] by means of parametric gain curves and its properties are discused there in detail. As noted above this gain function depends on the variances of the signal and noise spectral components, or equivalently, on the a-priori SNR $\xi_{k,n}$ and the noise spectral component

© Technion - Israel Institute of Technology, Elyachar Central Library

variance $\lambda_d(k,n)$. In this paper we assume that the noise is stationary and that the variances of its spectral components are known. Under this assumption $\lambda_d(k,n)=\lambda_d(k)$ for all n, and the problem of estimating the k-th speech spectral component variance $\lambda_x(k,n)$ can be considered in terms of the a-priori SNR $\xi_{k,n}$.

A very important result to the subject of this paper, is that a finite number of samples of the optimal gain function $G_{opt}(\xi,\gamma)$ can be used, with a negligible loss in performance [2]. For example, when the SNR of the input speech is between -5dB and 5dB, it suffices to consider 961 values of $G_{opt}(\xi,\gamma)$ obtained from a uniform sampling of the square region $-15dB \leq \xi,\ \gamma-1 \leq 15dB$. This means that the estimation problem of $\xi_{k,n}$ can be eased, and that only a finite set of its quantized values should be estimated. This observation leads to a very efficient implementation of the MAP estimator, by using the efficient dynamic programming algorithm of Viterbi [5,6].

MAP Estimation

The estimation problem of a speech spectral component variance $\lambda_x(k,n)$, expressed in terms of the a-priori SNR $\xi_{k,n}$, is formulated here as follows: Given a sequence of noisy observations $\underline{\gamma}_{k,N} \triangleq (\gamma_{k,1}, \gamma_{k,2}, \ldots, \gamma_{k,N})$, find the MAP estimate of the sequence $\underline{\xi}_{k,N} \triangleq (\xi_{k,1}, \xi_{k,2}, \ldots, \xi_{k,N})$. This is the problem of maximizing the a-posteriori probability density function (PDF) of $\underline{\xi}_{k,N}$ given $\underline{\gamma}_{k,N}$. The MAP sequence estimate $\hat{\underline{\xi}}_{k,N}$ is obtained by:

$$\begin{aligned} \hat{\underline{\xi}}_{k,N} &= arg \max_{\underline{\xi}_{k,N}} p(\underline{\xi}_{k,N} | \underline{\gamma}_{k,N}) \\ &= arg \max_{\underline{\xi}_{k,N}} p(\underline{\gamma}_{k,N} | \underline{\xi}_{k,N}) p(\underline{\xi}_{k,N}) \end{aligned} \qquad (7)$$

where $arg \max_x f(x)$ is the argument which maximizes $f(x)$, and the last equation follows from the fact that the maximization is done for a given sequence $\underline{\gamma}_{k,N}$.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

To solve the above problem, $p(\underline{\gamma}_{k,N}|\underline{\xi}_{k,N})$ and $p(\underline{\xi}_{k,N})$ should be known. $p(\underline{\gamma}_{k,N}|\underline{\xi}_{k,N})$ is easily obtained from the assumed Gaussian model and the additional assumption that the random variables constituting $\underline{\gamma}_{k,N}$ are statistically independent. The last assumption is reasonable when the analysis (i.e., the spectral decomposition of the noisy speech) is done on non-overlapping frames. However, in the system proposed in [2] the analysis is done with overlap. Nevertheless we continue with this assumption since the statistical dependence is difficult to be modeled and handled.

To obtain the PDF $p(\underline{\xi}_{k,N})$, some knowledge on the process by which the k-th speech spectral component variance $\lambda_x(k,n)$ is generated, is needed. We consider here a very simple model which reflects our belief in how $\lambda_x(k,n)$ could be generated. We assume the following Markov model;

$$\lambda_x(k,n+1) = \lambda_x(k,n) + \mathrm{w}(k,n) \qquad (8)$$

where $\mathrm{w}(k,n)$ is a sequence (in n) of statistically independent random variables. The PDF of $\mathrm{w}(k,n)$ is chosen so that given $\lambda_x(k,n)$, which is assumed to be in the range $[\lambda_{min},\lambda_{max}]$, $\lambda_x(k,n+1)$ has a prescribed probability p_k of remaining in a neighborhood of $\lambda_x(k,n)$, and probability of $1-p_k$ to be in the rest of the range $[\lambda_{min},\lambda_{max}]$. The values of $\lambda_x(k,n+1)$ within each of the two regions, are assumed to be equally likely. The values used for λ_{min} and λ_{max} are for example those which result in the a-priori SNR values of -15 and 15dB respectively, in the system from [2]. The boundaries $[\lambda_1,\lambda_2]$ of the neighborhood of $\lambda_x(k,n)$ are chosen to be proportional to $\lambda_x(k,n)$, and are determined as follows:

$$\lambda_1 = \begin{cases} b_1 \triangleq (1-\alpha_{k,n})\lambda_x(k,n) & \textit{if } b_1 \geq \lambda_{min} \\ \lambda_{min} & \textit{otherwise} \end{cases} \qquad (9)$$
$$\lambda_2 = \begin{cases} b_2 \triangleq (1+\alpha_{k,n})\lambda_x(k,n) & \textit{if } b_2 \leq \lambda_{max} \\ \lambda_{max} & \textit{otherwise} \end{cases}$$

where $\alpha_{k,n}$ depends on $\lambda_x(k,n)$ and satisfies $0 \leq \alpha_{k,n} \leq 1$. Its specific value was determined on the basis of informal listening, as will be described in section III.

Fig. 1 shows three typical cases of the PDF $p_{\lambda_x(k,n+1)|\lambda_x(k,n)}(\lambda)$ of $\lambda_x(k,n+1)$ given $\lambda_x(k,n)$.

It is interesting to note that the above PDF (of $\lambda_x(k,n+1)$ given $\lambda_x(k,n)$) can be considered as a weighted sum of two uniform PDF given by:

$$p_1(\lambda) = \begin{cases} \frac{1}{\lambda_{max}-\lambda_{min}} & \lambda_{min} \le \lambda \le \lambda_{max} \\ 0 & otherwise \end{cases} \tag{10}$$

$$p_2(\lambda) = \begin{cases} \frac{1}{\lambda_2-\lambda_1} & \lambda_1 \le \lambda \le \lambda_2 \\ 0 & otherwise \end{cases}$$

where $p_{\lambda_x(k,n+1)|\lambda_x(k,n)}(\lambda)$ is given by:

$$p_{\lambda_x(k,n+1)|\lambda_x(k,n)}(\lambda) = q_k p_1(\lambda)+(1-q_k)p_2(\lambda), \quad 0 \le q_k \le 1. \tag{11}$$

This interpretation of the proposed PDF is very useful in explaining our motivation for choosing that specific form. By so doing we assume that given $\lambda_x(k,n)$, $\lambda_x(k,n+1)$ is generated in accordance with one of two possible hypotheses, whose probabilities are q_k and $(1-q_k)$ respectively. The first hypothesis is defined by a situation in which a transition has occurred like from voiced speech in the n-th frame to unvoiced in the $n+1$-th frame, or from non-activity of speech (silence) to speech activity, etc. When this hypothesis occurs, it is reasonable to assume that $\lambda_x(k,n+1)$ takes any value in the range $[\lambda_{min},\lambda_{max}]$ with equal probability, and that this value is independent of the previous value $\lambda_x(k,n)$. The second hypothesis is defined by a situation in which speech activity of the same type (e.g., voiced, unvoiced) takes place during the n-th and the $n+1$-th frames. When this hypothesis happens, it is plausible to assume that $\lambda_x(k,n+1)$ strongly depends on $\lambda_x(k,n)$ and therefore should remain in its neighborhood. The probabilities q_k and p_k are related as can be easily seen. The value of p_k is chosen by informal listening as explained in section III.

