תכן מערכי מסננים ספרתיים בעלי זכרון סופי

חבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר

דוקטור למדעים

מאת

אמיר דמבו

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

ניסן תשמ"ו חיפה מאי 1986

תכן מערכי מסננים ספרתיים בעלי זכרון סופי

חבור על מחקר
לשם מילוי חלקי של הדרישות לקבלת תואר
דוקטור למדעים

מאת

אמיר דמבו

הספריה המרכזית ע"ש אלישר
מספר מערכת

הוגש לסנט הטכניון - מכון טכנולוגי לישראל
ניסן תשמ"ו חיפה מאי 1986

203110
000000208437
2 00/C

מסנן FIR ספרתי

תכנון מערכי מסננים

מערכי מסננים אחידים אופטימליים

כימוי (קונטיזציה)

עבוד ספרתי של אותות

621.372.54.037.37:621.391:621.3.037.37

BC208437

המחקר נעשה בהנחיית פרופ' דויד מלאך בפקולטה להנדסת חשמל

אני מודה לקרן גוטווירט וליחידת המדען הראשי במשרד התקשורת על התמיכה הנדיבה בהשתלמותי.

תודתי נתונה לפרופ' דויד מלאך על הנחייתו המסורה, למר יורם אור-חן על עידודו בכל שלבי העבודה, ולמר צליל סלע ודוקטור משה דובינר על הדיונים המועילים.

לדפנה ואדר היקרים

תוכן העניינים

	עמוד
תקציר	1
רשימת סמלים וקיצורים	4
פרק 1 : מבוא	5
1.1 תאור הנושא וחשיבותו	5
1.2 מטרת ומבנה העבודה	12
פרק 2 : סקר תקורות ספרות	14
2.1 שיטות לתכנון מסנן FIR ספרתי יחיד	14
2.2 שיטות לתכנון מערכי מסננים	17
2.3 שיטות לתכנון מערכות אנליזה / סינתזה	20
פרק 3 : תכנון מערכי מסננים עם תגובה כוללת מוכתבת	28
3.1 תכנון מערכים אופטימליים בקריטריון WMMSE	28
3.2 קיום, יחידות ותכונות של המערכים האופטימליים בקריטריונים שונים	37
פרק 4 : שיטות לתכנון מערכי מסננים אחידים	42
4.1 תכונות כלליות של מערכי מסננים אחידים אופטימליים	42
4.2 תכנון מערכי מסננים אחידים אופטימליים בקריטריון WMMSE	43
4.3 תכנון מערכי מסננים אחידים אופטימליים בקריטריון Min-Max	46
4.4 שיטת ה"חלון" המוכללת ושימוש ב"חלון" אופטימלי לתכנון המערכים	47
פרק 5 : תכנון מערכות אנליזה-סינתזה אופטימליות הכוללות כימות (קונטיזציה)	53
5.1 תאור המודל הסטטיסטי ומדדי השגיאה	53
5.2 מסנני סינתזה אופטימליים עבור כימות עדין, ועבור כימות באמצעות ספרי-קוד	58
5.3 מערכות אנליזה-סינתזה אופטימליות עבור כימות עדין	61
פרק 6 : מערכי מסננים לסינתזה אופטימלית של אותות לאחר אנליזה ומודיפיקציה	64
6.1 תנאים לקיום מערכות אנליזה-סינתזה שהן מערכות יחידה ללא מודיפיקציה	64
6.2 סינתזה אופטימלית בקריטריון WMMSE עבור אות סופי	65
6.3 הסינתזה האופטימלית עבור אות אין-סופי (פתרון במצב מתמיד)	67
6.4 תנאים לקיום מערכות יחידה המכילות מודיפיקציה לינארית.	70

תוכן העניינים (המשך)

עמוד

פרק 7 : סיכום ותאור בעיות פתוחות 73

7.1 סיכום 73

7.2 תאור בעיות פתוחות 75

רשימת מקורות ספרות 76

נספחים 83

נספח א' : תכנון מערכי מסננים אופטימליים בקריטריון WMMSE עם תגובה כוללת מוכתבת 83

נספח ב' : קיום, יחידות ותכונות כלליות של מחלקה של בעיות אפרוקסימציה וקטורית. 111

נספח ג' : תכנון מערכי מסננים אחידים אופטימליים עם תגובה כוללת מוכתבת 150

נספח ד' : תכנון סטטיסטי של מערכות אנליזה-סינתזה עם כימות 173

נספח ה' : סינתזה אופטימלית של טרנספורם לזמן-קצר שעבר מודיפיקציה 217

תקציר

מערכי מסננים ספרתיים בעלי תגובה לדגם יחידה סופית (FIR) נפוצים מאד בעיבוד אותות ספרתי ובמיוחד בשימושי עיבוד אותות דיבור. תכנון מסננים אלה הוא אחד השטחים המרכזיים בתחום של עיבוד ספרתי ועבודות רבות נעשו בתחום זה בעשור האחרון. המחקרים הראשונים התרכזו בפתרון הבעיה הקלסית של תכנון מסנן FIR יחיד שיקרב תגובת תדר רצויה. תחילה הוצעו שיטות תת-אופטימליות שיתרונן בפשטותן הנומרית ואחר כך הוצגו שיטות מתוחכמות יותר שאיפשרו פתרון אופטימלי תחת נורמה מתאימה במרחב תגובות התדר.

לאחר פתרון הבעיה ה"קלסית" התפתח התחום של אנליזה וסינתזה של אותות דיבור וחלק גדול מהמערכות לסינון, קידוד וזהוי דיבור מכילות בתוכן מערכי מסננים המאפשרים אנליזת פורייה לזמן קצר. הופעת מערכות אלו, דרשה כלים לתכנון מערכי מסננים שמקיימים בנוסף לאופטימליות של כל מסנן ומסנן גם אילוץ משותף על המערך כולו. כאשר מערכי המסננים משמשים לאנליזה של האות שאין לאחריה סינתזה שלו (למשל: בזיהוי דיבור, זיהוי דובר, מיצוי פרמטרים בפסי תדר נפרדים למטרות קידוד), האילוץ המקובל הוא אילוץ של תגובת יחידה (קרי סכום תגובות המסננים שווה ל-1 בכל תדר ותדר). מאידך במערכות של אנליזה המלווה בסינתזה (בהן קיימים שני מערכי מסננים שונים) מקובל לדרוש אילוץ של מערכת יחידה. אילוץ זה מבטיח שללא מודיפיקציה, האות במוצא מערך מסנני הסינתזה זהה לאות בכניסת מערך מסנני האנליזה (עד כדי השהייה קבועה), ולכן ההפרדה לרכיבי תדר של האות לא יצרה עוות באות. קיים כיום ענין רב במערכי מסננים העונים לדרישות אלה. עם זאת עקב הקושי האנליטי והנומרי בפתרון בעיות התכנון שתוארו לעיל קיימות כיום תשובות חלקיות בלבד (תת-אופטימליות) למרביתן.

בעבודת מחקר זו מתוארים פתרונות אפטימליים מקוריים לבעיות התכנון של מערכי מסנני FIR המשמשים לאנליזה וסינתזה של אותות תחת אילוצים משותפים שונים. התוצאות העיקריות שהושגו בעבודה הן:

1. עבור מערכי מסננים המשמשים לאנליזה מתוארת שיטת התכנון של מערך מסננים אופטימלי בנורמת L_2, תחת אילוץ תגובה כוללת המוכתבת על ידי הסטייה המותרת מהתגובה האידיאלית. לנורמת L_2 יש יתרונות מתמטיים המקלים בפתרון אנליטי של הבעיה, וכן קיימת אינטרפרטציה סטטיסטית של שיטת התכנון הקושרת אותה לתכנון אופטימלי במובן של מינימום וריאנס השגיאה במוצא המערך. שיטת התכנון המתוארת היא כללית ביותר ומאפשרת אילוץ תגובה כוללת כלשהי (לאו דוקא תגובת יחידה), כאשר מערך המסננים אינו בהכרח מורכב ממסנני FIR, כי אם מקומבינציה לינארית כלשהי של אבני בניין נתונות. לכך יש יתרון גדול משיקולי ממוש יעיל בחומרה של מערך המסננים.

2. תכן מערכי מסננים כלליים, אופטימליים בנורמות אחרות (כגון: L_∞), בעלי אילוץ תגובה כוללת, מציב קשיים מתמטיים ניכרים. בעית תכנון זו מובילה לבעית אפרוקסימציה שלה לא קיים לעת עתה פתרון סגור. כצעד ראשון לקראת פתרון הבעיה נוסחו תנאים מספיקים לקיום פתרון אופטימלי וליחידותו. בנוסף הוכחו מספר תכונות כלליות של הפתרון האופטימלי, כגון: ממשיות מקדמי המסננים, תנאים לפזה לינארית של המסננים, והקשר בין גודל הסטיה המותרת מהתגובה הכוללת לטיב הקרוב של תגובות התדר של המסננים הבודדים.

3. המקרה הפרטי של מערכי מסננים אחידים (בהם אורך כל המסננים זהה וכן רוחב פס המעבר שלהם זהה) הוא בעל חשיבות רבה בעיבוד אותות עקב אפשרויות המימוש היעיל שלו בעזרת FFT. עבור מקרה זה מוכח שהפתרון האופטימלי נגזר על ידי הזזות בתדר של מסנן אב-טיפוס ממשי אופטימלי ואלגוריתם התכנון הכללי מפושט ומופעל ישירות על האב-טיפוס. יתר על כן מפותחת שיטה לתכנון מערך מסננים אחיד אופטימלי בנורמת L_∞, בעזרת שימוש בתכנות לינארי. פתרון תת-אופטימלי, הקרוב מאד לפתרון האופטימלי מוצג בהתבסס על הכללות של שיטת ה"חלון" לתכנון מסנני FIR.

4. במערכת אנליזה-סינתזה הכוללת כימות (קוונטיזציה) לא ניתן לקבל מערכת יחידה, כי הכימות הוא פעולה לא-הפיכה. מערכת אנליזה-סינתזה היא מערכת משתנה בזמן ולכן מוצא המערכת אינו תהליך אקראי סטציונרי במובן הרחב. בעבודה מתואר פיתוח של קריטריון שגיאה סטטיסטי לאותות לא-סטציונריים, המשמש להגדרת המערכים האופטימליים. שני מודלים שונים של כימות נתונים בעבודה - כימות עדין הממודל על ידי רעש-אדיטיבי, וכימות על ידי ספר-קוד. עבור שניהם מוצגת שיטה לתכנון האב-טיפוס האופטימלי של מערך מסנני הסינתזה. במקרה של כימות עדין שבו תלות אות המוצא במסנני האנליזה היא בקירוב לינארית מוצגת גם שיטה לתכנון מערכות אנליזה-סינתזה אופטימליות. מוכח בעבודה שכשרמת הרעש הממדל את הכימות שואפת לאפס מערכות האנליזה-סינתזה האופטימליות מתכנסות למערכות יחידה, ובמספר דוגמאות פרטיות חשובות ניתנת אינטרפרטציה של מסנני-הסינתזה כמסנני **Wiener**.

5. עבור מערכת אנליזה-סינתזה כללית (לאו-דוקא כזו המשמשת לקידוד), הורחב הדיון למערכות המבוססות על טרנספורם לינארי רגולרי לזמן-קצר (לאו דוקא אנליזת פוריה לזמן קצר) ומוצגים התנאים לקיום מערכת יחידה. בעית הסינתזה האופטימלית בנורמת L_2, של אות נתון לאחר מודיפיקציה (לא-ידועה) ואנליזה, נפתרה הן עבור אות סופי והן עבור אות אינסופי (קרי פתרון במצב-מתמיד), עבור טרנספורם כללי. תוצאות אלו מכלילות מספר מקרים פרטיים שנפתרו קודם לכן והן בעלות חשיבות רבה בשימושי סינון דיבור ושינוי ציר הזמן שלו.

בנוסף לכך מאופיינת מחלקת המודיפיקציות הלינאריות שעבורן קיימת סינתזה שמאפשרת לאתר אנליזה שניה, מודיפיקציה הפוכה וסינתזה שניה, שחזור האות המקורי ללא שגיאה. לאיפיון זה חשיבות רבה בין השאר לתכנון מערכות הסתרת דיבור.

נותרו עדיין מספר בעיות פתוחות בנושאים אלו המתוארות בפרק הסיכום של העבודה.

רשימת סמלים וקיצורים

FIR - Finite Impulse Response - תגובה לדגם יחידה סופית

LPF - Lowpass Filter - מסנן מעביר נמוכים

WOLA - Weighted Overlapp Add

DFT - Discrete Fourier Transform

DCT - Discrete Cosine Transform

FFT - Fast Fourier Transform -שיטה מהירה לממוש DFT.

WMMSE - Weighted Minimum Mean Square Error - קריטריון שגיאה-רבועית משוקללת.

QMF - Quadrature Mirror Filters

STFT - Short Time Fourier Transform - התמרת פורייה לזמן קצר.

DSTFT - Discrete Short-Time Fourier Transform-דגימות בתדר של התמרת פורייה לזמן קצר.

FBS - Filter Bank Summation

OLA - Overlapp and Add

MMSE - Minimum Mean Square Error - שגיאה רבועית ממוצעת מינימלית

IIR - Infinite Impulse Response - תגובה לדגם יחידה אינסופית

SNR - Signal to Noise Ratio - יחס אות לרעש

PTFB - Prototype Translated Filter Bank

CUFB - Complex Uniform Filter Bank - מערך מסננים אחיד קומפלקסי

RUFB - Real Uniform Filter Bank - מערך מסננים אחיד ממשי

AOW - Approximate Optimal Window - חלון, אופטימלי מקורב

A/S - Analysis/Synthesis - אנליזה סינתזה

FQ - Fine Quantization - כימות עדין

MQ - Matrix Quantization - כימות מטריצי

DSTT - Discrete Short Time Transform -דגימות בתדר של התמרה (כלשהי) לזמן קצר

PCM - Pulse Code Modulation

DPCM - Differential Pulse Code Modulation

DSTV - Discrete Short Time Vectors

MDSTV - Modified Discrete Short Time Vectors

MDSTT - Modified Discrete Short Time Transform

US - Unity System - מערכת יחידה

aUS - Almost Unity System - כמעט מערכת יחידה

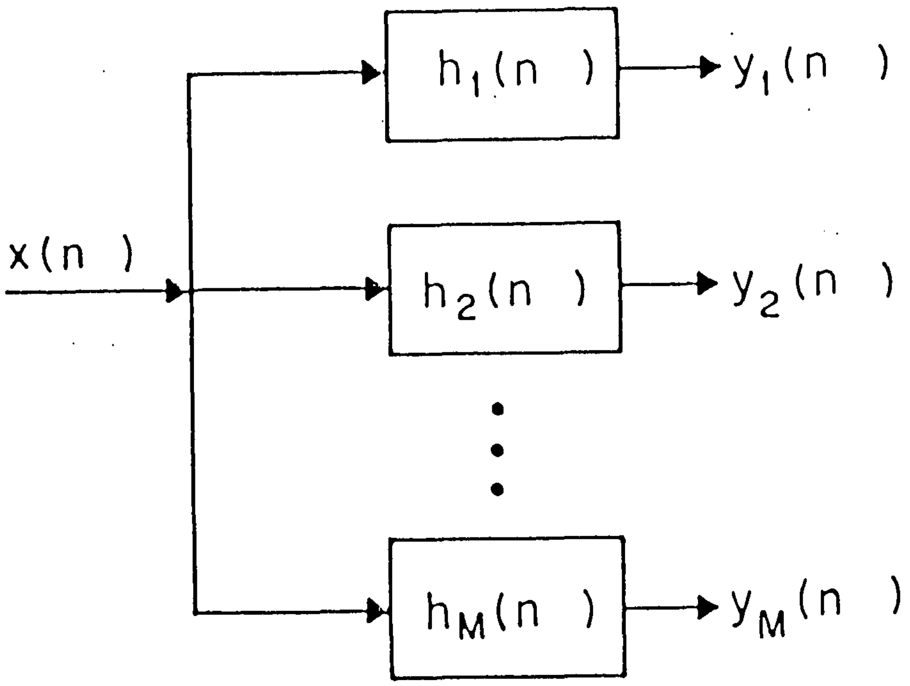
ftUS - Finite Time Unity System - מערכת יחידה בזמן סופי

DUS - Dual Unity System - מערכת יחידה דואלית

פרק 1 : מבוא

1.1 תאור הנושא וחשיבותו

מערך מסננים ספרתי מתואר סכמתית בציור מס' 1.1. הכניסה למערך היא אות ספרתי (סדרת דגימות) וכל אחד מהמסננים במערך, מסנן (על ידי קונוולוציה דיסקרטית) חלק מרכיבי התדר של האות. לפיכך, במוצא המערך מתקבלת סדרת וקטורים המאפיינת את הפילוג התדירותי לזמן קצר של האות בכניסה.



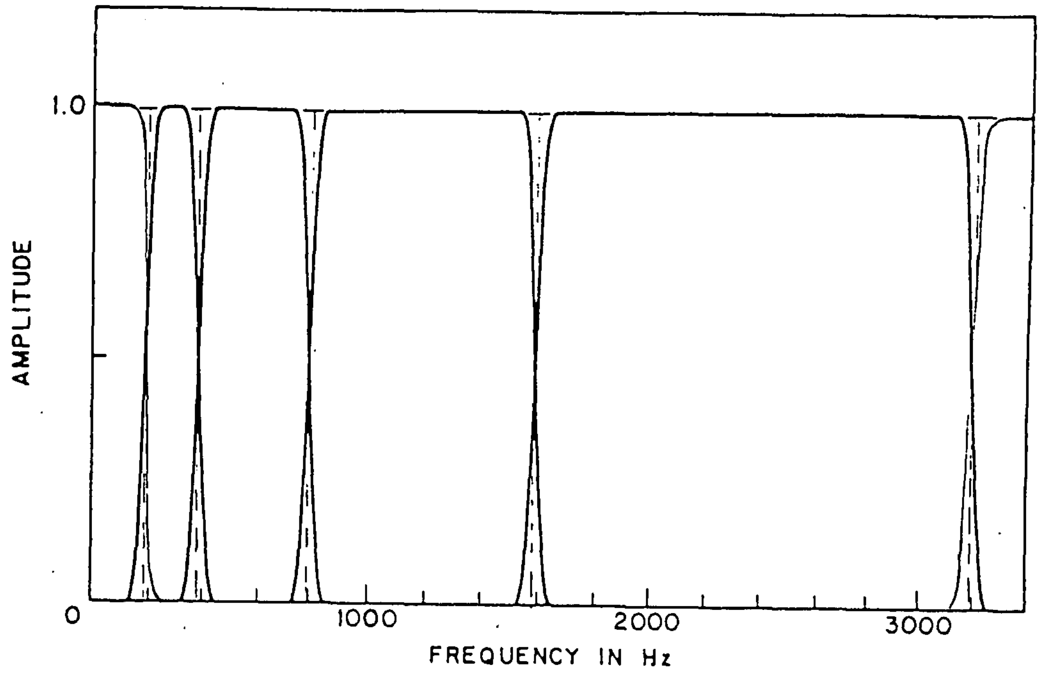
ציור מס' 1.1: מערך מסננים ספרתי (תאור סכמתי).

Fig. 1.1: Schematic Description of a Digital Filter Bank.

מערכים כאלו שימושיים מאד בעיבוד אותות ספרתיים, ובפרט בעיבוד אותות דיבור. מקובל להשתמש במוצא המערך כוקטור פרמטרים למערכות זיהוי דיבור [1], וככניסה למערכות סינון [2], ערבול [3] וקידוד [4] דיבור. מקובל לדרוש שסכום היציאות של מסנני המערך זהה לאות הכניסה, וזאת ע"מ להבטיח שכל רכיבי התדר של האות יופיעו במוצא המערך. דרישה זו מכתיבה במיוחד אילוצים על צורת תגובת התדר של המסננים בתחום המעבר (Transition-Band), שבו בדרך כלל לא ניתנות ספציפיקציות. דרישה זו היא דרישה לתגובה כוללת שהיא תגובת יחידה ומשמעה שסכום תגובות המסננים שווה ל-1 בכל תדר ותדר.

במקרים רבים האות בכניסה נדגם במקורו בתדר גבוה מתדר Nyquist ולכן ישנם רכיבי תדר שלו (התדרים העליונים) שידוע א-פריורי שהם אפס. כפועל יוצא ניתן לוותר על חלק מהמסננים במערך ולדרוש תגובה כוללת שאינה תגובת יחידה, כי אם תגובת מסנן מעביר נמוכים (LPF) רחב סרט.

על מנת שכל רכיבי התדר במוצא המערך יתייחסו לאותה נקודת זמן באות הכניסה, נדרש שכל המסננים במערך יהיו בעלי פזה לינארית (ובשיפוע זהה בכולם). דרישה זו מאלצת את המסננים במערך להיות בעלי תגובת דגם יחידה סופית (FIR). לפיכך לא ניתן לקבל הפרדת תדר אידיאלית, כי אם רק קרוב שלה (כמתואר סכמתית בציור מס' 1.2).

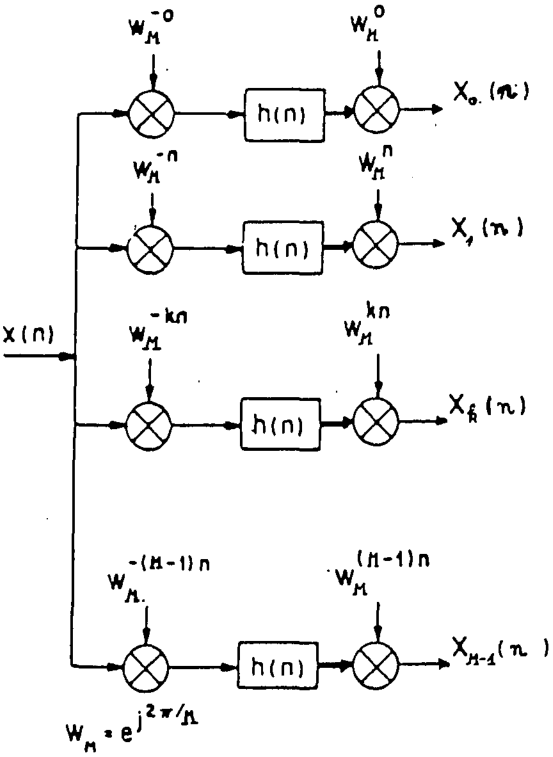


ציור מס' 1.2: הפרדת תדר אידיאלית וקרוב שלה.

Fig. 1.2: Ideal Frequency Response and its Approximation.

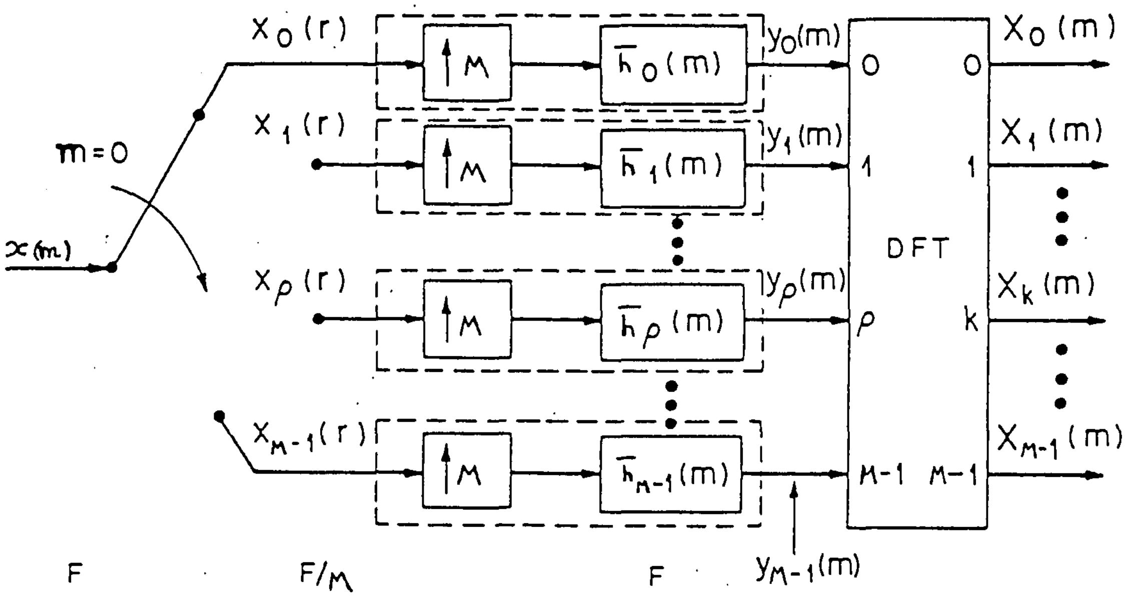
הקשר שבין מערך מסנני FIR ספרתיים, לבין אנליזת פורייה לזמן קצר מתואר בפרוט ב-[5], כמו גם נושאים נוספים הקשורים לשימוש ותכנון מערכי מסננים אלו.

מערכי מסננים אחידים הם תת-מחלקה חשובה של מערכי מסננים ספרתיים שבהם אורך כל מסנן זהה וכן רוחב פס-ההעברה (Passband) שלהם זהה. למרות שמבחינה פיסיקלית אין סיבה להעדיף השימוש במערכים אלו לצורך מיצוי פרמטרים של אות דיבור, הרי יש לכך סיבות חישוביות. כפי שמוראה בפרוט רב ב-[6], ניתן לממש מערכי מסננים אלו ביעילות רבה תוך ניצול ה-DFT (FFT), ובלבד שתגובות התדר של המסננים במערך נגזרות מתוך הזזות בתדר של מסנן אב-טיפוס (בדרך כלל זהו LPF). בציור 1.3 מתואר מערך מסננים אחיד, הנגזר ממסנן אב-טיפוס ובציור 1.4 מתואר מימוש יעיל שלו בעזרת DFT.



<u>ציור 1.3</u>: מערך מסננים אחיד הנגזר מהזזות של אב-טיפוס.

<u>Fig. 1.3</u>: **Uniform Filter Bank, Obtained by Frequency Translations of a Prototype Filter.**



<u>ציור מס' 1.4</u>: מימוש יעיל של המערך בציור 1.3, על ידי DFT.

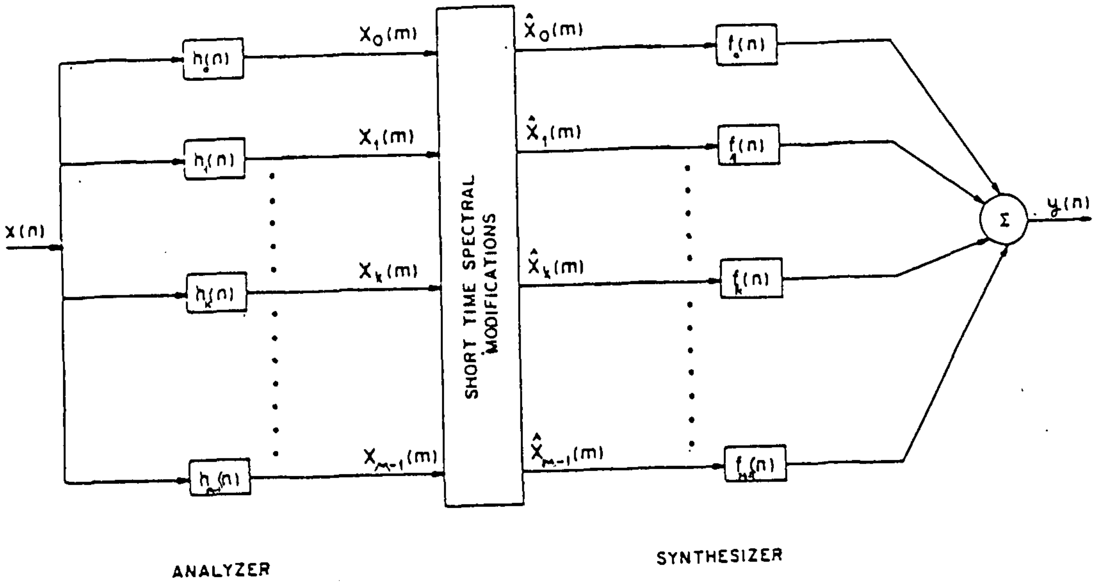
<u>Fig. 1.4</u>: Efficient Implementation of the Filter Bank in Fig. 1.3, using DFT.

מקדמי מסנן האב-טיפוס, פועלים במימוש היעיל כמקדמי "חלון" המופעל על אות-הכניסה קודם לביצוע ה-FFT ולכן משתמשים לפעמים במינוח "חלון" אנליזה במקום מסנן אנליזה.

לאור החשיבות הרבה של מערכים כאלו יש ערך גם בשיטות תכנון שהן ייחודיות עבורן, וכן קיימת חשיבות רבה לתשובה על השאלה מתי למערך המסננים האחיד <u>האופטימלי</u> ישנה התכונה שהוא מורכב מהזזות תדר של מסנן אב-טיפוס. על מנת להדגים את השימוש הנרחב במערכי מסננים אחידים נעיר שהם מופיעים באפליקציות תקשורתיות כגון: מעבר מריבוב זמן (TDM) לריבוב תדר (FDM) [7], באפליקציות שבהן מעורב שינוי בתדר הדגימה [8], ובאפליקציות של זיהוי [9] ודחיסת [10] דיבור.

במערכות לקידוד וסינון דיבור מבוצעת מודיפיקציה על סדרת הוקטורים שבמוצא מערך מסנני האנליזה ויש "לסנתז" אות זמני מתוך סדרת הוקטורים לאחר המודיפיקציה. הסינתזה נעשית על ידי מערך שני של מסננים.

כל אחד מהמסננים מוזן על ידי סדרה שונה של אברים מהוקטורים הנ"ל. מסנני הסינתזה מסלקים (על ידי קונוולוציה דיסקרטית) רכיבי תדר לא רצויים שמופיעים בסדרות הכניסה שלהם עקב המודיפיקציה. סיכום מוצאי מסנני הסינתזה יוצר אות זמני שהוא האות במוצא מערכת האנליזה - סינתזה. ציור 1.5 מתאר סכמתית מערכת אנליזה - סינתזה כזו, תאור מפורט יותר ניתן למצוא ב-[6].



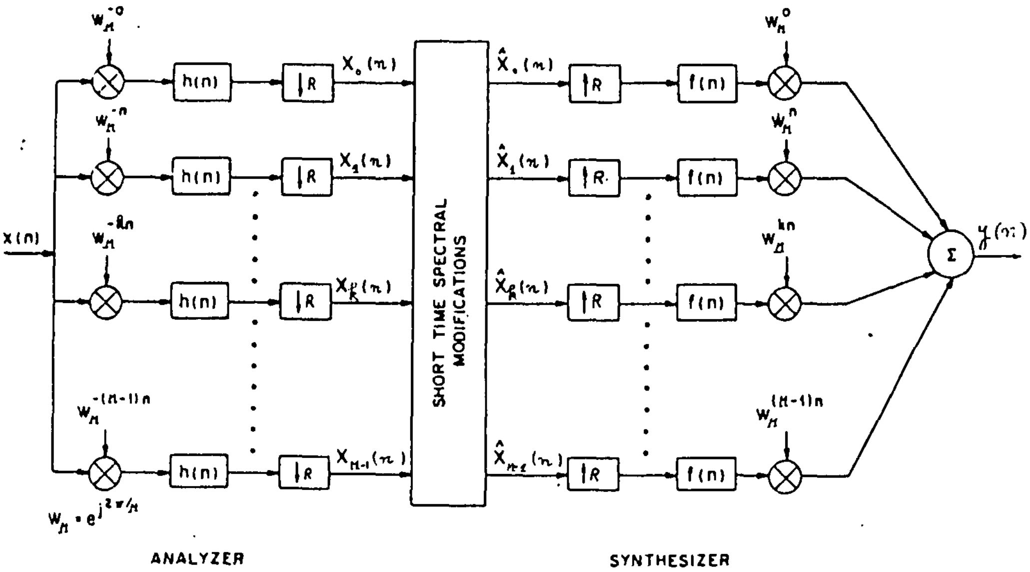
<u>ציור מס' 1.5</u>: מערכת אנליזה - סינתזה עם מודיפיקציה, ושני מערכי מסננים כלליים ללא דצימציה / אינטרפולציה.

<u>Fig. 1.5</u>: Analysis / Synthesis System including Modification and Two General Filter Banks without Decimation / Interpolation.

משיקולי סיבוכיות רצוי במקרים רבים לדגום את מוצא מערך מסנני האנליזה בקצב נמוך יותר על מנת לחסוך בפעולות.

בשימושים של קידוד, בהם מוצא מערך מסנני האנליזה משודר לאחר כימותו יש כמובן לדגום אותו בקצב המינימלי האפשרי, וזאת לא רק משיקולי סיבוכיות, כי אם על מנת לחסוך בקצב השידור וברוחב הסרט בערוץ. מסיבות אלו מקובל לדגום את מוצא מסנני האנליזה בקצב נמוך מקצב הדגימה המקורי. יתר על כן, נפוץ מאד השימוש במערכי מסנני אנליזה וסינתזה אחידים שעבורם ניתן להשתמש בקצב דגימה אחיד עבור מוצאי כל המסננים, קרי לבצע הורדת קצב דגימה (דצימציה) ביחס קבוע בכל המערכת.

בציור 1.6 מתוארת מערכת כזו שבה M מסננים אחידים, שמוצאם עובר דצימציה ביחס R:1, מודיפיקציה ואחר כך העלאת הקצב ביחס 1:R קודם לכניסה למערך M מסנני סינתזה אחידים המשמשים גם כמסנני אינטרפולציה.