© Technion - Israel Institute of Technology, Elyachar Central Library

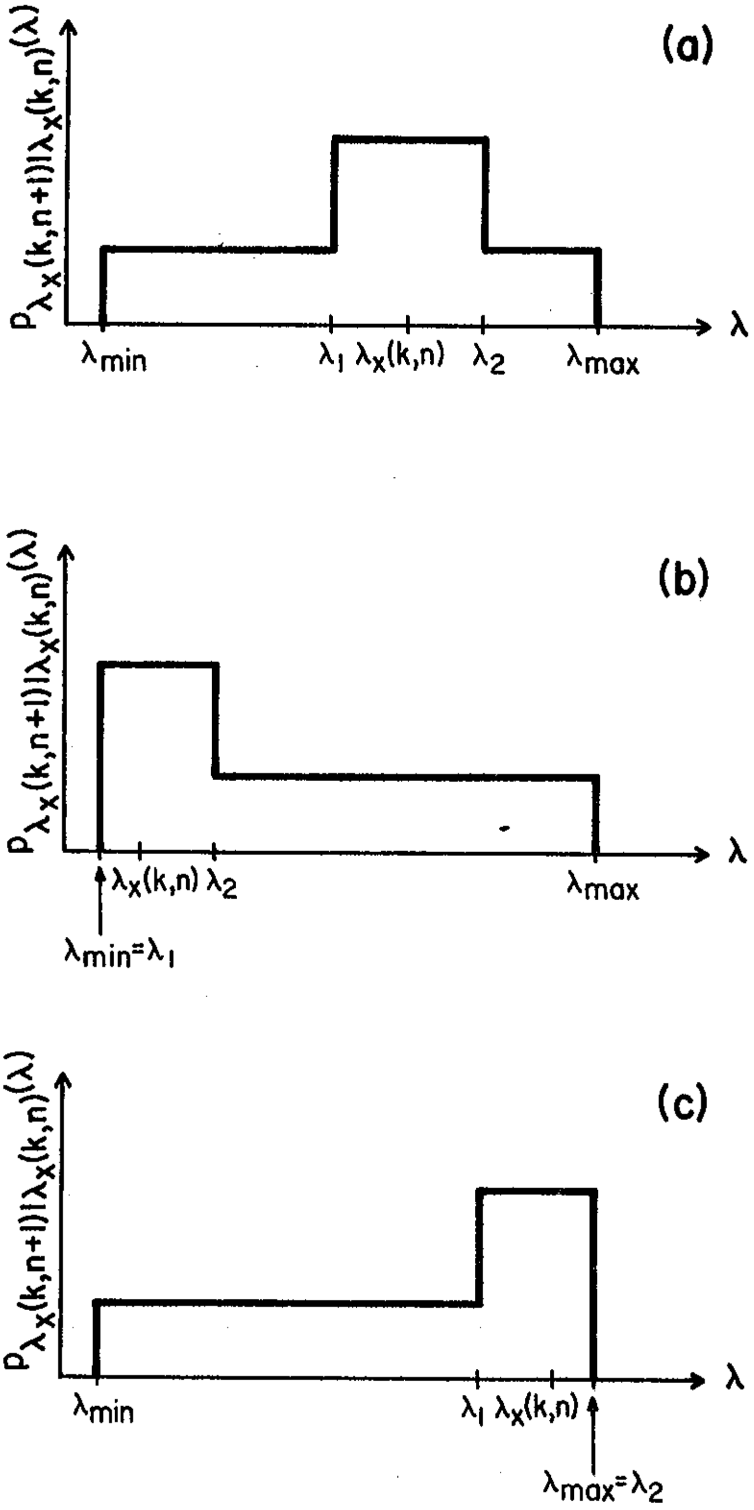


Fig. 1: Typical probability density functions of $\lambda_x(k,n+1)$ given $\lambda_x(k,n)$:

<u>ציור 1</u> : פונקציות פילוג סגולי אופייניות עבור $\lambda_x(k,n+1)$ בהינתן $\lambda_x(k,n)$.

(a) - $\lambda_{min}<\lambda_1$, $\lambda_2<\lambda_{max}$

(b) - $\lambda_{min}=\lambda_1$, $\lambda_2<\lambda_{max}$

(c) - $\lambda_{min}<\lambda_1$, $\lambda_2=\lambda_{max}$

© Technion - Israel Institute of Technology, Elyachar Central Library

It is worthwhile noting that other PDF types for $\lambda_x(k,n+1)$ given $\lambda_x(k,n)$ were examined. For example, a Gaussian PDF centralized at $\lambda_x(k,n)$ with variance proportional to $\lambda_x(k,n)$. However, the PDF given in (11) was found to be most successful in our application.

Now we are in a position to pursue the derivation of the a-priori SNR sequence estimator. It should be clear that since the noise variance $\lambda_d(k,n)$ equals to $\lambda_d(k)$ (i.e., time-invariant) and assumed to be known, the above model for generating $\lambda_x(k,n)$ is valid for the generation of $\xi_{k,n}$, when $\lambda_x(k,n)$ is appropriately normalized by $\lambda_d(k)$.

On the basis of the above assumptions, i.e., the independence of the random variable $(\gamma_{k,1}, \ldots, \gamma_{k,N})$ given $(\xi_{k,1}, \ldots, \xi_{k,N})$, and the Markov model assumed for the generation of $(\xi_{k,1}, \ldots, \xi_{k,N})$, $p(\underline{\gamma}_{k,N} | \underline{\xi}_{k,N})$ and $p(\underline{\xi}_{k,N})$ are given by:

$$p(\underline{\gamma}_{k,N} | \underline{\xi}_{k,N}) = \prod_{n=1}^{N} p(\gamma_{k,n} | \xi_{k,n}) \tag{12}$$

$$p(\underline{\xi}_{k,N}) = \prod_{n=1}^{N} p(\xi_{k,n} | \xi_{k,n-1}) \tag{13}$$

where $p(\xi_{k,1} | \xi_{k,0}) \stackrel{\Delta}{=} p(\xi_{k,1})$. Without any other a-priori information, $\xi_{k,1}$ is chosen to be uniformly distributed on some prescribed region (i.e., on $[\lambda_{min}/\lambda_d(k), \lambda_{max}/\lambda_d(k)]$). On substituting (12) and (13) into (7) we obtain:

$$\hat{\underline{\xi}}_{k,N} = arg \max_{\underline{\xi}_{k,N}} \prod_{n=1}^{N} p(\gamma_{k,n} | \xi_{k,n}) p(\xi_{k,n} | \xi_{k,n-1}) \tag{14}$$

$$= arg \max_{\underline{\xi}_{k,N}} \sum_{n=1}^{N} \ln p(\gamma_{k,n} | \xi_{k,n}) + \ln p(\xi_{k,n} | \xi_{k,n-1})$$

On the basis of the Gaussian statistical model assumed for the spectral components, it is easy to see that,

$$p(\gamma_{k,n} | \xi_{k,n}) = \frac{1}{1+\xi_{k,n}} \exp\left(-\frac{\gamma_{k,n}}{1+\xi_{k,n}}\right) \qquad \gamma_{k,n} \geq 0 \tag{15}$$

In addition, $p(\xi_{k,n} | \xi_{k,n-1})$ is specified above by (11) (with the suitable normalization by $\lambda_d(k)$). So in principle the problem of the MAP estimation as defined

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

by (14), is completely specified.