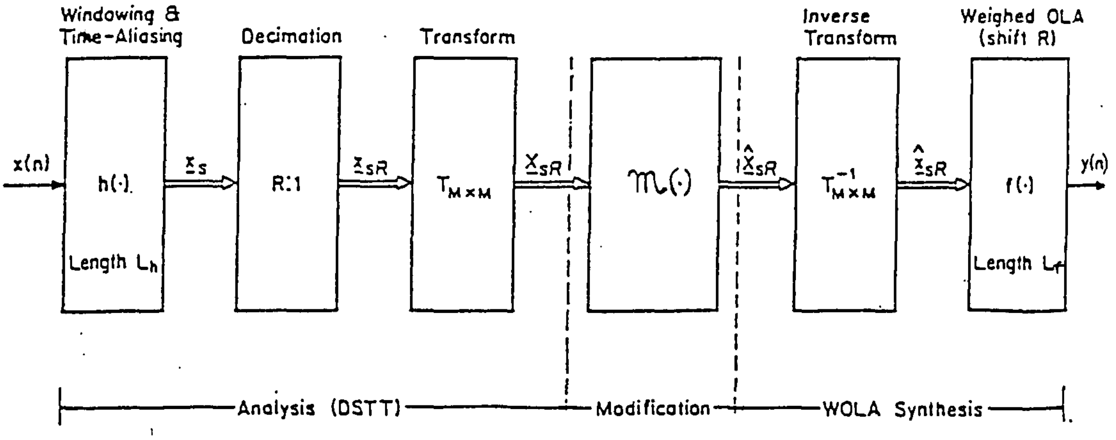
<u>ציור מס' 1.6</u>: מערכת כבציור 1.5, עם דצימציה / אינטרפולציה ומערכי מסננים אחידים.

<u>Fig. 1.6</u>: System as in Fig. 1.5, with Decimation / Interpolation and Uniform Filter Banks.

מערכות אלו שימושיות ביותר בעבוד אותות ספרתיים, ומתוארות בפרוט ב-[6]. דוגמאות לשימוש במערכות כאלו ניתן למצוא ב-[3,4,10,11,12].

בנוסף לפשטות המושגת על ידי שימוש ביחס דצימציה קבוע, קיים ממוש יעיל של מערכות אלו בעזרת FFT, בדומה לממוש היעיל של מסנני אנליזה שהוזכר כבר. הממוש הנ"ל מוצג בהרחבה ב-[6, Chapter 7], וכאשר מנצלים ממוש זה מתקבלת מערכת סינתזה הידועה בכינוי WOLA (Weighted Overlapp - Add).

בציור 1.7 מתוארת מערכת אנליזה - סינתזה בשיטת WOLA [13], כפי שמתקבלת על ידי שימוש ב-FFT.



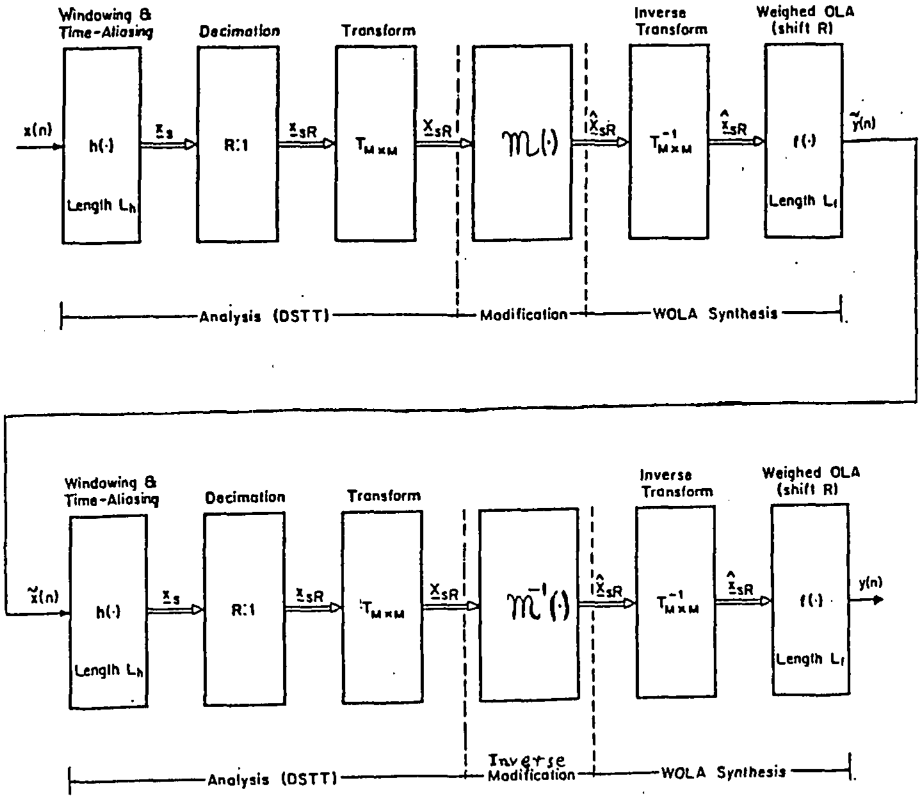
ציור מס' 1.7: מערכת אנליזה-סינתזה עם טרנספורם לינארי, מודיפיקציה, וסינתזה בשיטת WOLA.

Fig. 1.7: Analysis/Synthesis System Based on Arbitrary, Linear Transform, Modification, and WOLA Synthesis.

המערכת המתוארת בציור זה היא מערכת כללית יותר שבה ישנו טרנספורם לינארי רגולרי כלשהו המחליף את ה-DFT. בשימושי קידוד מקובל מאד להחליף את ה-DFT ב-DCT [14] או בהתמרת Hadamard, ולכן הסכמה המתוארת בציור 1.7, מאפשרת טיפול אחיד בכל המקרים הללו.

בעוד שביישומי קידוד וסינון האות במוצא מערכת הסינתזה הוא האות הסופי הרי שביישומים של שינוי ציר הזמן של אות הדיבור [10] וערבול [3], מוצא מערכת הסינתזה הוא האות המשודר בערוץ. במקלט ישנה מערכת אנליזה-סינתזה שניה המכילה בתוכה מודיפיקציה הפוכה והיא שמשחזרת את האות המקורי הרצוי.

ניתוח מערכת כזו מחייב לכן לנתח את בצועי שתי מערכות אנליזה-סינתזה המחוברות בטור זו לזו. ציור 1.8 מתאר את המערכת הכוללת במקרה זה.



ציור 1.8: שתי מערכות בשיטת WOLA, האחת מכילה מודיפיקציה והשניה את המודיפיקציה ההפוכה.

Fig. 1.8: Two Systems with WOLA Synthesis; The First Contain Certain Modification, whereas the Second Contain the Inverse Modification.

1.2 מטרת ומבנה העבודה

מטרת העבודה היא לפתח שיטות לתכנון אופטימלי של מערכי מסננים לצורך אנליזה וסינתזה של אותות ספרתיים, עם דגש על שימושים לאותות דיבור. ניתן לחלק את העבודה באופן כללי לשני חלקים עיקריים:

א. תכנון מערכי-מסננים לצורך אנליזת אותות ספרתיים כשמוכתבת התגובה הכוללת של המערך, נושא זה נדון בפרקים 3 ו-4 של העבודה.

פרק 3 עוסק בתכנון מערכי מסננים עם תגובה כוללת מוכתבת, בעוד שפרק 4 מתמקד בשיטות לתכנון מערכי מסננים אחידים הנפוצים מאד בשימושים השונים, כפי שתואר לעיל. עבור מקרה זה מפותחות שיטות תכנון מיוחדות שאינן ישימות במקרה הכללי.

ב. תכנון מערכי-מסננים במערכות אנליזה וסינתזה הכוללות מודיפיקציה של האות. בנושא זה הדגש הוא על תכנון מערכי מסנני סינתזה המביאים לשחזור אופטימלי של האות.

פרקים 5 ו-6 של העבודה עוסקים בסוגיה זו. פרק 5 דן בגישה סטטיסטית לתכנון מערכות אנליזה-סינתזה אופטימליות הכוללות כימות, ואילו פרק 6 דן בבעיה הכללית יותר של סינתזה אופטימלית של אותות לאחר אנליזה ומודיפיקציה.

השיטות המתוארות בפרק 5 מנצלות את האפיון הסטטיסטי של המקודדים השונים על מנת להשיג תוצאות שאינן ישימות למודיפיקציה כללית (שאינה בעלת אפיון סטטיסטי ברור).

בפרק 2 נסקרים המקורות שעסקו בשתי בעיות תכנון אלה, ובפרק 7 ישנו סיכום התוצאות שהוצגו בעבודה ותאור הבעיות שנותרו פתוחות.

השיטות המתוארות בפרקים 5 - 3 הן במלואן פיתוח מקורי, ואילו הנושאים המתוארים בפרק 6 מהווים הרחבה של מקורות [17, 16, 15] הנסקרים בפרק 2.

פרק 2 : סקר מקורות ספרות

2.1 שיטות לתכנון מסנן FIR ספרתי יחיד

בסעיף זה נסקור בקצרה שלוש שיטות מקובלות לתכנון מסנן FIR ספרתי יחיד. למרות שתכנון מסנן יחיד אינו מנושאי עבודה זו, הרי שיטות התכנון הנ"ל יובילו לשיטות לתכנון מערכי מסננים עם תגובה כוללת שיפותחו בפרקים 3 ו-4.

השיטה הראשונה מכונה שיטת ה"חלון" (window-method) והיא שיטה לא אופטימלית המבוססת על נימוקים היוריסטיים בלבד. השיטה השניה היא שיטת Min-Max שבה נבחר המסנן הקרוב ביותר לתגובת התדר הרצויה בקריטריון L_∞ עם פונקצית משקל מתאימה, והשיטה השלישית המכונה WMMSE מבוססת על בחירת המסנן הקרוב ביותר לתגובת התדר הרצויה בקריטריון L_2 עם פונקצית משקל מתאימה.

שיטות אלה הן הנפוצות ביותר כשיטות לתכנון מסנני FIR ספרתיים.

<u>(א) שיטת ה"חלון"</u>

שיטה זו מתוארת בפרוט ב-[18], נתאר כאן את עיקרי השיטה, כדלקמן:

תהא $D(f)$ תגובת התדר הרצויה. נניח כי תגובת דגם היחידה האינסופית $d(n)$ הקשורה ל-$D(f)$ ניתנת לחישוב אנליטי (למשל: עבור LPF אידיאלי קל למצוא נוסחא סגורה ל-$d(n)$). תהא נתונה סדרת מקדמי "חלון" $w(n)$ סופית באורך L (הזהה לאורך הרצוי של מסנן ה-FIR). עתה, מסנן FIR שיתוכנן בשיטת ה"חלון" יהא בעל תגובת דגם יחידה $h(n) = d(n)w(n)$, ומאחר ו-$w(n)$ סופית הרי $h(n)$ היא בעלת משך סופי (FIR).

מתכונות התמרת פוריה קל לראות שתגובת התדר $H(f)$ של המסנן תהא קונוולוציה של התגובה הרצויה $D(f)$ עם תגובת התדר של סדרת מקדמי ה"חלון". "חלון" טוב יהא בעל תגובת תדר המהווה קירוב לפונקצית ההלם של דירק.

כמובן שדרישה זו סותרת את הדרישה ל"חלון" בעל תגובת דגם יחידה סופית, ונעשו מספר מחקרים על בחירת ה"חלון" ותכונותיו (למשל [19, 23, 24]). ה"חלון" המקובל ביותר כיום לתכנון מסנני FIR הוא "חלון" Kaiser (ראה [20]). זו למעשה משפחה של "חלונות" כאשר קיים פרמטר רציף המאפשר Trade-off, בין רוחב האונה הראשית של תגובת התדר של ה"חלון" לבין גובה אונות הצד שלה.

מרבית ה"חלונות" המקובלים בספרות נוצרים על ידי דגימת פונקציה רציפה וסופית $w_a(t)$ בדגימה אחידה, היוצרת L דגימות בסדרת מקדמי ה"חלון". לפיכך סדרות "חלון" עבור אורכים שונים של מסנני FIR מחושבות בקלות מתוך אותה נוסחא בסיסית המגדירה את $w_a(t)$. דרגות החופש בתכנון הן בחירת $w_a(t)$, בחירת אורך המסנן L ותגובת התדר הרצויה $D(f)$.

במרבית מסנני ה-FIR השימושיים תגובת התדר **D(f)** היא קבועה למקוטעין. במקרה זה תחום התדר $0. \le f \le 0.5$ מחולק למס' אינטרוולים שבהם הגבר המסנן הרצוי חיובי (אינטרוולים אלו קרויים תחומי העברה - **Passbands**), למס' אינטרוולים שבהם הגבר המסנן הרצוי הוא אפס (הקרויים תחומי הנחתה - **Stopbands**) ובין כל זוג אינטרוולים ישנו תחום בו **H(f)** אינו מוכתב (קרוי תחום מעבר - **Transition Band**).

בשיטת החלון המקובלת נקודת אי-הרציפות של **D(f)** נבחרת תמיד במרכז תחום המעבר [18], אולם בפיתוח מקורי שלנו שמתואר ב-[21], נמצא שניתן לשפר את ביצועי שיטת ה"חלון" על ידי קביעה אופטימלית של מקום נקודה זו בתוך תחום המעבר.

אנליזת ביצועי שיטת ה"חלון" המתוארת בפרוט רב ב-[22] וששימשה בסיס לשיטת קביעת המיקום האופטימלי של נקודת אי-הרציפות של **D(f)** , היא שהובילה לפיתוח ה"חלון האופטימלי" המשמש לתכנון מערכי מסננים בעבודה זו (סעיף 4.4).

(ב) <u>שיטת Min-Max</u>

שיטה זו מתוארת בפרוט רב ב-[18]. היא מבוססת על בחירת המסנן בעל תגובת דגם יחידה באורך L, שתגובת התדר שלו קרובה ביותר בנורמת L_∞ (משוקללת בתדר) לתגובת התדר הרצויה.

קריטריון שגיאה זה מקובל ביותר בתכנון מסננים מכיון שהוא מאפשר למשתמש להגדיר את דרישותיו באופן נוח ביותר (על ידי הגדרת ה-**tolerance** של תגובת התדר סביב התגובה האידיאלית).

מאחר ופתרון אנליטי לבעית קירוב זו אינו ידוע, מוצאים פתרון נומרי על ידי דיסקרטיזציה של תחום התדר ומציאת המסנן האופטימלי על ידי פתרון בעית תכנות לינארי מתאימה (ראה פרוט ב-[26, 25, 18]). על מנת לקבל פתרון הקרוב לאופטימום נדרשת דיסקרטיזציה עדינה של תחום התדר וכשאורך המסנן הרצוי (קרי מס' הנעלמים בבעית התכנות הלינארי) גדל נוצרות בעיות זמן חישוב במימוש שיטה זו.

קיימת דרך מהירה יותר לפתרון בעית הקירוב הנ"ל על ידי האלגוריתם של **Remez**. טכניקה זו הוצעה לראשונה ב-[27], ושופרה והואצה ב-[30, 29, 28], אולם היא שימושית <u>רק</u> כאשר אין כל אילוצים נוספים במישור הזמן או התדר על המסנן הדרוש. במקרה וקיימים אילוצים כאלה חייבים לחזור לשיטת התכנות הלינארי, האיטית יותר (ראה [18]).

(ג) <u>שיטת WMMSE</u>

שיטת WMMSE לתכנון מסנן FIR יחיד מבוססת למעשה על מסנן **Wiener** דיסקרטי [31]. השיטה מתוארת בפרוט ב-[32,33], כאשר במאמר הראשון מוצגת האינטרפרטציה הסטטיסטית שלה ובשני האינטרפרטציה הדטרמניסטית.

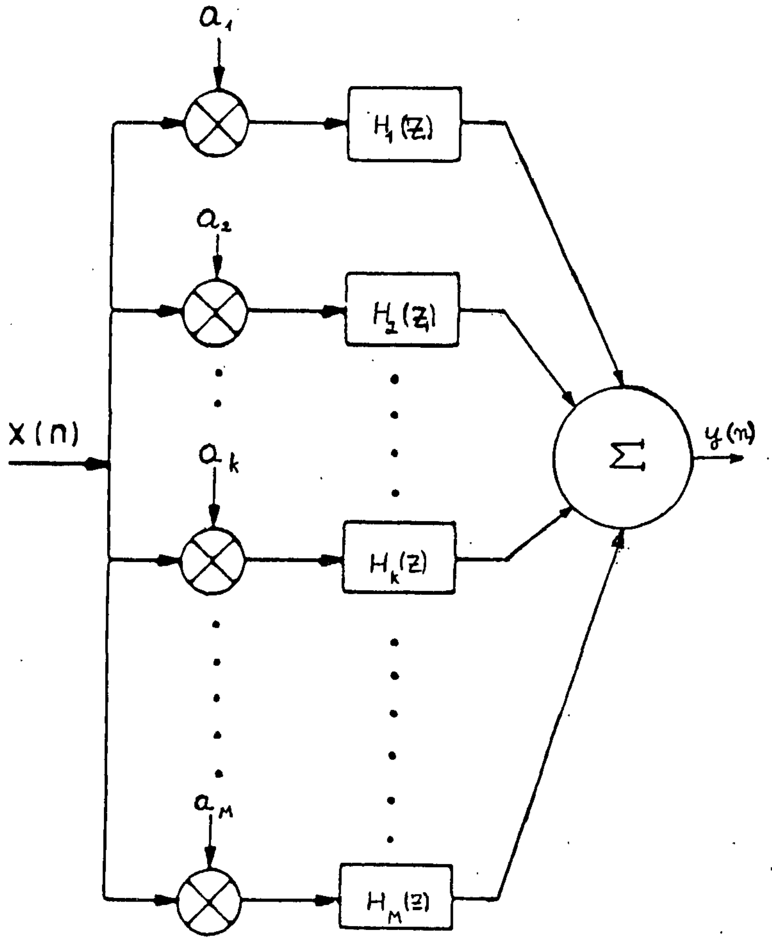
נתאר כאן בקצרה את עיקרי השיטה בגישה הדטרמניסטית:

נתונה תגובת תדר רצויה D(f) ופונקצית משקל של השגיאה W(f). נחפש מסנן FIR באורך נתון, שיביא למינימום את נורמת L_2 של הפונקציה W(f)(H(f)-D(f)), כאשר H(f) זו תגובת התדר של המסנן.

מאחר ונורמת L_2 היא תבנית ריבועית חיובית ממש (עבור W(f) שאינה אפס בתחום בעל מידה שאינה אפס), בנעלמים שהם אברי תגובת דגם יחידה של המסנן, הרי ניתן למצוא את המינימום הגלובלי (קרי - המסנן האופטימלי) על ידי פתרון מערכת משוואות לינאריות. משוואות אלה קרויות גם משוואות נורמליות. מאחר שהמטריצה המייצגת אותן היא מטריצת Toeplitz קיימות שיטות מהירות לפתרונן (למשל ב-[34]).

משיקולים מעשיים עדיף במקרים רבים להשתמש במסנן ספרתי המורכב משרשרת של חוליות בסיסיות (שאינן בהכרח יחידות השהייה), על פני מסנן FIR קלסי.

מקרה כזה מתואר בהרחבה ב-[35, 36]. ניתן אז להסב את שיטת WMMSE, כך שתאפשר מציאת המסנן האופטימלי (קרי - ערכים אופטימליים למכפלים החיצוניים שאינם חלק מהחוליות הבסיסיות, ראה למשל בציור מס' 2.1).



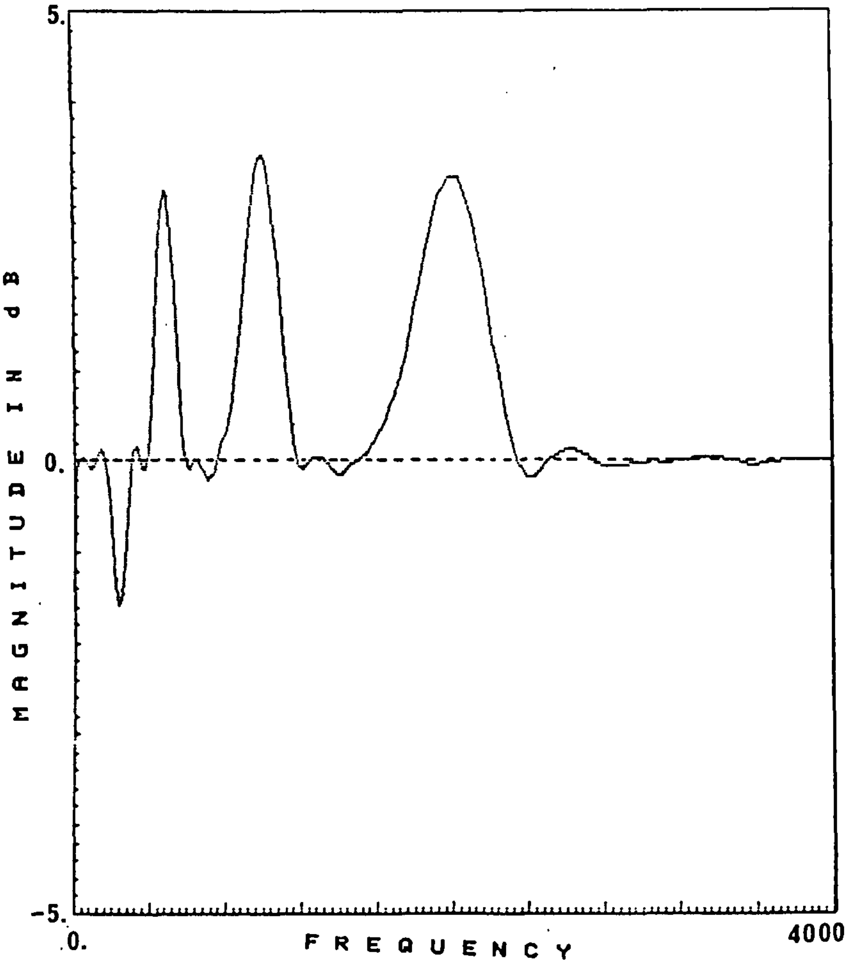
ציור מס' 2.1: מסנן ספרתי המורכב מחוליות כלליות (תאור סכמתי).

Fig. 2.1: Schematic Description of a Digital Filter with General Basic Elements.

שיטת WMMSE לתכנון מערכי מסננים אופטימליים עם תגובה כוללת מוכתבת שתוצג בהמשך, מפותחת כך שניתן להשתמש בה לתכנון מערכי מסננים המורכבים מחוליות בסיסיות כלשהן.

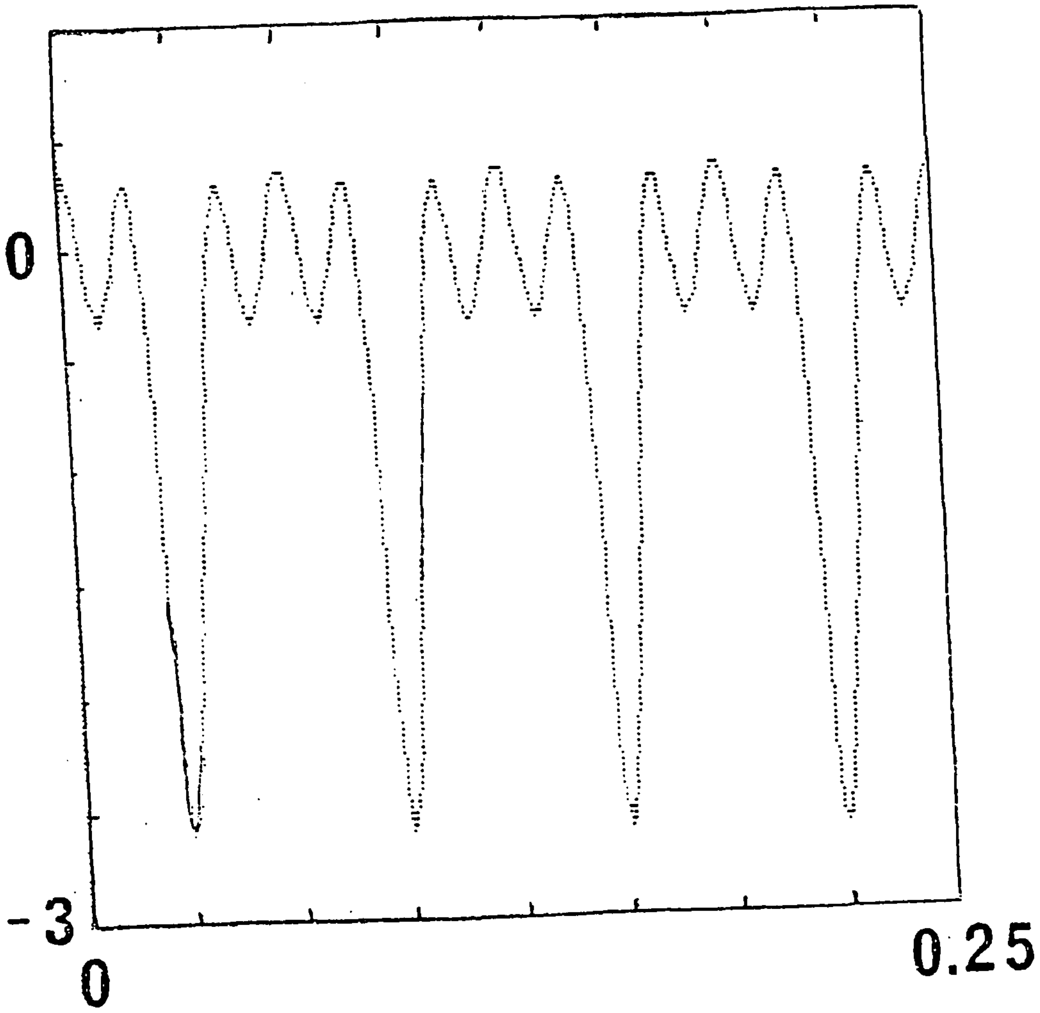
2.2 שיטות לתכנון מערכי מסננים

כמובן שניתן לתכנן מערכי מסננים על ידי כל אחת משלושת השיטות לתכנון מסנן יחיד שתוארו לעיל. אולם כיון שבמקרה זה מפעילים את שיטות התכנון באופן בלתי-תלוי עבור כל אחד מהמסננים המרכיבים את המערך, אין שליטה נוחה על התגובה הכוללת של המערך. ואכן, במספר דוגמאות טיפוסיות שנבדקו, הן בשיטת Min-Max והן בשיטת WMMSE, התקבלה תגובה כוללת גרועה כמודגם למשל בציורים 2.2 ו-2.3.



ציור מס' 2.2: תגובה כוללת של מערך מסננים טיפוסי שתוכנן בשיטת WMMSE המקובלת.

Fig. 2.2: Composite Frequency Response of a Typical Filter Bank Obtained by using the Conventional WMMSE Design Method.



ציור מס' 2.3: תגובה כוללת של מערך מסננים טיפוסי שתוכנן בשיטת Min-Max המקובלת.

Fig. 2.3: Composite Frequency Response of a Typical Filter Bank Obtained by Using the Conventional Min-Max Design Method.

בשיטת ה"חלון" ניתן להבטיח תגובה כוללת שהיא תגובת יחידה ובלבד שיתקיימו התנאים הבאים [5,37]:

(א) כל המסננים בעלי אורך זהה.

(ב) אותה סדרת מקדמי "חלון" משמשת לתכנון כל המסננים.

(ג) למערך המסננים האידיאלי תגובה כוללת שהיא תגובת יחידה (קרי סכום תגובות התדר הרצויות של המסננים הוא 1 בכל תדר ותדר).

הטענה דלעיל נובעת מהעובדה שפעולות החיבור והקונוולוציה הן פעולות מתחלפות. מסיבה זו שיטת ה"חלון" היא אחת השיטות השימושיות ביותר לתכנון מערכי מסנני FIR לאנליזת אותות דיבור.

החסרון העיקרי בשיטה זו הוא שהפרדת התדרים המתקבלת היא לא אופטימלית (ראה למשל: [22, 21, 18]). חסרון נוסף הוא המגבלה שארכי כל המסננים יהיו זהים.

לפיכך, בתכנון מערכי מסננים עם תגובה כוללת מוכתבת, נפוצה השיטה ההיוריסטית הבאה: כל מסנן מתוכנן בשיטת Min-Max באופן בלתי תלוי. במידה והתגובה הכוללת המתקבלת אינה טובה משנים מעט את הספציפיקציות של תחומי-התדר השונים או את פונקצית המשקל שבשימוש כך שמתקבל מערך-חדש. חוזרים על התהליך מספר פעמים עד שמתקבלת תוצאה סבירה (ראה למשל ב-[10]).

ב-[38] מוצעת אסטרטגיה לתהליך ממוחשב מצורה זו ומודגמים ביצועיו. הבעיה היא שתהליך זה (כמו זה הלא ממוחשב המקביל לו) סובל מנטיות להתבדרות, במיוחד כשמספר המסננים במערך גדול (ראה שם), וסיבוכיותו עשויה להיות עצומה כשמדובר במספר רב של מסננים באורך טיפוסי של מספר מאות מקדמים כל אחד.

במקרה הפרטי של מערכי מסננים אחידים, הרי עקב הרצון לממש ביעילות בעזרת FFT כמתואר ב-[6], הם מתוכננים מראש מתוך מסנן אב-טיפוס (בדרך כלל LPF). כפועל יוצא מכך אורכם של כל המסננים במערך זהה לארכו של מסנן האב-טיפוס, וניתן לקבל תגובה כוללת שהיא תגובת יחידה, ובלבד שמקדמי מסנן האב-טיפוס מקיימים מספר אילוצים המוכתבים על ידי משוואות לינאריות מסוימות במקדמי המסנן.
כאשר אורך המסנן הוא אי-זוגי, משוואות אלו הופכות לאילוצי איפוס כל דגימה M-ית בתגובת דגם יחידה שלו (כש-M הוא מימד ה-FFT), כמתואר ב-[6], וכן בנספח ב' של עבודה זו. לפיכך עבור המקרה הפרטי הנ"ל הוצעה ב-[39] השיטה התת-אופטימלית הבאה:
תכנן מסנן אב-טיפוס אופטימלי בשיטת Min-Max. המסנן המתקבל אינו יוצר מערך מסננים בעל תגובה כוללת שהיא תגובת יחידה, אך על ידי איפוס הדגימות המתאימות בתגובת דגם יחידה של מסנן האב-טיפוס ניתן לקבל מערך מסננים (תת-אופטימלי) שבו התגובה הכוללת היא תגובת יחידה. ב-[39] מפותחים חסמים על השגיאה המקסימלית בתדר הנוצרת על ידי תהליך איפוס הדגימות.

ב-[40] הוצע (בהקשר לבעית תכנון מסנני אינטרפולציה), אלגוריתם של תכנות-לינארי המאפשר תכנון מסנן אב-טיפוס אופטימלי (בנורמת L_∞), עם אילוצי אפסים על אברים מסוימים בתגובת דגם היחידה שלו. שיטה זו משווית עם זו שהוצעה ב-[39] ונמצאת טובה ממנה (ראה ב-[40]), אולם אלגוריתם התכנות הלינארי המוצע, מוגבל מבחינה מעשית לתכנון מסננים באורך של מאה מקדמים או פחות (ניתוח סיבוכיות מפורט מופיע ב-[26]).

טבלא מס' 2.1 מסכמת את היתרונות והחסרונות של ארבע השיטות לתכנון מערכי מסננים שנסקרו כאן (הכוונה לשיטות שהופיעו ב-[40, 39, 38, 37]), וכן של התכנון של כל אחד מהמסננים המרכיבים את המערך באופן בלתי-תלוי.

<u>טבלה 2.1</u>: היתרונות והחסרונות של חמש גישות לתכנון מערכי מסננים.

<u>Table 2.1</u>: Comparison of Five Filter Banks' Design Methods.

השיטה	יתרונות	חסרונות
1. שיטת ה"חלון" [37].	* ניתן לשלוט בתגובה הכוללת. * פשטות בתכנון.	* תגובות התדר של המסננים שבמערך אינן אופטימליות. * אורכי כל המסננים זהים. * אותו "חלון" משמש לתכנון כל המסננים.
2. קריטריון Min-Max מקורב [39].	* ניתן להשיג תגובה כוללת שהיא תגובת יחידה. * תכנון פשוט יחסית.	* מוגבלת לתגובה כוללת שהיא תגובת יחידה. * מוגבלת לתכנון מערכי מסננים אחידים, ולמסננים באורך אי-זוגי. * תגובות התדר של המסננים שבמערך הן תת-אופטימליות.
3. אלגוריתם איטרטיבי בשיטת Min-Max [38].	* ניתן להשיג תגובת תדר אופטימלית של המסננים במערך ובו זמנית לאלץ את התגובה הכוללת.	* האלגוריתם מסובך מאד, ולכן המסננים בד"כ חייבים להיות קצרים. * תיתכן התבדרות של האלגוריתם.
4. קריטריון Min-Max [40].	* ניתן להשיג תגובת תדר אופטימלית של המסננים במערך ובו זמנית לאלץ את התגובה הכוללת.	* מוגבל לתכנון מערכי מסננים אחידים ולמסננים באורך אי-זוגי. * השימוש בתכנות-לינארי מחייב יותר כושר חישוב.
5. תכנון כל מסנן ומסנן באופן בלתי-תלוי.	* תגובות התדר של המסננים השונים הן אופטימליות. * תכנון פשוט יחסית.	* התגובה הכוללת היא בדרך כלל גרועה מאד.

2.3 <u>תכנון מערכות אנליזה - סינתזה</u>

על פי הגדרתן, מערכות אנליזה-סינתזה מכילות שני מערכי-מסננים כלליים (לאו דוקא אחידים). האחד משמש לאנליזת האות והשני לסינתזה שלו (בדרך כלל לאחר שעבר מודיפיקציה כלשהי).

מרבית מערכות האנליזה - סינתזה המקובלות הן מאחת מתוך שתי הצורות הבאות:

(א) מערכות המבוססות על מערכי מסננים אחידים ולפיכך מימושן הוא על ידי FFT.

(ב) מערכות המבוססות על עצים בינריים של מסננים, כשבכל דרגה מופיע צמד מסנני QMF (Quadrature Mirror Filters), כפי שיוסבר בהמשך.

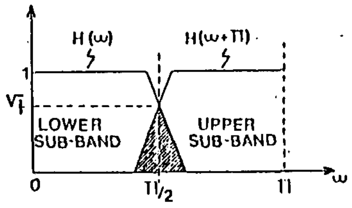
בעבודה זו נתייחס אך ורק לתכנון מערכות אנליזה - סינתזה מהצורה הראשונה (קרי מבוססות על מערכי מסננים אחידים)' אולם על מנת להשלים את התמונה הכללית נסקור ראשית בקצרה את הגישה האלטרנטיווית של מסנני QMF.