The Viterbi Algorithm

The maximization of (14) by simply taking the partial derivative and equating to zero, results in a set of non-linear equations which are difficult to solve. However, by taking advantage of the fact that for our purposes it suffices to estimate finite set of quantized values of $\xi_{k,n}$, we can apply here the efficient Viterbi algorithm to solve this maximization problem. For this reason, from now on we consider the estimation of a discretized version of $\xi_{k,n}$. We assume that each $\xi_{k,n}$ can take M equally spaced values in the range of $[\xi_{\min},\xi_{\max}]$. In the system discussed in [2], $\xi_{\min}$ and $\xi_{\max}$ correspond to -15dB and 15dB respectively, and M=31. Since $\xi_{k,n}$ is discretized, the PDF $p(\xi_{k,n}\,|\,\xi_{k,n-1})$ should be replaced by an appropriate probability measure. This can be done, for example, by discretizing the PDF $p(\xi_{k,n}\,|\,\xi_{k,n-1})$ as follows [6]: Let $\bar{p}(\xi_{k,n}\,|\,\xi_{k,n-1})$ denote the probability of $\xi_{k,n}$ to be equal to some value β_m, given that $\xi_{k,n-1}$ equals β_l. Then $\bar{p}(\xi_{k,n}\,|\,\xi_{k,n-1})$ can be obtained from $p(\xi_{k,n}\,|\,\xi_{k,n-1})$ by

$$\bar{p}(\xi_{k,n}=\beta_m\,|\,\xi_{k,n-1}=\beta_l) = \delta_l\, p(\xi_{k,n}=\beta_m\,|\,\xi_{k,n-1}=\beta_l) \tag{16}$$

where δ_l is chosen to satisfy:

$$\sum_{m=1}^{M}\bar{p}(\xi_{k,n}=\beta_m\,|\,\xi_{k,n-1}=\beta_l) = 1, \quad l=1,2,\ldots,M \tag{17}$$

By this way of discretization, $\bar{p}(\xi_{k,n}\,|\,\xi_{k,n-1})$ is completely defined by an $M\times M$ Markov matrix.

To show how (14) can be solved by the VA, we define a target function:

$$\Gamma_{k,N} \triangleq \sum_{n=1}^{N}\ln p(\gamma_{k,n}\,|\,\xi_{k,n})+\ln\bar{p}(\xi_{k,n}\,|\,\xi_{k,n-1}) \tag{18}$$

and re-write the maximization problem (14) as:

$$\max_{\{\xi_{k,n}\}_{n=1}^{N}} \Gamma_{k,N} \tag{19}$$

where $\{\xi_{k,n}\}_{n=1}^{N} = \underline{\xi}_{\,k,N}$. Since

$$\Gamma_{k,N} = \begin{cases} \Gamma_{k,N-1} + \ln p(\gamma_{k,N} \mid \xi_{k,N}) + \ln \bar{p}(\xi_{k,N} \mid \xi_{k,N-1}) & N>1 \\ \ln p(\gamma_{k,1} \mid \xi_{k,1}) + \ln \bar{p}(\xi_{k,1} \mid \xi_{k,0}) & N=1 \end{cases} \tag{20}$$

the maximization problem for $N>1$ can be written as [6]:

$$\max_{\{\xi_{k,n}\}_{n=N-1}^{N}} \{ \max_{\{\xi_{k,n}\}_{n=1}^{N-2}} \Gamma_{k,N-1} + \ln p(\gamma_{k,N} \mid \xi_{k,N}) + \ln \bar{p}(\xi_{k,N} \mid \xi_{k,N-1}) \} \tag{21}$$

where

$$\ln p(\gamma_{k,N} \mid \xi_{k,N}) + \ln \bar{p}(\xi_{k,N} \mid \xi_{k,N-1}) \tag{22}$$

is called the path-metric in the VA terminology. This form yields the basis for the VA. According to (21), the maximizing trajectory $\{\hat{\xi}_{k,n}\}_{n=1}^{N}$ while passing through $\hat{\xi}_{k,N-1}$ on its way to $\hat{\xi}_{k,N}$, should arrive at $\hat{\xi}_{k,N-1}$ along a route $\{\hat{\xi}_{k,n}\}_{n=1}^{N-2}$ that maximizes $\Gamma_{k,N-1}$. Otherwise, $\hat{\xi}_{k,N-1}$ and $\hat{\xi}_{k,N}$ could be retained, and $\{\hat{\xi}_{k,n}\}_{n=1}^{N-2}$ could be replaced with a different sequence which would increase the value of $\Gamma_{k,N}$.

The maximization procedure by the VA is well documented in [5,6], and therefore will be explained here very briefly. The algorithm begins by calculating $\Gamma_{k,1}$ (see (20)) for each of the M states of $\xi_{k,1}$, based on the first noisy measurement $\gamma_{k,1}$. $\bar{p}(\xi_{k,1} \mid \xi_{k,0})$ is connected with the initial conditions, and if no a-priori information in favor of a specific state of $\xi_{k,1}$ is known, all states are assumed to be equally likely, and therefore $\bar{p}(\xi_{k,1} \mid \xi_{k,0}) = 1/M$. After completing this stage, the measurement $\gamma_{k,1}$ is not needed anymore, and therefore can be discarded. As the new measurement $\gamma_{k,2}$ arrives, the algorithm calculates $\Gamma_{k,2}$ for each terminating state β_l, $l=1,\ldots,M$ at $N=2$, and every beginning state β_m at $N=1$. Let the resulting values be denoted by $\Gamma_{k,2}(\beta_l,\beta_m)$. Then the algorithm selects M sequences $(\beta_l,\hat{\beta}_m(l))$ (called survivors) which are terminating at β_l, $l=1,\ldots,M$, and starting at $\hat{\beta}_m(l) \triangleq arg \max_{\beta_m} \Gamma_{k,2}(\beta_l,\beta_m)$. All other sequences (β_l,β_m), $\beta_m \neq \hat{\beta}_m(l)$ are discarded, as they are not candidates for the maximizing trajectory. The above procedure can of course be done for each state β_l separately, where for each l one survivor is chosen. At the end of this stage, the

measurement $\gamma_{k,2}$ can be discarded. The procedure is continued in a similar fashion as each new measurement $\gamma_{k,n}$ arrives.

It is well known that the estimated sequence $\{\hat{\xi}_{k,n}\}_{n=1}^{N} \triangleq (\hat{\xi}_{k,1}(N), \hat{\xi}_{k,2}(N), \ldots, \hat{\xi}_{k,N}(N))$ obtained by VA on the basis of N measurements, may differ from the estimated sequence $\{\hat{\xi}_{k,n}\}_{n=1}^{N+1} \triangleq (\hat{\xi}_{k,1}(N+1), \hat{\xi}_{k,2}(N+1), \ldots, \hat{\xi}_{k,N}(N+1))$ obtained on the basis of $N+1$ measurements for all $n=1,\ldots,N$ [5-7]. However, if the decision on an estimated value is done with delay (say n_0), then with high probability this estimated value will be the true MAP estimate. This is due to the fact that for high values of N, the "tail" of all survivors tend to merge. Therefore it is a common practice to use the VA with a delayed decision, i.e., the estimated value obtained at time n is:

$$\hat{\xi}_{k,n-n_0}(n), \quad n=n_0+1,\ n_0+2, \ldots \tag{23}$$

III. System Description and Performance Evaluation

The MAP estimator of the a-priori SNR $\xi_{k,n}$ (assuming the noise spectral component variance is known and is time-invariant) described in this paper, was examined in the speech enhancement system used in [2]. It is utilized for estimating the a-priori SNR when either the MMSE-STSA estimator (1) or the Wiener STSA estimator (6) is used.

In this section we first briefly describe the above system. Then we give the exact conditions used for the application of the VA. Finally, we describe the performance of that system and compare it with the one described in [2].

System Description

The input to the speech enhancement system used here, is an 8kHz sampled speech of 0.2-3.2kHz bandwidth, which were degraded by uncorrelated additive noise. Each analysis frame of that input signal, which consists of 256

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

samples and overlaps the previous frame by 192 samples, is spectrally decomposed by means of a discrete short-time Fourier transform (DSTFT) analysis [8], using a Hanning window. The amplitude of each DSTFT sample is estimated, on the basis of the estimated values of the noise spectral component variance and the a-priori SNR. Each estimated spectral amplitude is combined with the complex exponential of the noisy DSTFT sample, to produce the estimate of the speech DSTFT sample. The estimated DSTFT samples in each analysis frame are used for synthesizing the enhanced signal, by using the well known weighted overlap and add method [8].

In the experiments done here, a stationary white noise is used. The variances of its spectral component are estimated only once, from an initial non-speech segment of 320 msec of duration. In addition, the input SNR examined here was 5 and 0dB.

The VA is applied for estimating the a-priori SNR of each DSTFT sample, under the following conditions:

- Lower bound for a-priori SNR value $\xi_{\min} = -15dB$.
- Upper bound for a-priori SNR value $\xi_{\max}=15dB$.
- Number of linear quantization levels in the range $[\xi_{min}, \xi_{\max}]$ is $M=51$.
- Delay of decision is $n_0=2$.
- The value used for the parameter $\alpha_{k,n}$ in (9), is determined as follows:

$$\alpha_{k,n} = 0.8 + \frac{0.8-0.2}{\xi_{\max}-\xi_{\min}}(\xi_{\min}-\xi_{k,n}) \tag{24}$$

  and there are M different values, since $\xi_{k,n}$ can take M values only.
- The probability p_k of $\lambda_x(k,n+1)$ to remain in the neighborhood $[\lambda_1,\lambda_2]$ of $\lambda_x(k,n)$ equals 0.5.

The above parameters were chosen on the basis of informal listening, and are the best ones found here. We found that both parameters $\alpha_{k,n}$ and p_k are

very dominant here. For example, if $p_k \to 1$ and $\alpha_{k,n} \ll 1$, that is if $\lambda_x(k,n+1)$ is constrained to remain in some neighborhood of $\lambda_x(k,n)$, then a very reverberant enhanced speech results. On the other hand, if $\alpha_{k,n} \to 1$ then the distribution of $\lambda_x(k,n+1)$ given $\lambda_x(k,n)$ approaches a uniform one, and the MAP estimator turns out to be a ML estimator. As was noted in section I, the latter estimator results in musical residual noise in the enhanced speech, which is very undesirable.

Performance Evaluation

The above system was examined in enhancing speech degraded by uncorrelated white additive noise with SNR of 5 and 0dB. The optimal MMSE STSA estimator (1) and the Wiener estimator (6), operating with the MAP estimator of the a-priori SNR, were examined. The resulting enhanced speech was compared with that obtained in [2], where a "decision-directed" a-priori SNR estimator is used with each of the two STSA estimators.

In general, when the MAP estimator is used rather than the "decision-directed" one, with either the optimal MMSE or the Wiener STSA estimators, a slightly better enhanced speech quality is obtained. Specifically, a lower residual noise level is obtained in some cases. In addition, that residual noise sounds more uniform (white) in all cases examined here. The speech itself sounds very similar when both a-priori SNR estimators are used.

IV. Summary and Conclusions

In this paper we deal with the problem of estimating the variance of a speech spectral component. Such an estimator is needed in the application of MMSE estimators, like that of the STSA (1) or of the STFT (6). The MAP estimation approach proposed here, incorporates for the first time a model for generating the time-varying variance of a speech spectral component. While this

© Technion - Israel Institute of Technology, Elyachar Central Library

model has a heuristic basis, it is found to be reliable, in light of the good enhanced speech quality obtained here. The above approach is basically different from the methods used so far in the context of speech enhancement (e.g., the power spectral subtraction method [4] and its extension in [2]), which inherently belong to spectral estimation methods which are based on the periodogram [4,9]. In this paper we develop the MAP estimator and describe its efficient implementation by the VA.

Utilizing the above estimator rather than a previously developed "decision-directed" estimator, does not provide a significant improvement in the enhanced speech quality. This fact implies perhaps on some limit of our basic model (in which the spectral component of speech and noise are assumed to be statistically independent random variables), which was already reached by the very simple "decision-directed" estimator. We conjecture that removing the statistical independence assumption may improve the speech enhancement results.

© Technion - Israel Institute of Technology, Elyachar Central Library

References

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using Optimal Non-Linear Spectral Amplitude Estimation", in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 1118-1121, Apr. 1983.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Spectral Amplitude Estimator", submitted for publication in IEEE Trans. Acoust., Speech, Signal Proc., May 1983.

[3] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", Proc. IEEE, Vol. 67, pp. 1586-1604, Dec. 1979.

[4] R.J. McAulay and M.L. Malpass, "Speech Enhancement Using a Soft Decision Noise Suppression Filter", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 137-145, Apr. 1980.

[5] G.D. Forney, "The Viterbi Algorithm", Proc. IEEE, Vol. 61, No. 3, pp. 268-278, March 1973.

[6] L.L. Scharf, D.D. Cox, J. Masreliez, "Modulo- 2π Phase Sequence Estimation", IEEE Trans. Inform. Theory, Vol. IT-26, No. 5, pp. 615-620, Sept. 1980.

[7] C.R. Cahn, "Phase Tracking and Demodulation With Delay", IEEE Trans. Inform. Theory, Vol. IT-20, pp. 50-58, Jan. 1974.

[8] R.E. Crochiere, "A Weighted Overlap-Add Method for Short-Time Fourier Analysis/Synthesis", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 99-102, Feb. 1980.