<u>(א) תכנון מערכת אנליזה - סינתזה בעזרת מסנני QMF</u>

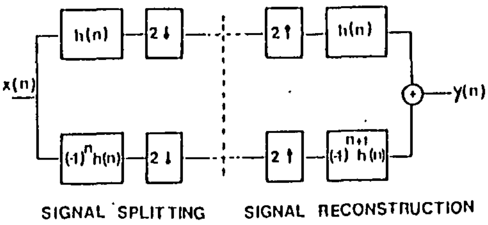
מסנני QMF מקובלים בעיקר במערכות אנליזה - סינתזה המשמשות לקידוד. במקרה זה על מנת לחסוך בכמות סיביות הקוד מקובל לבצע דצימציה ביחס המקסימלי האפשרי (קרי מס' פסי התדר M שווה ליחס הדצימציה R), ואזי ידוע שבמערכת אנליזה - סינתזה המבוססת על מערך מסננים אחיד לא ניתן לקבל מערכת - יחידה כאשר אורכי המסננים גדולים ממימד ה-DFT (ראה [16, 15, 12]). מאידך ידוע שעל מנת לקבל הפרדת תדר טובה בעת האנליזה, יש להשתמש במסנני אנליזה ארוכים ממימד ה-DFT (ראה [6]).

במערכות כאלו, השגיאה הנוצרת בשחזור האות נובעת בין היתר מקיפולי תדר (Aliassing) עקב הדצימציה. אפקט זה פוגם באיכות השמיעה של אותות דיבור, במידה רבה יותר מאשר עוותי אמפליטודה קבועים בתדר, מכיון שהאוזן מסתגלת לעוות קבוע של הדיבור, בעוד שהיא רגישה מאד לעוותים המשתנים בזמן [42, 41]. עובדה זו שאומתה בנסיונות רבים, הובילה לחיפוש מערכת אנליזה - סינתזה שבה שגיאת השחזור אינה מכילה קיפולי תדר. עבור המקרה הפרטי של R = M = 2, קרי חלוקה של תחום התדר לשני פסים (bands) בעלי רוחב זהה, הוצגה מערכת כזו ב-[43].

המערכת המבוססת על צמד מסנני-אנליזה בעלי תגובות תדר שהן שיקוף אחת של השניה סביב f = 0.25 (ולכן קרויים QMF) מתוארת בציור 2.4.



Half-band QMF filters.

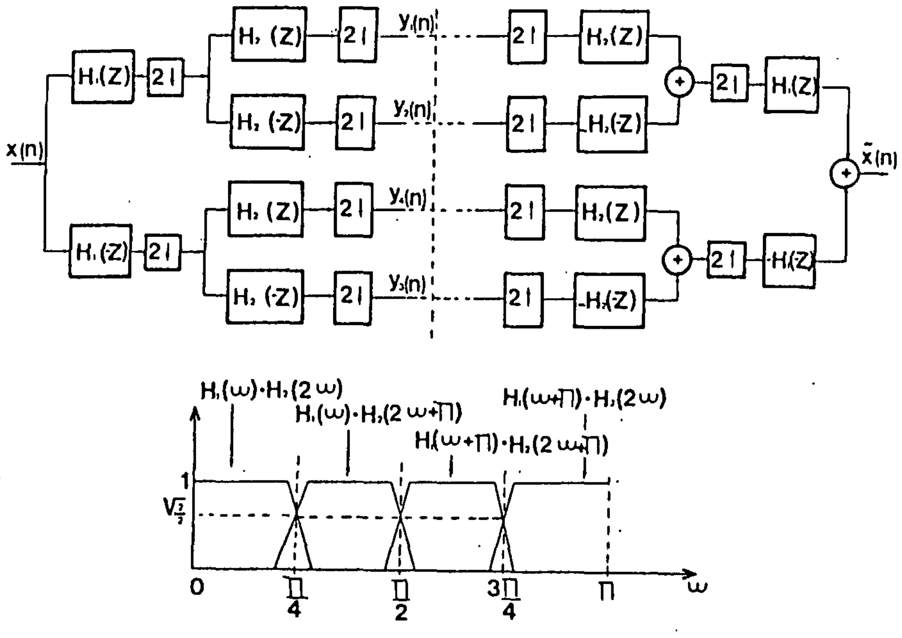


<u>ציור מס' 2.4</u>: מערכת אנליזה - סינתזה בשיטת QMF עבור R = M = 2.

<u>Fig. 2.4</u>: Analysis / Synthesis System using QMF for R = M = 2.

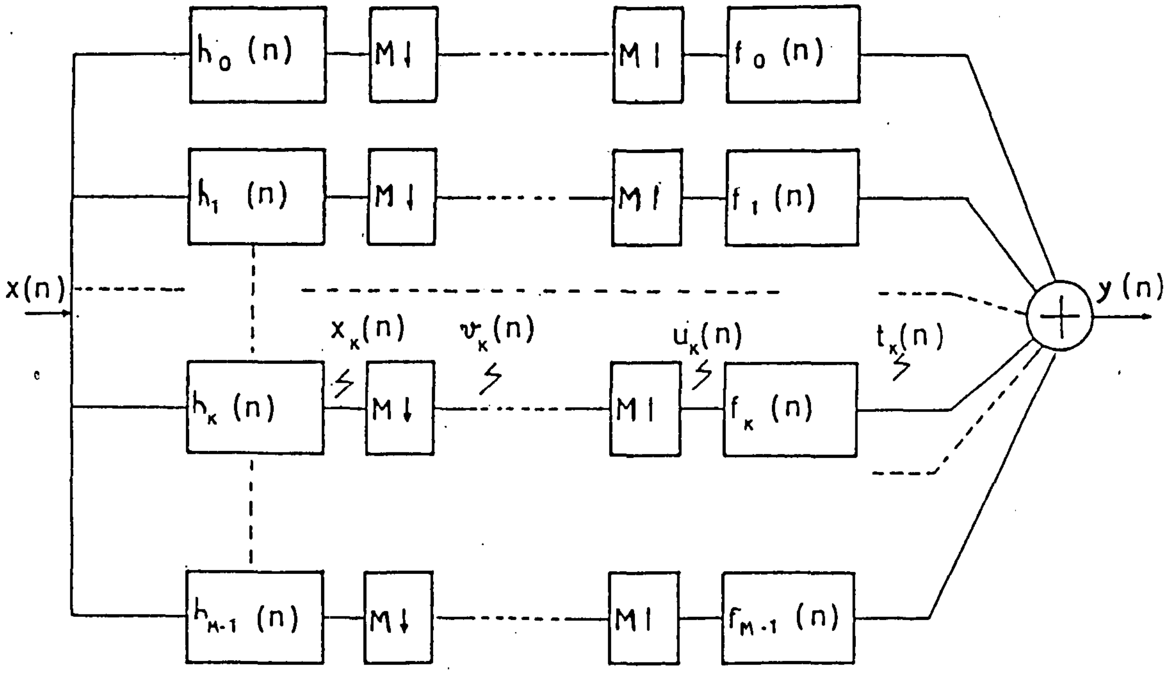
שיטות לתכנון מסנני QMF אופטימליים המבוססות על אלגוריתמי אופטמיזציה הוצעו ב-[44,45]. מאחר ושיטת ה-QMF מוגבלת למקרה של M = 2 , הוצע עבור n > 1 , $M = 2^n$ להשתמש בעץ של מסנני QMF כשבכל דרגה נעשית חלוקה נוספת ביחס 2 : 1, ובסינתזה להפעיל עץ מקביל של מסננים [47, 46].

בציור 2.5 מתוארת המערכת עבור M = 4 והדרישות מתגובות התדר של המסננים במקרה זה. השימוש בעץ של מסננים מחייב בקרה מורכבת של המערכת ולכן הוצעה גם גישה אלטרנטיבית של מימוש העץ הנ"ל על ידי מערך מסננים מקבילי [48], כמודגם בציור 2.6, אך חסרונה העיקרי בכך שהמסננים הנוצרים הם ארוכים מאד ולכן סיבוכיות המימוש גדולה (ראה השוואה ב-[46]).



ציור מס' 2.5: תאור של עץ מסנני QMF עבור M = 4.

Fig. 2.5: QMF Tree Structure for M = 4.



<u>ציור מס' 2.6</u>: תאור של מערך מקבילי של מסנני QMF.

<u>Fig. 2.6</u>: Perrallel Implementation of QMF Filters for $M > 2$.

לפיכך הוצעו שיטות אלטרנטיביות לתכנון מערכי מסננים מקביליים עבור $M > 2$, שלהן בקירוב יש את תכונת ביטול הקיפול בתדר (המאפיינת את ה-QMF עבור $M = 2$) [49, 50]. המחקר בנושא זה עודנו אקטיבי.

חשוב לציין את שני החסרונות הבאים של המערכות המבוססות על מסנני QMF:

(א) במערכות אלו נוצר סיבוך יתר בהשוואה למערכות המבוססות על מערכי מסננים אחידים הממומשים בעזרת FFT. הסיבוך הנ"ל אינו מוצדק במקרה שבו $M > R$, קרי בשימושים כגון שינוי ציר-הזמן של האות, ערבול דיבור וסינון דיבור. ואכן בכל השימושים הנ"ל נפוצים מערכי המסננים האחידים ולא מערכות מסנני QMF.

(ב) במקרה של $R = M$, כפי שנדרש בשימושי קידוד דיבור, המוטיבציה לשימוש במסנני QMF היא תכונת ביטול קיפול התדר. תכונה זו מתקיימת כמובן ללא כימות, וכאשר מבוצע כימות של הדיבור (לצורך קידוד), הרי נוצרים עוותים לא-ליניאריים בשחזור הכוללים כמובן קפול בתדר. ככל שהקידוד נעשה עם קצב סיביות לשניה נמוך יותר, כן השפעת הכימות ניכרת יותר.

לפיכך נראה היה לנו שמוטב להתרכז במערכות אנליזה - סינתזה המבוססות על מערכי מסננים אחידים ולתכנן מערכות אופטימליות עבור מקודדים נתונים, תוך ניצול האפיון הסטטיסטי של המקודדים והאותות.

(ב) <u>תכנון מערכות אנליזה - סינתזה עם מערכי מסננים אחידים</u>

מערכות אנליזה - סינתזה עם מערכי מסננים אחידים צמחו מתוך התמרת פורייה לזמן - קצר (STFT), שמשמשת כלי בסיסי בניתוח אותות דיבור, ואותות קווזי-סטציונריים אחרים (ראה ב-[53, 52, 51, 5]).

בעוד שהתמרת פורייה לזמן-קצר מהווה מוטיבציה פיסיקלית לשימוש במערכות אנליזה - סינתזה עם מערכי מסננים לעיבוד אותות דיבור, הרי השימוש במערך מסננים אחיד מונע בעיקר משיקולי סיבוכיות המימוש.

מימוש יעיל של מערכים כאלו שהוצג לצורך מעבר מריבוב זמן לריבוב תדר [6,7], תואר בציור מס' 1.4. ממוש זה של מערכי המסננים מתאים לדגימות של ה-STFT בסריג-אחיד של M נקודות תדר, והטרנספורם המתקבל קרוי התמרת פורייה דיסקרטית לזמן קצר (DSTFT). תחילה מומשו מערכות אנליזה סינתזה עם מסנני אנליזה בלבד כשבסינתזה נעשה סיכום של דגימות התמרת פוריה לזמן-קצר. שיטה זו מכונה Filter Bank Summation (FBS).

במקביל פותחה שיטת סינתזה שניה שבה נעשתה התמרת פורייה הפוכה ואחר כך סוכמו הוקטורים המתקבלים עבור וקטורי התמרה מזמנים שונים. שיטה זו מכונה Overlapp and Add (OLA). ב-[13] הוצגה שיטת WOLA המאחדת שתי שיטות אלה על ידי הכנסת מערך מסנני סינתזה המשמשים לאינטרפולציה (במקרה של R > 1) ולסילוק רכיבי תדר לא רצויים כאשר האות עובר מודיפיקציה לאחר האנליזה.

ב-[6] מתוארת שיטת סינתזה זו בפרוט ומודגם שהן ה-FBS והן ה-OLA הם מקרים פרטיים שלה.

השיטה הוצגה במקור עבור מסנני אנליזה שארכם קטן ממימד הטרנספורם (קרי M) והרחבה אחר כך למקרה הכללי יותר ב-[10].

במקביל, בהתבסס על האינטרפרטציה של מערכות אנליזה-סינתזה ככלי לחישוב ה-STFT, הוצגו ב-[15] תנאים לקבלת מערכת יחידה בשיטת WOLA, תנאים אלה קרויים תנאי פורטנוף.

ב-[12] הוצגו תנאים מספיקים והכרחיים על מסנני האנליזה המאפשרים קבלת מערכת יחידה על ידי **<u>סינתזה לינארית כלשהי</u>** תוך שימוש בגישה אלגברית לתאור פעולת ה-DSTFT והסינתזה של אות זמני מתוכו.

בטבלה 2.2 מתוארים התנאים הללו באופן תמציתי.

טבלה מס' 2.2: התנאים שבהם קיימת מערכת יחידה.

Table 2.2: Conditions for Existence of Unity Systems.

התנאים	R = M	R < M	R > M
1. $L_h = M$ ואינו מכיל אפסים	* מערכת יחידה אחת בדיוק קיימת.	* קיימות אינסוף מערכות יחידה.	* אין מערכת יחידה.
2. $L_h > M$ וכל polyphase מכיל לפחות דגם אחד השונה מאפס.	* אין מערכת יחידה. (כשתנאי ההתחלה לא ידועים).	* קיימות אינסוף מערכות יחידה	* אין מערכת יחידה.

מכיוון ועבור המקרה המעניין של $R = M$ וארכי מסנני אנליזה הגדולים ממימד הטרנספורם, לא קיימת מערכת יחידה בשיטת WOLA [15, 12], נדרשה שיטה לתכנון המערכת האופטימלית למקרה זה.

ב-[54] פותחה שיטה כזו, כשקריטריון התכנון הוא מינימיזצית האנרגיה של התגובה לדגם יחידה[1] של מערכת אנליזה-סינתזה בשיטת WOLA, תחת שני האילוצים הבאים:

1. האנרגיה הממוצעת של מסנן האב-טיפוס של האנליזה בתחום ההנחתה שלו מוכתבת (קטנה). קרי מערך מסנני האנליזה מבצע הפרדת תדר מספקת.

---

1 מערכת אנליזה-סינתזה היא מערכת משתנה בזמן. לפיכך התגובה לדגם יחידה שלה תלויה מפורשות **בזמן** שבו הופיע בכניסה הדגם הנ"ל. מאחר והתלות בזמן היא תלות מחזורית במערכות אלו, המזעור ב-[54] הוא של האנרגיה הממוצעת של התגובה לדגם יחידה, כשהממוצע הוא על פי מחזור אחד של התגובה הנ"ל. לפרוט ראה ב-[6].

2. על המערכת הכוללת להוות קירוב למערכת יחידה, פרט להשהייה קבועה Mr_0 המוכתבת על ידי המשתמש. לפיכך הערך של הדגם ה-Mr_0-י בתגובה לדגם יחידה של המערכת מאולץ להיות 1.

האלגוריתם המתקבל הוא אלגוריתם איטרטיבי והוא סבוך למדי עבור M גדול. בנוסף לכך, האופטימליות של הפתרון היא בהנחה שאין מודיפיקציה בתוך המערכת. האלגוריתם אכן נוסה רק עבור המקרה של M = 2 [55] ולמעשה המוטיבציה לפיתוחו היתה עבור המקרה של מסנני QMF, כמתואר ב-[45].

באופן בלתי תלוי פותחה שיטת תכנון אלטרנטיבית עבור המקרה שבו קיימת מודיפיקציה של האות ב-[17, 16]. שיטה זו מוגבלת למקרה שארכי המסננים קטנים או שווים למימד הטרנספורם ולכן אין חפיפה בין השיטות. מאידך בשיטה זו מניחים קיומה של מודיפיקציה כללית כלשהי של האות לאחר אנליזה, ומחפשים שחזור על ידי סדרה זמנית של המודיפיקציה של ה-DSTFT (MDSTFT), כך שאחרי אנליזה של הסדרה המשוחזרת הנ"ל יתקבל ה-DSTFT הקרוב ביותר ל-MDSTFT הנתון במובן של שגיאה ריבועית ממוצעת מינימלית (MMSE).

הן ב-[16] והן ב-[17] מתקבלת אותה תוצאה. מסנן אב-טיפוס הסינתזה האופטימלי המתקבל, מתאים לסינתזה בשיטת WOLA והוא גם מבטיח מערכת יחידה כאשר אין מודיפיקציה של האות לאחר האנליזה שלו. בצורה זו ניתנת גם שיטה אנליטית לבחירת "מערכת היחידה האופטימלית", מבין כל אלו המקיימות את תנאי פורטנוף.

ההבדל בין שתי העבודות הנ"ל הוא בהנחות תחתן פותח הפתרון, כדלקמן:

(א) ב-[16] מגבילים את הדיון לסינתזה-לינארית, אולם מתירים אפסים בתגובת דגם יחידה של מסנן האב-טיפוס של האנליזה ומניחים את ההנחה (המעשית) שנתונה ה-MDSTFT בלבד.

(ב) ב-[17] הדיון הוא עבור סינתזה כלשהי (כולל גם סינתזה לא-לינארית), אך ישנן שתי הנחות מגבילות והן שתגובת דגם היחידה של מסנן האב-טיפוס של האנליזה אינה מתאפסת, וכן שנתונה כל המודיפיקציה של ה-STFT ולא רק הדגימות שלה.

הפתרון לבעית הסינתזה האופטימלית עבור מסנני אנליזה ארוכים ממימד הטרנספורם אינו ידוע (ראה [12]), וכן לא בדור משיטות אלה, איך לנצל ידע סטטיסטי על המודיפיקציה על מנת לשפר את ביצועי המערכת.

נושא נוסף שנותר פתוח הוא אפיון פשוט של מודיפיקציות שתחתן ה-MDSTFT הוא DSTFT "חוקי", במובן שקיימת סדרה זמנית שה-DSTFT שלה מתלכד עם ה-MDSTFT הנתון.

עבור מודיפיקציות (הפיכות) כאלה ניתן לשחזר את האות המקורי ללא שגיאה, על ידי סכמה כזו המתוארת בציור מס' 1.8. הפתרון לשתי בעיות הללו מוצג בפרק 6 של העבודה.
טבלה 2.3 מסכמת את היתרונות והחסרונות בשלוש הגישות העקרוניות לתכנון מערכות אנליזה / סינתזה שנסקרו כאן והן: מסנני QMF, שיטת התכנון על פי [54] ושיטת התכנון על פי [16,17].

טבלה 2.3: היתרונות והחסרונות של שלוש גישות לתכנון מערכות אנליזה-סינתזה.

Table 2.3: Comparison of Three Design Methods for Analysis/Synthesis Systems.

השיטה	יתרונות	חסרונות
1. מסנני QMF	* ביטול קיפולים בתדר	* מוגבלת לקצב דצימציה קריטי * מסובך למימוש עבור M >> 2. * מתעלמת מהשפעת המודיפיקציה.
2. מערכות יחידה מקורבות [54]	* קירוב אופטימלי למערכת יחידה. * תקפה בקצב דצימציה קריטי * המימוש פשוט יחסית.	* מוגבלת לקצב דצימציה קריטי. * מתעלמת מהשפעת המודיפיקציה. * אין ביטול קיפולים בתדר.
3. סינתזה אופטימלית MDSTFT-מ [16,17].	* תקפה לכל קצב דצימציה. * המימוש פשוט יחסית. * מתחשבת בקיום מודיפיקציה.	* לא ידוע פתרון סגור עבור אורכי מסננים הגדולים מהמימד הטרנספורם. * אינה מנצלת פרמטרים ספציפיים של המודיפיקציה. * אין ביטול קיפולים בתדר.

פרק 3 : תכנון מערכי מסננים עם תגובה כוללת מוכתבת

3.1 תכנון מערכים אופטימלים בקריטריון WMMSE

כאשר התגובה הכוללת של מערך המסננים מוכתבת, בדרך כלל לא ניתן להשיגה על ידי תכנון נפרד של כל מסנן ומסנן, כפי שהוסבר והודגם בסעיף 2.2. התכנון המשולב של כל המערך יוצר בעית אופטימיזציה ממימד גדול (מס' הנעלמים הטיפוסי הוא כאלף).

יתר על כן, מכיון שקריטריון הטיב הוא על תגובות התדר של המסננים, גישה ישירה לתכנון אופטימלי בקריטריון Min-Max מחייבת בדרך כלל מספר אינסופי של אילוצים (אילוץ בכל תדר ותדר).

שיטת ה"חלון" נותנת תשובה חלקית לבעיה, אך היא מוגבלת למערכי מסנני FIR באורך זהה ונותנת הפרדת תדר לא-אופטימלית (ראה סעיפים 2.1, 2.2). לפיכך הכללנו את שיטת WMMSE ששמשה עד כה לתכנון מסנן FIR יחיד, לתכנון מערכי מסננים אופטימליים עם תגובה כוללת מוכתבת. השיטה פותחה עבור מקרה כללי יותר המאפשר תכנון מערכי מסננים עם חוליות בסיסיות כלשהן (כגון אלו המופיעות ב-[35,36]), ובפרט כל מסנן במערך יכול להיות בעל אורך שונה. יתר על כן, החוליות הבסיסיות (המוכתבות מראש) יכולות להיות אף מסנני IIR (Infinite Impulse Response), ובלבד שהאופטימיזציה תבוצע רק על מקדמי הקומבינציה הלינארית של החוליות הללו (ראה ציור מס' 2.1).

המקרה הפשוט ביותר הוא כשהתגובה הכוללת מוכתבת כאילוץ. בניגוד למקרה של קריטריון Min-Max האילוץ משרה למעשה מספר סופי של אילוצים לינאריים על מקדמי המסננים במערך [2], ויש לפתור בעית WMMSE עם מספר סופי של אילוצי שיוויון לינאריים.

בטרם נדון בפתרון נבהיר את מדד השגיאה.

בסעיף 2.1 ניסחנו את קריטריון WMMSE לתכנון מסנן יחיד. באופן דומה מדד השגיאה עבור מערך המסננים יהא הנורמה האוקלידית של וקטור ערכי נורמות L_2 של שגיאות-התדר של המסננים המרכיבים את המערך לאחר שעברו שקלול (המוכתב על ידי המתכנן). מערך המסננים האופטימלי מביא למינימום את מדד השגיאה על פני כל מערכי המסננים המקיימים את אילוצי התגובה הכוללת המוכתבת.

---

2 במסגרת הכללית יותר שבה המסננים מורכבים מחוליות בסיסיות כלשהן, הרי הנעלמים הם הגברי החוליות ולא מקדמי תגובת דגם יחידה של המסננים. אם זאת למטרות הבהירות של הצגת השיטה נכנה אותם מקדמי המסננים. כינוי זה מדויק עבור מסנני FIR מקובלים - בהם החוליות הבסיסיות הן אלמנטי השהיה.

שיטת הפתרון במקרה זה היא כדלקמן:

(א) ממירים את בעית המינימיזציה של הנורמה האוקלידית של נורמות L_2 במינימיזציה של תבנית ריבועית תחת אילוצים לינארים, על ידי חישוב אנליטי של התמרות פורייה הפוכות של פונקציות המשקל ותגובות התדר הרלוונטיות.

(ב) את בעית המינימיזציה המאולצת שנוצרה, ניתן לפתור באופן אנליטי על ידי גזירה, תוך שימוש בכופלי לגרנז' [56].

(ג) התוצאה הסופית של תהליך זה היא שמקדמי המערך האופטימלי ניתנים על ידי פתרון (נומרי) של מערכת משוואות לינאריות שמימדה כמס' הנעלמים הכולל במערך (קרי, סכום אורכי המסננים השונים).

על ידי ניצול המבנה המיוחד של מטריצת המקדמים של מערכת המשוואות ניתן לפרק את הבעיה לפתרון של מספר מערכות משוואות ממימד קטן יותר. ליתר דיוק, יהא N מס' המסננים במערך ו- M_i מספר הנעלמים במסנן ה-i-י ($1 \leq i \leq N$). אזי בעוד שמימד מערכת המשוואות המקורית הוא $\sum_{i=1}^{N} M_i$, הרי הפרוק מוביל ל-N מערכות משוואות כשמימד המערכת ה-i-ית הוא M_i.

ניסוח מתמטי מדויק של מקרה זה ושיטת הפתרון עבור מסנני FIR המרכבים מאלמנטי השהייה הוצגו על ידינו ב-[57]. בדומה לקריטריון WMMSE לתכנון מסנן FIR יחיד, יש לקריטריון WMMSE לתכנון מערכי מסננים עם תגובה כוללת מוכתבת אינטרפרטציה סטטיסטית. תחת אינטרפרטציה זו מערך המסננים האופטימלי הוא זה הממזער את הסכום המשוקלל של וריאנסי השגיאה במוצא המערך תחת אילוץ התגובה הכוללת. ניתן גם להראות שזה הוא המערך המביא למקסימום את הממוצע ההרמוני של יחסי אות-לרעש (SNR) במוצא המסננים השונים. בלבוש זה שיטת התכנון מתאימה גם לבעיות בתחום התקשורת, וכך היא הוצגה על ידינו ב-[58].

נדון כעת במקרה הכללי יותר שבו התגובה הכוללת מוכתבת באחת מהצורות הבאות:

(א) נורמת L_2 של שגיאת תגובת התדר הכוללת ביחס לתגובת התדר הכוללת הרצויה, מוכפלת בגורם שקלול מתאים (שיסומן להלן K_{N+1}^2), ומוספת למדד השגיאה של WMMSE שתואר לעיל.

(ב) מוכתב אילוץ tolerance - כלומר נורמת L_2 של שגיאת תגובת התדר הכוללת ביחס לתגובת התדר הכוללת הרצויה צריכה להיות קטנה מגודל מסוים המוכתב על ידי המתכנן (שיסומן להלן על ידי η).

בנספח א', סעיף 1 מנוסחות באופן מתמטי מדויק שתי בעיות התכנון הנ"ל ומוצג פתרונן. המקרה של תגובה כללית המוכתבת על ידי אילוץ שתואר לעיל הוא מקרה פרטי של בעיות אלה, המוראה שם. נתאר להלן את עיקרי הפתרון.

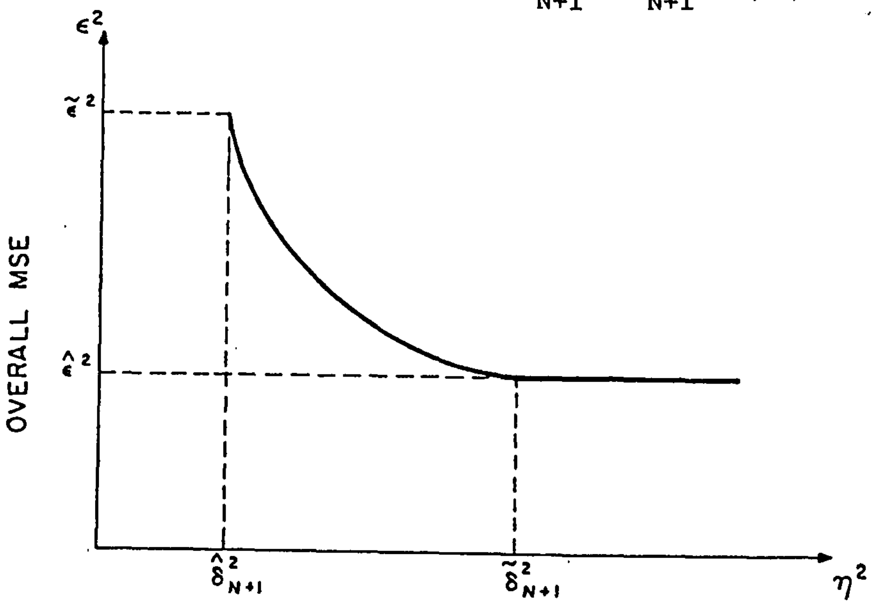
בהסתמך על [56], שתי בעיות התכנון שקולות במובן הבא: לכל ערך של η עבורו קיים פתרון לבעית התכנון (ב), ישנו ערך $K^2_{N+1}(\eta)$ כך שהמערך האופטימלי המתקבל כפתרון בעית התכנון (א) בשמוש ב-$K^2_{N+1}(\eta)$ הוא גם הפתרון של בעית התכנון (ב) עבור η, ולהיפך.

לפיכך מספיק להציג שיטה לפתרון בעית התכנון (א) (שהיא קלה יותר לפתרון) ואלגוריתם לחישוב $K^2_{N+1}(\eta)$.

בנספח א', סעיף 1 מתואר באופן מפורט פתרון בעית התכנון (א). נעיר כאן רק שהפתרון דומה מאוד לפתרון שתואר לעיל עבור המקרה שבו התגובה הכוללת מוכתבת כאילוץ.

האלגוריתם לחישוב $K^2_{N+1}(\eta)$ מבוסס על לכסון סימולטני של שתי מטריצות הרמיטיות (ראה [59]), ופתרון משוואה סקלרית לא-לינארית בשיטות של תכנות לא-לינארי (ראה [56]). בנספח א', סעיף 1 מתואר אלגוריתם זה בפרוט.

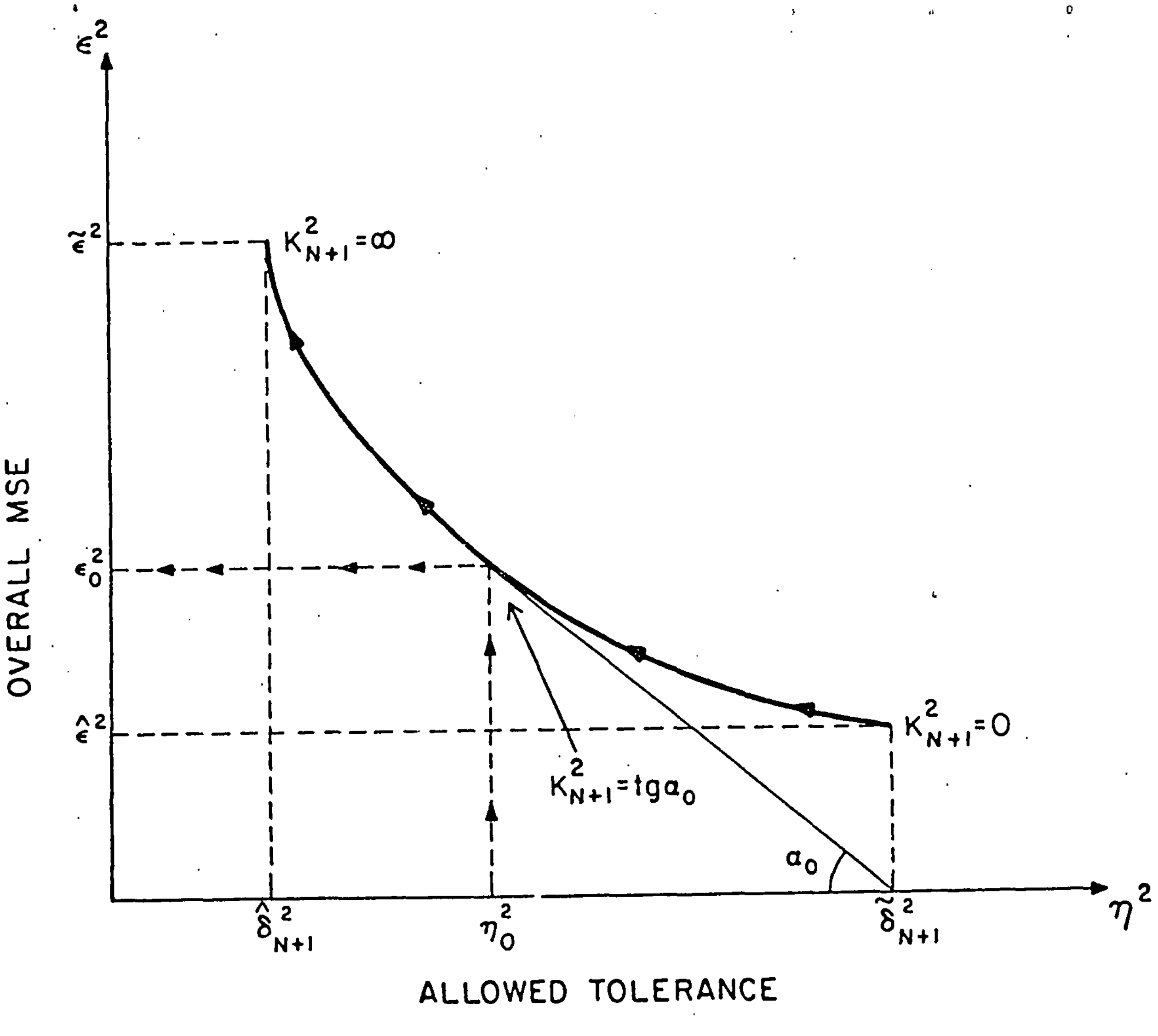
יהא ε מדד השגיאה המושג על ידי מערך המסננים האופטימלי. בציור מס' 3.1a מתואר עקום טיפוסי של ε^2 כפונקציה של ה- tolerance η^2. זהו עקום מונוטוני לא-עולה וקמור, (convex), כפי שמוכח בנספח א', סעיף 1. נקודת הקיצון $(\tilde{\varepsilon}^2 , \hat{\delta}^2_{N+1})$ בציור זה מייצגת את המקרה הפרטי של אילוץ התגובה הכוללת שתואר ב-[57], ואילו נקודת הקיצון $(\hat{\varepsilon}^2 , \tilde{\delta}^2_{N+1})$ מייצגת את המקרה הפרטי של תכנון ללא ספציפיקציות לגבי התגובה הכוללת, תוך שימוש בשיטת WMMSE לתכנון המסננים המרכיבים את המערך, כמתואר בסעיף 2.1. ביטויים אנליטיים לערכי $\hat{\varepsilon}^2, \tilde{\varepsilon}^2$, $\hat{\delta}^2_{N+1}$ ו-$\tilde{\delta}_{N+1}$ מפותחים בנספח א', סעיף 1.



ציור מס' 3.1a: עקום אופייני של ה-MSE של מערך המסננים האופטימלי (ε^2) כפונקציה של ה- tolerance המותר בתגובה הכוללת (η^2).

Fig. 3.1a: Typical trade-off curve of the MSE ε^2 as function of the allowed tolerance from the specified composite response (η^2).

בציור מס' 3.1b ניתנת אינטרפרטציה גאומטרית ל-(η) K^2_{N+1} , שכן זהו בדיוק שפוע המשיק לעקום של ε^2 כפונקציה של η^2. יתר על כן שתי נקודות הקיצון בעקום זה שתוארו לעיל מתאימות ל- $K^2_{N+1} = \infty$ ו- $K^2_{N+1} = 0$ בהתאמה. עד כאן לתאור פתרון שתי בעיות התכנון ותכונות העקום של ε^2 כפונקציה של η^2 .