[9] W.B. Davenport and W.L. Root, An Introduction to the Theory of Random Signals and Noise, McGraw-Hill, 1958, New York, pp. 105-108.

© Technion - Israel Institute of Technology, Elyachar Central Library

<u>נספח ד'</u> - שלוב הדגשה וקדוד מסתגל בתחום התדר של אותות דבור רועשים

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

COMBINED ENHANCEMENT AND ADAPTIVE TRANSFORM CODING OF NOISY SPEECH.

Abstract

This paper deals with the problem of improving the quality of reconstructed speech, obtained by using an adaptive transform coder (ATC) which operates on noisy speech. We propose to estimate the short-time spectral amplitude of the speech signal, and to utilize the noisy phase, prior to the encoding process. The appropriate minimum mean square error estimator is derived in this paper. The above approach significantly improves the quality of the reconstructed speech, although it loses some of its crispness. The input noise is suppressed, and the irregularities characteristic to the reconstructed noisy speech almost disappear.

© Technion - Israel Institute of Technology, Elyachar Central Library

I. Introduction

This paper deals with the problem of improving the quality of reconstructed speech, obtained by using an adaptive transform coder (ATC) which operates on noisy speech. The ATC is a waveform coder which was found to be very efficient for encoding speech at rates of $7.2-16kb/s$. At the rate of $16kb/s$ or above it gives toll quality, while at the rate of $7.2kb/s$ it results in communication quality [1-3].

The ATC quantizes the speech spectral components in each analysis frame, in accordance with a dynamic bit allocation and a variable quantization step-size. Thus, the perceptually more important spectral components can be traced and better quantized. The bit assignment and the step-size used for each spectral component are determined on the basis of knowledge of its variance. These variances of the spectral components are obtained from a parametric estimated spectrum of the speech signal in that frame. This parameterized spectrum is encoded and transmitted as side-information to the receiver.

Although the ATC belongs to the class of waveform coders, which are supposed to be robust, it turns out that it is sensitive to background noise. Unlike other speech waveform coders (e.g. PCM, DPCM, etc) in which the input noise is reflected to the reconstructed output speech, here the noise has additional effect: It adversely influences the extraction of the parameters representing the speech short-time spectrum. This results in wrong bit allocation and quantization step-size, which cause further degradation of the reconstructed speech.

The structure of the ATC calls for a very convenient way for improving the quality of the reconstructed speech, obtained under noisy environment. Specifically, since the encoding is already done in the frequency

domain, we can optimally estimate the perceptually important short-time spectral amplitude (STSA) of the speech signal, and use the noisy phase, prior to the encoding process. This way is similar to that taken in [4], but here the transform is the discrete cosine transform (DCT) rather than the discrete Fourier transform (DFT). This fact does not change the perceptual significance of the STSA, since as is well known [3] both transforms have the same spectral envelope. However, the appropriate estimator should be re-derived.

The paper is organized as follows: In section II we briefly describe the ATC scheme used in this work. Then in section III we formulate the estimation problem of the STSA, and derive its minimum mean square error (MMSE) estimator. In section IV we describe the performance of the ATC operating on noisy speech, with and without the above enhancement. In section V we discuss the results obtained here, and briefly describe some experiments we have done to reduce also the quantization noise level.

II. Adaptive Transform Coding

The ATC has been extensively investigated and several schemes were proposed [1-3]. In this work we focus on the so-called "speech-specific" ATC which is well documented in [3]. Before describing that scheme, we briefly discuss the DCT and some of its properties. The DCT is used by the ATC rather than other transforms like the DFT, since on the basis of a mean square error (MSE) criterion, the DCT was found to be nearly optimal for speech encoding, relative to the optimal Karhunen-Loeve transform [1,5]. In addition, on an experimental basis, the DCT was found to perform better than the DFT [1]. Finally, the DCT reduces block-end effect problems as explained in [3].

© Technion - Israel Institute of Technology, Elyachar Central Library

The DCT of a real M point sequence x_n is defined by:

$$X_k = \sum_{n=0}^{M-1} x_n c_k \cos[(2n+1)\pi k / 2M] \quad k=0,1,\ldots,M-1 \tag{1}$$

where

$$c_k = \begin{cases} 1 & k=0 \\ \sqrt{2} & k=1,2,\ldots,M-1 \end{cases} \tag{2}$$

The inverse DCT is defined similarly as:

$$x_n = \frac{1}{M}\sum_{k=0}^{M-1} X_k c_k \cos[(2n+1)\pi k / 2M] \quad n=0,1,\ldots,M-1 \tag{3}$$

The DCT can be efficiently calculated by using an FFT of M points [6]. However it can also be obtained by applying a $2M$ point DFT on the sequence u_n defined by:

$$u_n \triangleq \begin{cases} x_n & n=0,1,\ldots,M-1 \\ 0 & n=M,\ldots,2M-1 \end{cases} \tag{4}$$

By so doing it can be shown that the DCT X_k and the DFT U_k are related by:

$$X_k = c_k \mid U_k \mid \cos(\vartheta_k - \pi k / 2M) \quad k=0,1,\ldots,M-1 \tag{5}$$

where ϑ_k represents the phase of U_k. This form supplies an interesting spectral interpretation of the DCT. It shows that the DCT and the DFT have the same spectral envelopes (up to the constant c_k) [3].

The speech-specific ATC scheme is depicted in Fig. 1. According to this figure, each frame of the input speech is first cosine transformed. Then the parameters representing the estimated spectrum of the speech in that frame are extracted and quantized. These parameters include the linear prediction coefficients (LPC), the pitch period, and the gain, which appear in the basic model of speech production. These parameters are obtained by using an estimate of the speech autocorrelation function. This function is obtained by inverse transforming the square of the DCT components, and thus the fact that the DCT and the DFT have the same spectral envelope is utilized.