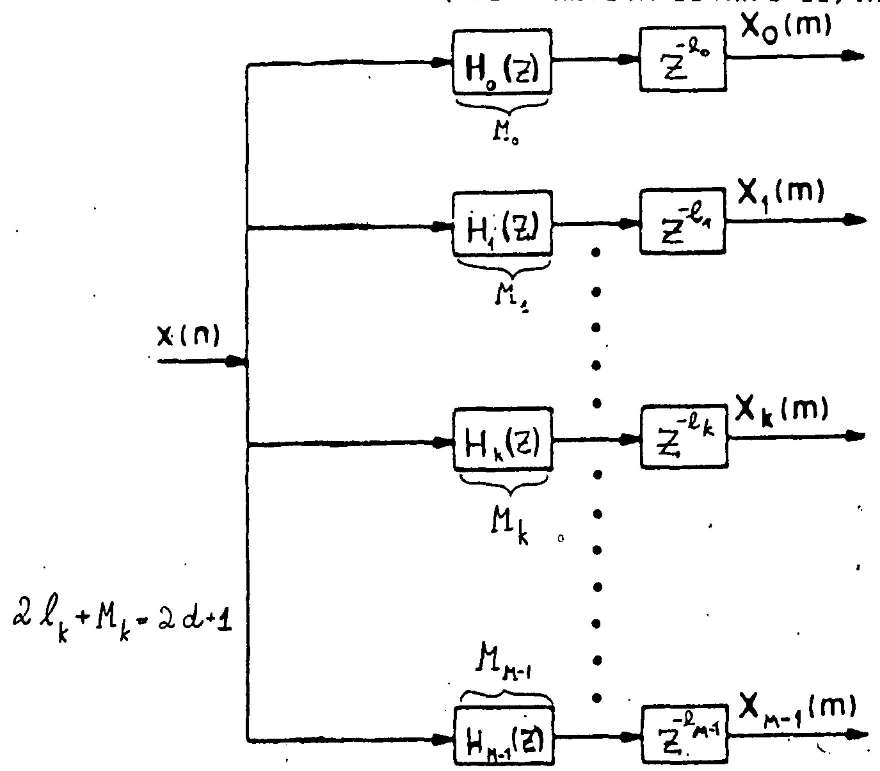
ציור מס' 3.1b: אינטרפרטציה גאומטרית של מקדם השקלול של התגובה הכוללת (K^2_{N+1}).

Fig. 3.1b: Geometrical interpretation of the weight constant K^2_{N+1}.

פיתוח שיטת התכנון נעשה עבור מערכי מסננים עם מקדמים קומפלקסיים, ואין דרישת פזה - לינארית של המסננים המתקבלים כחלק מדרישות התכנון. נשאלת השאלה מהם התנאים בהם מובטחת פזה - לינארית של כל המסננים שבמערך האופטימלי ומהם התנאים בהם מקדמי המסננים האופטימליים הם ממשיים. התשובה המדויקת לכך והוכחתה ניתנות בנספח א', סעיף II, ונתאר אותה כאן בקצרה, ולמטרות הפשטות רק עבור מערכי מסנני FIR שחוליותיהם הבסיסיות הן אלמנטי השהייה.

(א) אם כל פונקציות המשקל ותגובות התדר הרצויות, המוגדרות בתהליך התכנון, הן התמרות פורייה של סדרות ממשיות, אזי מקדמי כל המסננים במערך האופטימלי הם ממשיים.

(ב) אם לכל תגובות התדר הרצויות ישנה פזה-לינארית זהה המתאימה להשהייה של d דגימות, ואם לכל המסננים שבמערך מוסיפים את ההשהיות המתאימות (כמתואר בציור 3.2), אזי לכל מסנני המערך האופטימלי תהא פזה-לינארית המתאימה להשהייה של d דגימות (בש-d הוא כפולה שלמה של 1/2).



ציור מס' 3.2: מערך מסנני FIR, בעל השהיות מתאימות המבטיחות פזה לינארית.

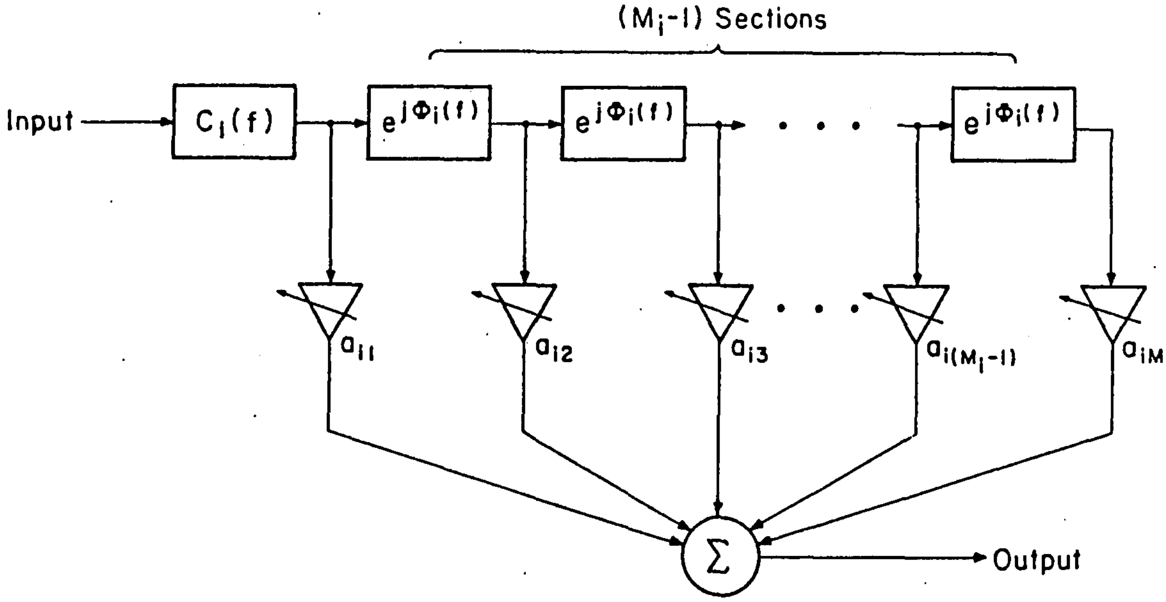
Fig. 3.2: A Digital Filter Bank Composed of FIR Filters, with Appropriate Delays that Guarantee a Linear Phase.

האינטרפרטציה הסטטיסטית של שיטת התכנון שתוארה לעיל, אנלוגית לזו שתוארה כבר עבור המקרה הפרטי של אילוץ התגובה הכוללת והמופיעה ב-[58]. הפיתוח המפורט של אינטרפרטציה זו במקרה הכללי ניתן בנספח א', סעיף III. יתרונה בשימושים בשטח התקשורת (כמתואר למשל ב-[7]).

בעת פיתוח שיטת תכנון WMMSE למערכי מסננים, הושם דגש על הורדת סיבוכיות התכנון במידת האפשר וזאת לאור מס' הנעלמים הרב בבעיות טיפוסיות. סיכום של שלבי התכנון של מערך מסננים בשיטה זו וניתוח מפורט של סיבוכיותם ניתן בנספח א', סעיף IV.

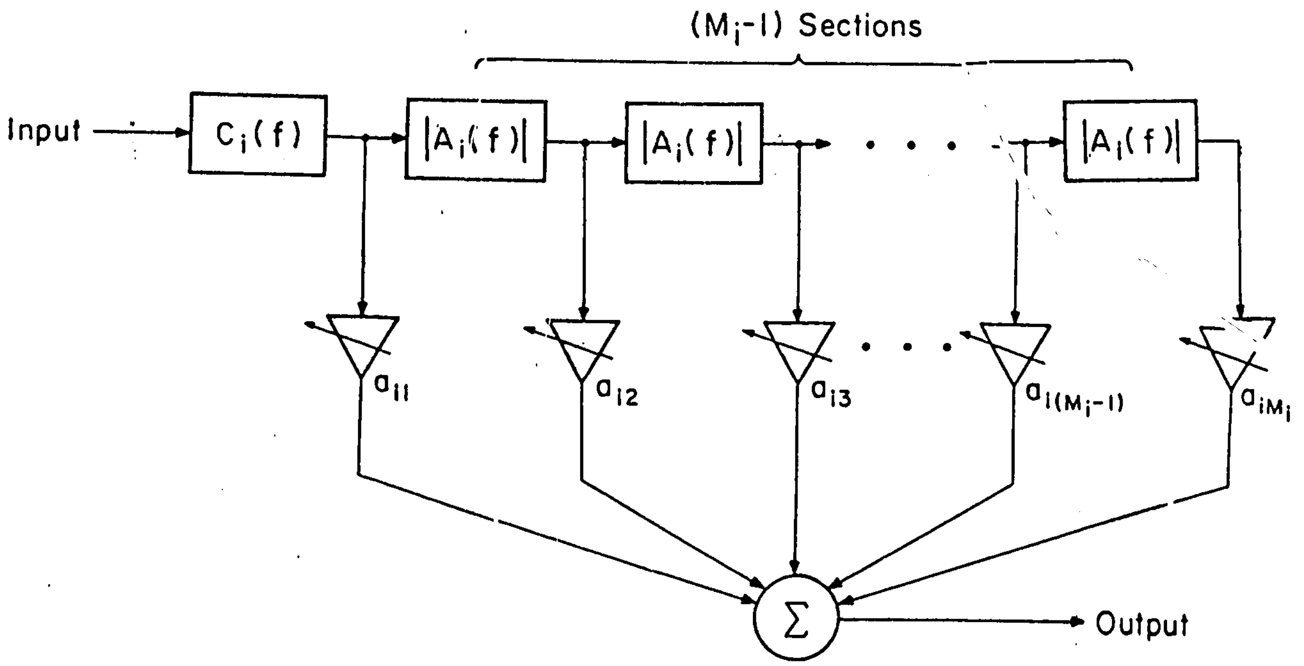
למטרות השלמות מתוארת בפרוט שיטת הלכסון הסימולטני של שתי מטריצות הרמיטיות המבוססת על [59,60] בסעיף VI באותו נספח, כולל ניתוח הסיבוכיות של שלב זה של התכנון. הסיכום של ניתוח הסיבוכיות הכוללת (כמתואר בטבלא A.1) הוא שתכנון המערך מערב פתרון (N+1) מערכות משוואות לינאריות נפרדות. ממד המערכת ה-i-ית הוא כמספר הנעלמים במסנן ה-i-י (עבור 1 ≤ i ≤ N) וממדי המערכת ה-(N+1)-ית הוא כמספר החוליות הבסיסיות השונות בכל מסנני המערך.

בפרט כאשר המסנן ה-i-י הוא מהצורה המתוארת בציור מס' 3.3 מתקבלת מערכת משוואות i-ית עם מטריצת מקדמים שהיא Toeplitz וכשהוא מהצורה המתוארת בציור מס' 3.4 מתקבלת מערכת משוואות i-ית עם מטריצת מקדמים שהיא Hankel. בשני המקרים קיימות שיטות לפתרון מהיר במיוחד של מערכות המשוואות (ראה למשל - [34] עבור מטריצות Toeplitz). נעיר שמסנן FIR עם חוליות בסיסיות שהן אלמנטי השהייה, הוא מקרה פרטי של הצורה שמתוארת בציור מס' 3.3 והמסננים שהוצעו ב-[35,36] הם מקרים פרטיים של הצורה המתוארת בציור מס' 3.4.



ציור מס' 3.3: מבנה המסנן ה-i-י במערך המסננים שעבורו מערכת המשוואות ה-i-ית היא עם מטריצת מקדמים Toeplitz.

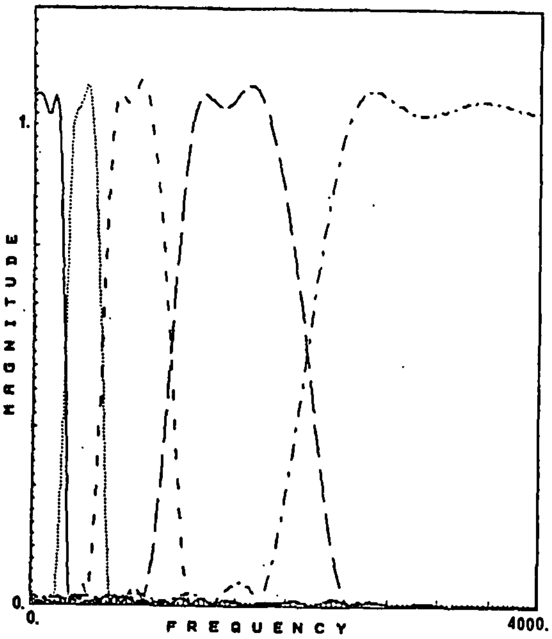
Fig. 3.3: The structure of the i-th filter for which the i-th system of equations has a Toeplitz matrix of coeficients.



ציור מס' 3.4: מבנה המסנן ה-i-י במערך המסננים שעבורו מערכת המשוואות ה-i-ית היא עם מטריצת מקדמים Hankel.

Fig. 3.4: The structure of the i-th filter for which the i-th system of equations has a Hankel matrix of coeficients.

בציור מס' 3.5a מתוארות תגובות התדר של מערך אופטימלי של חמישה מסנני FIR באורכים שונים שתוכנן עם אילוץ תגובה כוללת שהיא תגובת יחידה. בציור מס' 3.5b מוצגות למטרות השוואה תגובות התדר של מערך מסננים שתוכנן תחת אותן הדרישות למעט הסרת הספציפיקציות לגבי התגובה הכוללת שלו. תגובות התדר של המסננים המרכיבים מערך זה קרובות מעט יותר לתגובות הרצויות, אך מאידך התגובה הכוללת שלו המוצגת בציור מס' 2.2, היא גרועה.

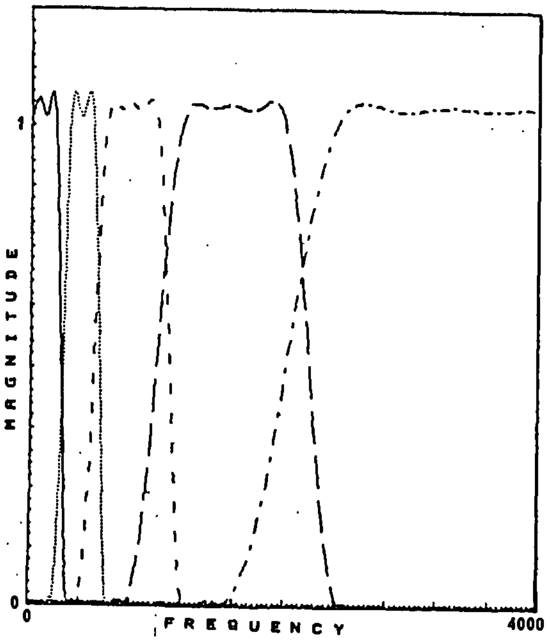


ציור 3.5a:

תגובת התדר של מערך אופטימלי המכיל חמישה מסנני FIR, שתוכנן עם אילוץ תגובה כוללת שהיא תגובת יחידה.

Fig. 3.5b:

Frequency response of an optimal filter bank with five FIR filters and unity composite response (linear magnitude scale).



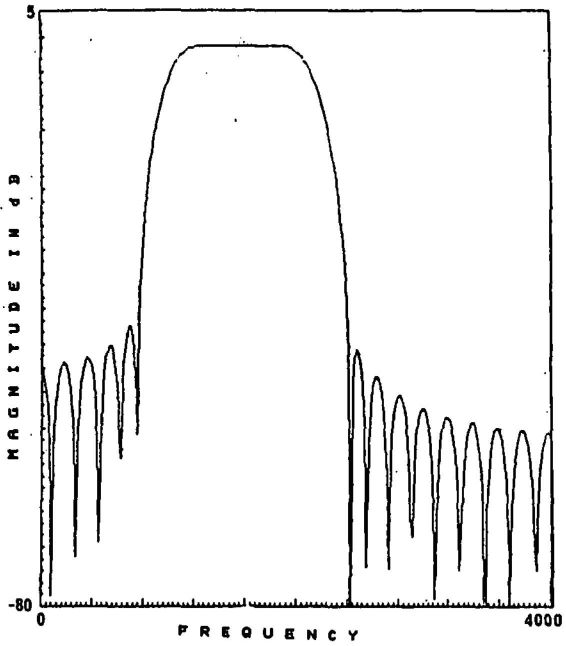
ציור 3.5b:

תגובת התדר של מערך שתוכנן תחת אותן דרישות תכנון כב-3.5a למעט התגובה הכוללת שאינה מוכתבת.

Fig. 3.5a:

Frequency response of a similar filter bank as in fig. 3.5a but with no constraint on the composite response (linear magnitude scale).

בציורים מס' 3.6a ו-3.6b מושוות תגובות התדר של המסנן הרביעי בשני המערכים בסקלה לוגריתמית. השוואה זו מחדדת את ההבדלים במידת הניחות בתחומי ההנחתה, אך עדיין נראה שההבדלים בין המסננים אינם ניכרים.