© Technion - Israel Institute of Technology, Elyachar Central Library

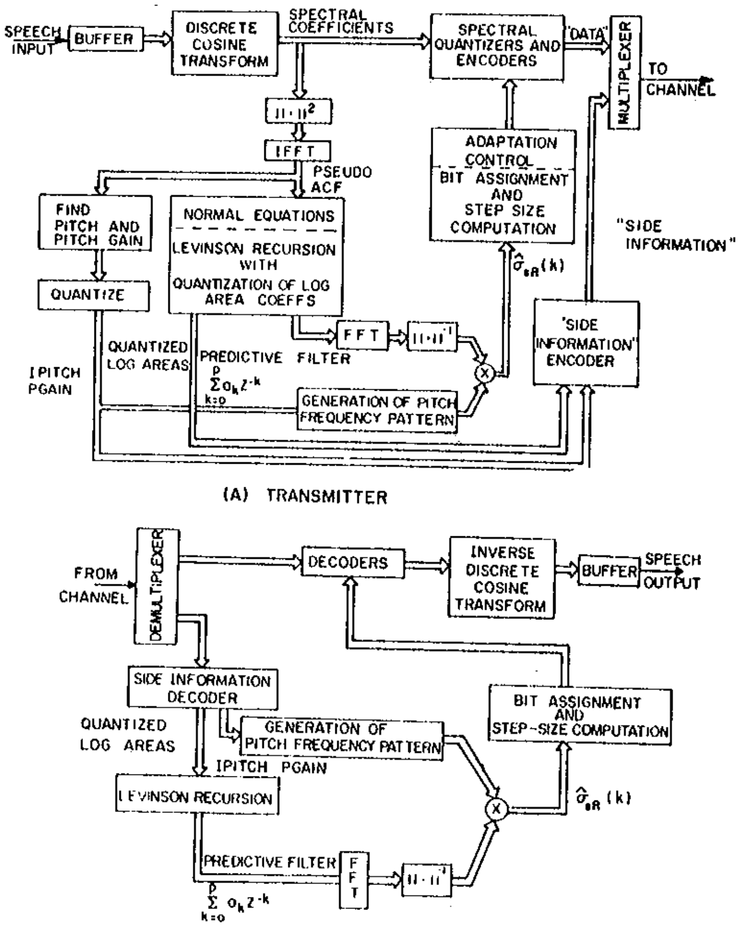


Fig. 1: **Block diagram of ATC.**

<u>ציור 1</u> : סכימת בלוקים של מקדד ה-ATC (הציור לקוח מ-[2]).

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

The LPC and the gain are used for estimating the speech spectral envelope, while the pitch period and the so-called "pitch-gain" [3] are used for estimating the fine details of the spectrum, called the pitch pattern. The spectral envelope and the pitch pattern are combined (via multiplication) to produce the estimated speech spectrum. This spectrum is used in the transmitter for allocating the B bits available for encoding each frame, and for normalizing the DCT components prior to their optimal quantization.

The optimal bit allocation and quantization of the DCT components are based on the observation that they are approximately Gaussian distributed [1,3]. The bit assignment rule results from the solution of the following optimization problem:

$$\min_{B_k} \frac{1}{M} \sum_{k=0}^{M-1} w_k \varepsilon_k \tag{6}$$

subject to: $\sum_{k=0}^{M-1} B_k = B$

where B_k is the number of bits assigned to the k-th DCT component, ε_k is the resulting distortion in that component, and w_k is a positive weighting function. B_k and ε_k are related by the rate-distortion function of a Gaussian source:

$$B_k = \frac{1}{2} \log_2 \frac{\sigma_k^2}{\varepsilon_k} \tag{7}$$

where σ_k^2 is the variance of the k-th DCT component. The optimal bit allocation and the resulting distortion are given by:

$$B_k = \bar{B} + \frac{1}{2} \log_2 \frac{w_k \sigma_k^2}{[\prod_{l=1}^{M} w_l \sigma_l^2]^{1/M}} \tag{8}$$

$$\varepsilon_k = 2^{-2\bar{B}} [\prod_{l=1}^{M} w_l \sigma_l^2]^{1/M} \; w_k^{-1} \qquad k=0,1,\ldots,M-1 \tag{9}$$

where $\bar{B} \triangleq B/M$. σ_k^2 is obtained as the k-th component of the estimated

spectrum. A useful weighting function was proposed in [3]. It is given by:

$$w_k = \sigma_{s_k}^{2\gamma} \qquad -1<\gamma<0 \tag{10}$$

where $\sigma_{s_k}^2$ is the estimated spectrum without the pitch pattern, i.e., it is the estimated spectral envelope. Since ε_k is proportional to w_k^{-1} (see (9)), this weighting function results in a quantization noise spectrum which follows that of the speech. As a consequence, low energy spectral components will not be masked by the quantization noise.

The quantization of each DCT component is done by first normalizing it by its estimated standard deviation, and then by using the optimal normalized quantization step-size for a Gaussian source derived by Max [7].

At the receiver the bit stream is decoded, and the spectrum is reconstructed from the side-information. With this spectrum available, the receiver can follow the bit allocation and the DCT normalization done in the transmitter.

In this work we examined the ATC at 12 and 16kb/s. The specific parameters of the coder used here, are those recommended in [3]. Specifically, the transform size M is 256; speech sampling rate is 8kHz; block overlap equals 16 samples, using a trapezoidal window; maximal number of quantizer bits is 4; quantizer loading parameter (to multiply Max quantization step-size) equals 1.3 for 12 kb/s, and 1.5 for 16 kb/s; noise shaping parameter $\gamma = -.125$; number of LPC is 9. In this work the side information parameters are not quantized but the number of bits needed for their quantization (44) was taken into account. We avoided from quantizing these parameters, as our main objective here is to examine the enhancement of the reconstructed noisy speech. In addition, very efficient ways for quantizing these parameters are available.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

III. MMSE Spectral Amplitude Estimator

As explained in section I, we propose to improve the quality of the reconstructed noisy speech, by using a MMSE estimate of the speech STSA prior to the encoding process. In the context of the ATC this means that we have to estimate the amplitude (i.e., the absolute value) of each speech DCT component, given the noisy DCT components.

The above estimation problem can be simplified by taking advantage of asymptotic statistical properties of the DCT components. Specifically, we can reasonably assume that the DCT components of the speech signal, as well as of the noise process, can be modeled as statistically independent Gaussian random variables. The Gaussian assumption is motivated by the central limit theorem. The statistical independence assumption results from the Gaussian model and the fact that the correlation between the DCT components reduces as the analysis interval length increases. It is satisfying to note that Zelinski and Noll [1,3] arrived at the same statistical model for the speech signal, on an experimental basis. This fact was already utilized in section II, where we discussed the optimal bit allocation and quantization of the speech DCT components. For the noise process one can obtain the above model or simply assume that the noise is Gaussian.

On the basis of the above statistical model, it is easy to see that the estimation problem reduces to that of estimating the amplitude of each speech DCT component, given the corresponding noisy DCT component. Let X_k, D_k, and Y_k denote respectively the DCT component of the speech, the noise, and the noisy process. Then the MMSE estimator of $|X_k|$ is given by:

$$
\begin{aligned}
|\hat{X}_k| &= E\{\,|X_k| \,\big|\, Y_0, Y_1, \ldots, Y_{M-1}\} \qquad (11)\\
&= E\{\,|X_k| \,\big|\, Y_k\}
\end{aligned}
$$

$$= \int_{-\infty}^{\infty} |x_k| p(x_k|y_k) dx_k$$

On the basis of the above Gaussian assumption, and by using [6:3.562.4,3.546.2], it can be shown that:

$$|\hat{X}_k| = \frac{\xi_k}{1+\xi_k}\left[\Phi\left(\sqrt{\frac{v_k}{2}}\right) + \sqrt{\frac{2}{\pi}\frac{1}{v_k}}\exp\left(-\frac{v_k}{2}\right)\right]|Y_k| \tag{12}$$

where $\Phi(\cdot)$ is defined by:

$$\Phi(x) \triangleq \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{13}$$

v_k is defined by:

$$v_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k \tag{14}$$

ξ_k and γ_k are defined by:

$$\xi_k \triangleq \frac{E\{|X_k|^2\}}{\lambda_d(k)} \tag{15}$$

$$\gamma_k = \frac{|Y_k|^2}{\lambda_d(k)} \tag{16}$$

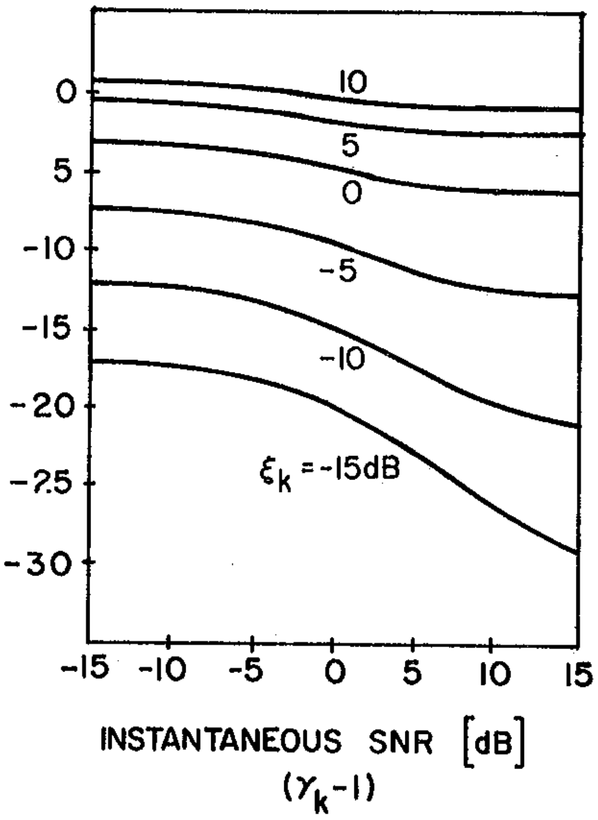
where $\lambda_d(k) \triangleq E\{|D_k|^2\}$. ξ_k and γ_k are interpreted as the a-priori and a-posteriori signal to noise ratio (SNR) respectively. The MMSE estimator (12) is conveniently described by a gain function defined by:

$$G(\xi_k, \gamma_k) \triangleq \frac{|\hat{X}_k|}{|Y_k|} \tag{17}$$

This gain function is described in Fig. 2 by parametric gain curves. The behavior of these gain curves is similar to that of the gain curves obtained in [4], where the amplitude of a DFT component is estimated. The explanation given there for the shape of the gain curves holds as well for the problem discussed here.

The estimate of the k-th speech DCT component is obtained by combining the above MMSE amplitude estimator (12), with the phase (i.e., the sign) of the k-th noisy spectral component. That is,

© Technion - Israel Institute of Technology, Elyachar Central Library



<u>ציור 2</u> : עקומי הגבר פרמטריים המתארים את משערך האמפליטודה הספקטרלית (12) של ה-DCT.

Fig. 2: Parametric gain curves describing the spectral amplitude estimator (12) of the DCT.

© Technion - Israel Institute of Technology, Elyachar Central Library

$$\hat{X}_k = |\hat{X}_k| \quad \frac{Y_k}{|Y_k|} \tag{18}$$

$$= G(\xi_k, \gamma_k)\ Y_k$$

To implement the above estimator, the noise variance $\lambda_d(k)$ and the a-priori SNR ξ_k should be known. In the experiments we carried out here we examined stationary noise, and estimated its variances once only from an initial non-speech interval of 640msec in duration. The a-priori SNR is estimated by the "decision-directed" estimator proposed in [4]. This estimator is given by:

$$\hat{\xi}_{k,n} = \alpha \frac{|\hat{X}_{k,n-1}|^2}{\lambda_d(k,n-1)} + (1-\alpha)P[\gamma_{k,n}-1] \tag{19}$$

where $\xi_{k,n}$, $|X_{k,n}|$, $\lambda_d(k,n)$, and $\gamma_{k,n}$ are respectively the a-priori SNR, the speech spectral amplitude, the noise variance, and the a-posteriori SNR of the corresponding k-th DCT component in the n-th analysis frame. $P[\cdot]$ is an operator defined by:

$$P[x] = \begin{cases} x & x \geq 0 \\ 0 & otherwise \end{cases} \tag{20}$$

and its function is to prevent $\hat{\xi}_{k,n}$ from being negative, if $(\gamma_{k,n}-1)$ is negative. α is an averaging parameter which is determined on the basis of informal listening. It was found that its best value for 16 kb/s is 0.94, whereas for 12 kb/s it is 0.85.

IV. Performance Evaluation

The STSA estimator (12) was applied in the "speech-specific" ATC described in section II, and the system was examined in encoding noisy speech. Speech signals which were degraded by uncorrelated additive wide-band noise, with SNR of 10 and 5dB, were examined.

When the ATC was operated on the noisy speech but no enhancement is done, a noisy reconstructed speech results. In addition, it has some

© Technion - Israel Institute of Technology, Elyachar Central Library

noticeable irregularities which strongly influences its quality and intelligibility. These irregularities are probably a result of using a wrong bit allocation and an incorrect quantization step-size, due to the poor estimate of the speech spectrum from the noisy input speech. By applying the STSA estimator (12) to the above ATC system, the quality of the reconstructed speech is greatly improved, although it loses some of its crispness. Specifically, the input noise is suppressed, and the above mentioned irregularities almost disappear.

V. Discussion

In this paper we examine the operation of the ATC under a noisy environment. It was noted that its performance significantly degrades when the input speech is noisy. By utilizing a MMSE STSA estimator, which is applied prior to the encoding process, we achieve a great improvement in the ATC performance. Since the ATC is already operated in the frequency domain, the above enhancement method is well suited to its structure and can be easily implemented. For example, a look-up table which contains a finite number of samples of the multiplicative enhancing gain function can be used.

It is worthwhile noting that in the course of this work we also examined the possibility of improving the ATC performance when it operates on clean speech. Specifically, we examined three alternative approaches. In the first, we tried to reduce the quantization noise level in each quantized DCT component, by estimating the speech DCT component. In the second alternative, we tried to decorrelate the speech and the quantization noise in each DCT component, by using dithered quantization [9,10]. In the third alternative, we unified the above two approaches and tried to estimate the speech DCT component from the dithered quantized noisy one. The first

© Technion - Israel Institute of Technology, Elyachar Central Library

and third approaches are reasonable of course, only if the quantizer used is not optimal (in Max sense [7]). However, this is the case here since a uniform quantizer is used. In both cases we obtained that the estimated speech DCT component given the quantized noisy component, is the centroid of the area of the probability density function of the speech DCT component, in the quantizer step-size interval. This is an interesting result since it coincides with that obtained by Max in designing the optimal non-uniform quantizer. However, in his work the optimization is also done on the quantization step-size interval. We note finally, that at this problem we estimate the speech DCT component rather than its absolute value, since in this application the sign of the DCT component is known exactly, and both approaches are identical.

Unfortunately, the above three approaches do not improve the ATC performance. One possible explanation is that on the basis of MSE criterion, the expected improvement is bounded by that which can be obtained by using Max optimal non-uniform quantizer. However, as can be seen from Fig. 5 of [7], the non-uniform quantizer can reduce the MSE (in comparison with the uniform quantizer) by at most 20%, if the number of bits is less than or equal to four (as is in the discussed case).

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

References

(1) R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, No. 4, pp. 299-309, Aug. 1977.

(2) J.L. Flanagan et al, "Speech Coding", IEEE Trans. Commun., Vol. COM-27, No. 4, pp. 710-736, April 1979.

(3) J.M. Tribolet and R.E. Crochiere, "Frequency Domain Coding of Speech", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, No. 5, pp. 512-530, Oct. 1979.

(4) Y. Ephraim and D. Malah, "Speech Enhancement Using an Optimal Non-Linear Spectral Amplitude Estimation", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Boston, pp. 1118-1121, 1983.

(5) J.J.Y. Huang and P.M. Schultheiss, "Block Quantization of Correlated Gaussian Random Variables", IEEE Trans. Commun. Syst., Vol. CS-11, pp. 289-296, Sept. 1963.

(6) M.J. Narasimha and A.M. Peterson, "On the Computation of the Discrete Cosine Transform", IEEE Trans. Commun., Vol. COM-16, pp. 934-936, June 1978.

(7) J. Max, "Quantizing for Minimal Distortion", IRE Trans. Inform. Theory, Vol. IT-6, pp. 7-12, March 1960.

(8) I.S. Gradshteyn and I.M. Ryzhik, Table of Integrals, Series, and Products, Academic Press Inc., 1980.

(9) L. Schuchman, "Dither Signals and their Effect on Quantization Noise", IEEE Trans. Commun. Tech., Vol. COM-12, pp. 162-165, Dec. 1964.

(10) J.S. Lim and A.V. Oppenheim, "Reduction of Quantization Noise in PCM Speech Coding", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, No. 1, pp. 107-110, Feb. 1980.

ENHANCEMENT OF NOISY SPEECH

Research Thesis

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Science

by

Yariv Ephraim

Submitted to the Senate of the Technion - Israel Institute of Technology

Sivan 5745 Haifa June 1984

© Technion - Israel Institute of Technology, Elyachar Central Library

This research was carried out under the supervision of Professor David Malah in the Signal Processing Laboratory of the Faculty of Electrical Engineering, Technion-Israel Institute of Technology.

Acknowledgement

I wish to express my sincerest thanks to Professor David Malah for introducing me to the speech enhancement problem, for his devoted guidance and invaluable help throughout all stages of this research.

I am also grateful to Professors Israel Bar-David, Moshe Zakai and Jacob Ziv for being always ready to advise me during the course of this work.

I am indebted to my friend Mr. Shlomo Shitz for the many interesting and fruitful discussions on filtering theory and thank him for his help.

Finally I would like to thank the Signal Processing Laboratory engineer Mr. Yoram Or-Chen and the programmer Ms. Zipora Portnoy for their help on various issues concerning the laboratory work.

© Technion - Israel Institute of Technology, Elyachar Central Library

© Technion - Israel Institute of Technology, Elyachar Central Library

Contents

			Page
Abstract			1
Notation List			2
Chapter	1 -	Introduction	4
	1.1 -	The nature of the problem	4
	1.2 -	Speech production and perception	7
	1.3 -	Previous works	9
	1.4 -	Screening of the work	12
	1.5 -	Contribution of the research	15
Chapter	2 -	Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator	19
Chapter	3 -	Combined Enhancement and Adaptive Transform Coding of Noisy Speech	28
Chapter	4 -	Discussion and Conclusions	32
Appendix	A -	Speech Enhancement Using a Minimum Mean Square Error Spectral Amplitude Estimator	36
	I -	Introduction	38
	II-	Optimal short-time spectral amplitude estimator	43
		-Derivation of amplitude estimator	43
		-Error analysis and sensitivity	49
	III-	Optimal amplitude estimator under uncertainty of signal presence	53
		-Derivation of amplitude estimator under signal presence uncertainty	56
	IV-	Optimal MMSE complex exponential estimator	
		-Derivation of complex exponential estimator	60
		-Optimal phase estimator	
	V -	Variance estimation of spectral components	64
		-Maximum likelihood estimation approach	64
		-Decision-directed estimation approach	66
	VI-	System description and performance evaluation	69
		-System description	70
		-Performance evaluation	71
	VII-	Summary and Discussion	75
		References	92
Appendix	B -	Speech Enhancement Using A Minimum Mean Square Error Log-Spectral Amplitude Estimator	94
	I -	Introduction	96
	II-	Derivation of optimal estimator	97
	III-	Derivation of optimal estimator under signal presence uncertainty	102

	IV-	System description and performance evaluation	103
		-System description	
		-Performance evaluation	104
	V -	Summary and Conclusions	106
		References	107
Appendix	C -	Signal to Noise Ratio Estimation For Enhancing Speech Using the Viterbi Algorithm	108
	I -	Introduction	109
	II-	MAP speech spectral component variance estimation	111
		-Map estimation	113
		-The Viterbi algorithm	118
	III-	System description and performance evaluation	120
		-System description	120
		-Performance evaluation	122
	IV-	Summary and Conclusions	124
		References	124
Appendix	D -	Combined Enhancement and Adaptive Transform Coding of Noisy Speech	125
	I -	Introduction	127
	II-	Adaptive transform coding	128
	III-	MMSE spectral amplitude estimator	133
	IV-	Performance evaluation	136
	V -	Discussion	137
		References	139

Abstract in English I

Contents in English III

© Technion - Israel Institute of Technology, Elyachar Central Library

Abstract

This research deals with the problem of enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available. The basic approach taken here capitalizes on the major importance of the short-time spectral amplitude (STSA) in speech perception. We develop minimum mean square error (MMSE) STSA estimators and examine them in enhancing speech signals. The estimators considered are a MMSE-STSA estimator and a MMSE log-STSA estimator. In addition, we extend these estimators to take into account the fact that the speech signal is not surely present in the noisy observations. These estimators are derived on the basis of a statistical model which utilizes asymptotic properties of the spectral components. Specifically we assume that the spectral components of the speech process, as well as of the noise process, can be modeled as statistically independent Gaussian random variables.

For constructing the enhanced signal, we examine the MMSE estimation of the complex exponential of the short-time phase. It is shown that the resulting estimator has a non-unity modulus, and therefore its combination with a MMSE STSA estimator, affects the STSA estimation. On the other hand, the MMSE complex exponential estimator which is constrained to have a unity modulus, and therefore does not affect the STSA estimation, is the complex exponential of the noisy phase. For the above reason, the noisy phase is utilized in the proposed system.

The problem of estimating the signal to noise ratio (SNR) of each spectral component, which is needed in the application of the above mentioned STSA estimators, is extensively investigated in this work. We propose maximum likelihood, decision-directed, and maximum a-posteriori estimators. The latter estimator is implemented by using the Viterbi-algorithm.

© Technion - Israel Institute of Technology, Elyachar Central Library

The proposed system was examined in enhancing speech degraded by wide-band noise with SNR of 5, 0, and -5dB. This system significantly improves the quality of the noisy speech, by suppressing the background noise level. The residual noise sounds colorless and was found to be much less annoying than the "musical noise" obtained in other commonly used systems. The complexity of the proposed system is similar to that of other currently used systems.

The above method of enhancing speech was also successfully applied to improve the quality of the reconstruced speech, obtained by using an adaptive transform coder whose input is noisy speech. In this application the STSA of the speech signal is optimally estimated, before the encoding is done.

© Technion - Israel Institute of Technology, Elyachar Central Library