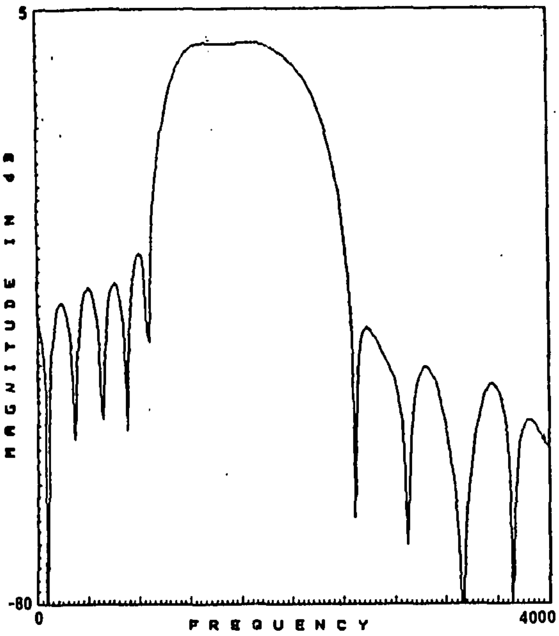


<u>ציור מס' 3.6a</u>:

תגובת התדר של המסנן הרביעי מתוך המערך שתואר בציור 3.5a, בסקלה לוגריתמית.

<u>Fig. 3.6a</u>:

Frequency response of the fourth filter in the filter bank of fig. 3.5a (in dB).



<u>ציור מס' 3.6b</u>:

תגובת התדר של המסנן הרביעי מתוך המערך שתואר בציור 3.5b, בסקלה לוגריתמית.

<u>Fig. 3.6b</u>:

Frequency response of the fourth filter in the filter bank of fig. 3.5b (in dB).

תוצאות דומות התקבלו בדוגמאות תכנון נוספות שנבדקו ומאשרות את כדאיות השימוש בשיטה המוצעת. התאור המפורט של דוגמאת התכנון המוצגת בציורים 3.6, 3.5, 2.2 ניתן בנספח א', סעיף V, כולל הערכים המדויקים של מדדי השגיאה בכל מסנן ומסנן במערך (ראה בטבלא A.2).

3.2 קיום, יחידות ותכונות של המערכים האופטימליים בקריטריונים שונים

בסעיף הקודם תארנו בהרחבה את שיטת התכנון של מערכי מסננים אופטימליים בקריטריון WMMSE כשהתגובה הכוללת מוכתבת.

האלגוריתם שהתקבל הוא בברור ספציפי לתכנון תחת קריטריון WMMSE, אולם מקצת מתכונות הפתרון (במיוחד, הקשר שבין שתי בעיות התכנון השקולות, והתנאים למחשיות המקדמים ופזה לינארית של המסננים במערך האופטימלי), עוררו את השאלה הבאה: האם תכונות אלו (ואחרות), הן אופייניות לבעיה הנדונה או שהן תלויות גם בקריטריון האופטימיזציה הספציפי שבו משתמשים.

מאחר ועבור קריטריוני אופטימיזציה אחרים, קשה מאד ובדרך כלל בלתי-אפשרי לרשום נוסחא סגורה לתאור הפתרון האופטימלי, יש להשתמש בניתוח מתמטי של מבנה הבעיה על מנת לענות על שאלה כזו. ניתוח כזה נעשה על ידינו והוביל למסקנה (הכללית) הבאה.

בבעיות של קירוב סימולטני במספר סופי (N) של תת-מרחבים לינאריים שכולם ממימד סופי, תחת האילוץ שקומבינציה לינארית נתונה של N הוקטורים המקרבים, תהווה קירוב טוב של וקטור אחר (לאו דוקא זהה לקומבינצית N הוקטורים הרצויים), ניתן לצפות לתוצאות אנלוגיות לאלו שפותחו בסעיף הקודם, ובלבד שקריטריוני טיב הקירוב יוגדרו על ידי סמי-נורמות (פונקציות ממשיות אי-שליליות, שהן לינאריות ביחס לכפל בסקלר ומקיימות את אי-שוויון המשולש).

נשים לב להכללות ביחס לקריטריון WMMSE שמשתמעות מהמסקנה דלעיל:

(א) קריטריוני השגיאה-בתדר יכולים להנתן על ידי נורמות L_q $1 \le q \le \infty$ עם פונקציות משקל כלשהן, כשניתן לכל מסנן להתאים נורמת L_q עם ערך אחר של q ועדיין התוצאות הנ"ל תקפות.

(ב) יתר על כן, אותן תוצאות תתקבלנה עבור בעיות קירוב דומות, שבכלל אינן מערבות מערכי מסננים ספרתיים, כגון קירוב סימולטני של N פונקציות ממשיות על ידי פולינומים מסדרים שונים, תחת אילוץ על סכום פולינומים אלו.

(ג) חלק מהתוצאות תתקיימנה גם עבור בעיות קירוב שאינן מכילות כלל פונקציות בתוכן, כמו קירובים במרחבי סדרות, או ב- $\mathbb{C}^n$, כפי שיפורט בהמשך.

בנספח ב', <u>סעיף I</u> מתוארת באופן מתמטי מדויק הבעיה הכללית אותה ניתחנו. <u>בסעיף II</u> של נספח זה מנוסחים המשפטים המציגים את אותן התכונות שהן אופיניות לבעיה האבסטרקטית ולא לקריטריוני האופטימיזציה הספציפיים. <u>בסעיף III</u> מנוסחות תכונות נוספות האופייניות לבעיה הכללית, אולם מתאימות רק למרחבים של העתקות (קרי מרחבי פונקציות, מרחבי סדרות, ומרחבי וקטורים - $\mathbb{C}^n$). <u>בסעיף IV</u> מוצגת

כדוגמה בעית הקירוב של פונקציות מדידות וחסומות על מעגל היחידה ב- $\mathbb{C}$, על ידי פולינומים טריגונומטריים. זו בדיוק הבעיה השקולה לבעית התכנון של מערכי מסננים ספרתיים, ועבורה מפורשות מקצת מהתוצאות הכלליות ומנוסחות במושגים המתאימים לבעית תכנון מערכי מסננים ספרתיים. בסעיף V מתוארת ההרחבה למקרה של מספר אילוצים (יותר מאילוץ אחד), ובסעיף VI של נספח ב' מרוכזות הוכחות כל המשפטים והטענות המופיעים בנספח זה.

כאן נסתפק בתאור התוצאות העיקריות, בשפה חופשית ולא מתמטית, וזאת בעיקר למטרות בהירות ההצגה שלהן.

ראשית נעיר לגבי שימוש במקצת מהתוצאות בהמשך העבודה.

משפטי הקיום שיוצגו כאן, מבטיחים (כמקרה פרטי) את קיומו של מערך מסננים אופטימלי בקריטריון Min-Max. שיטה לתכנון מערך כזה מוצגת בסעיף 4.3 עבור מערכי מסננים אחידים. משפטים אחרים שיוצגו כאן, מתארים מקצת מתכונות עקום ה-trade-off $\varepsilon(\eta)$ הקושר את טיב הקירוב בתגובה הכוללת של המערך עם שגיאת תגובות התדר של המסננים המרכיבים אותו. תוצאות אחרות מתארות תנאים מספיקים להבטחת ממשיות המקדמים ופזה לינארית של מסנני המערך האופטימלי בקריטריון Min-Max, שינוצלו בסעיף 4.3 בפיתוח שיטה לתכנון מערך מסננים זה.

(א) <u>משפטי קיום</u>:

לבעיה הנדונה קיים קירוב אופטימלי ללא אלוצים. נסמן את הערך המינימלי של קריטריון שגיאת הקרוב (המתקבל על ידי הקירוב האופטימלי ללא אילוצים) על ידי ε_m. קיים גם קירוב אופטימלי (בדרך כלל שונה מזה המוזכר לעיל), במובן של מינימיזצית שגיאת הקירוב של האבר ה-(N+1)-י על ידי קומבינצית האברים המקרבים. נסמן את הערך המינימלי של שגיאה זו על ידי η_m .

נתבונן בכל הקירובים שמגשימים את הערך המינימלי (ε_m) של שגיאת הקירוב, אזי קיים בתוכם אחד שהוא אופטימלי, בכך שמביניהם הוא בעל שגיאה מינימלית של קירוב האבר ה-(N+1)-י. נסמן ערך שגיאה זו על ידי η_M.

בבירור $\eta_m \leq \eta_M$. נתעניין במקרה (הכללי יותר) שבו $\eta_m < \eta_M$ במקרה זה, לכל $\eta_m \leq \eta \leq \eta_M$, תחת האילוץ ששגיאת הקירוב של האבר ה-(N+1)-י אינה עולה על η , קיים קירוב אופטימלי (שממזער את השגיאה המשוקללת בקירוב יתר N האברים). נסמן ב-$\varepsilon(\eta)$ את שגיאת הקירוב האופטימלי, תחת האילוץ דנן.

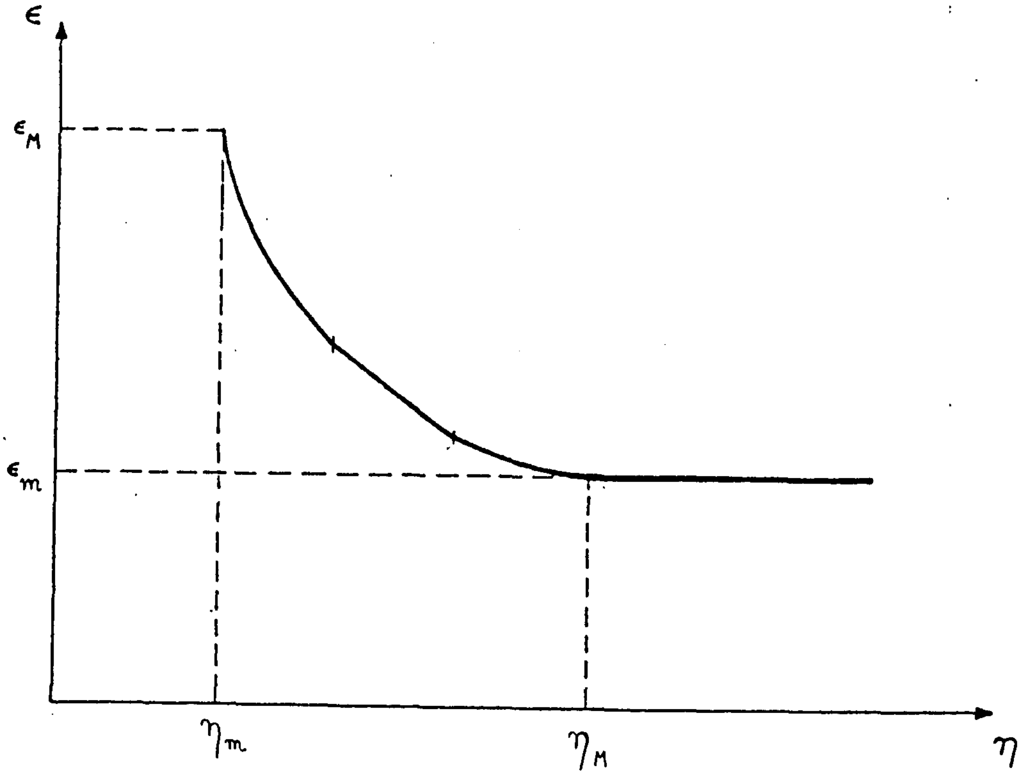
קבוצת וקטורי המקדמים של הקירוב $\underline{a}$ שעבורם מתקבל הערך של $\varepsilon(\eta)$, היא קבוצה קמורה (שתסומן על ידי $D(\eta)$, ולכל הוקטורים הללו שגיאת הקירוב של האבר ה-(N+1)-י היא בדיוק η .

(ב) <u>תכונות של הפונקציה $\varepsilon(\eta)$</u>:

הפונקציה $\varepsilon(\eta)$ מתארת את הקשר בין טיב הקירוב הסימולטני לבין טיב הקירוב של האבר ה-(N+1)-י.

זו פונקציה רציפה, חסומה, יורדת ממש וקמורה. יש לה נגזרות משמאל ומימין בכל נקודה, והנגזרת $\varepsilon'(\eta)$ מונוטונית לא-יורדת ואי-חיובית.

עקום טיפוסי של פונקציה זו מוצג בציור 3.7. לפונקציה זו יש את כל התכונות שהוכחנו בנספח א' באופן ספציפי עבור בעית תכנון מערך הסננים אופטימלי בקריטריון WMMSE (השווה לציור 3.1), למעט תכונות הקמירות ממש של $\varepsilon(\eta)$ ורציפות הנגזרת $\varepsilon'(\eta)$ שהן אכן תלויות בקריטריון האופטימיזציה הספציפי שבשימוש.



<u>ציור מס' 3.7</u>: עקום אופייני של הפונקציה $\varepsilon(\eta)$ עבור בעית הקירוב הכללית.

<u>Fig. 3.7</u>: Typical curve of the function $\varepsilon(\eta)$ for the general approximation problem.

(ג) <u>בעיה שקולה</u>:

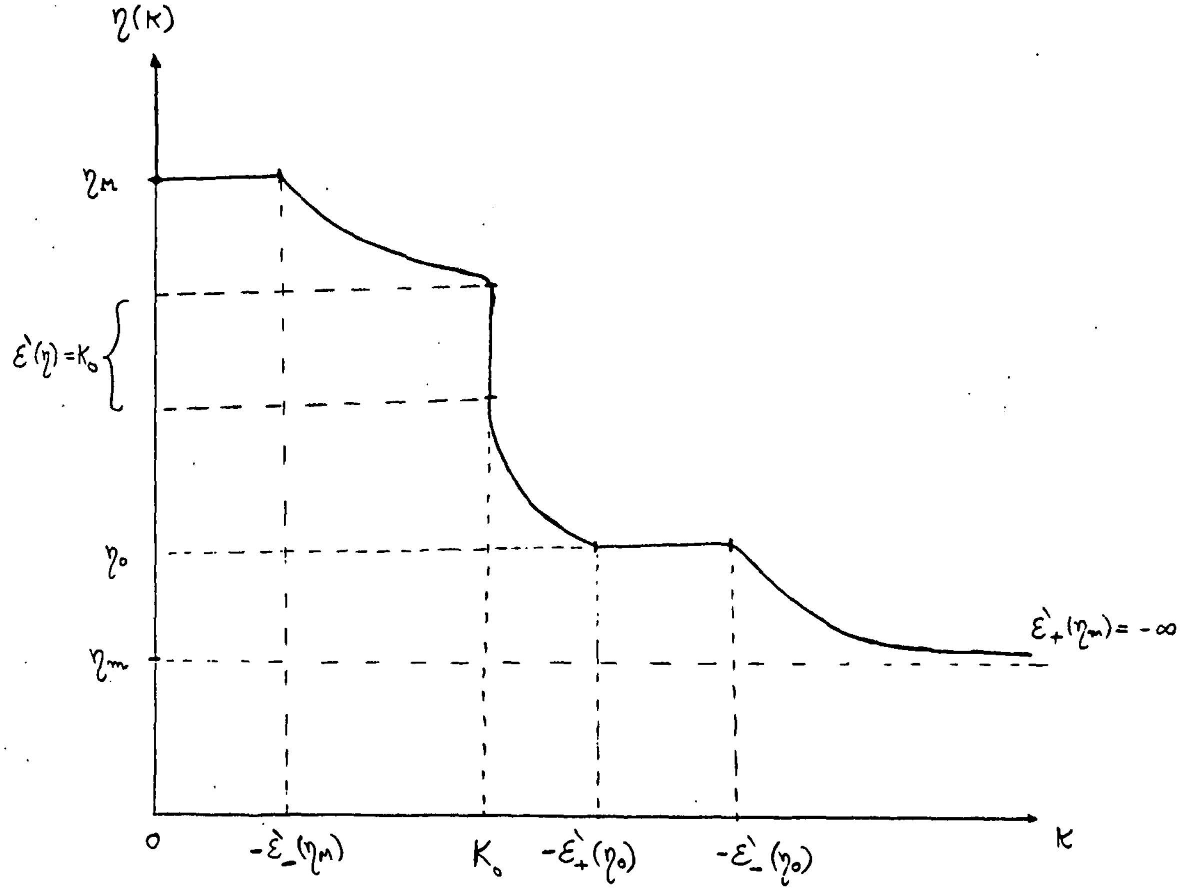
נגדיר בעית קרוב שניה, בה ממזערים את הסכום של שגיאת הקרוב הסימולטני ב-N תת-המרחבים ושל שגיאת קירוב האבר ה-(N+1)-י, כשזו האחרונה מוכפלת בקבוע משקול חיובי שיסומן על ידי K.

קל להראות שגם לבעיה זו קיים פתרון אופטימלי, ונסמן ב-$\xi(K)$ את קבוצת וקטורי המקדמים המשיגים את הערך האופטימלי עבור ערך נתון של $K \geq 0$.

בעיה זו שקולה לבעיה המקורית במובן הבא: $\xi(K) = \bigcup_{\eta \in I_K} D(\eta)$, כש- $I_0 = [\eta_M, \infty)$ ועבור $K > 0$ הרי $I_K = \{\eta \; ; \; \varepsilon_-'(\eta) \leq -K \leq \varepsilon_+'(\eta)\}$ הוא אינטרוול סגור המוכל ב- $[\eta_m, \eta_M]$ (כש- $\varepsilon_+'(\eta)$ מסמן את הנגזרת מימין של $\varepsilon(\eta)$ ו- $\varepsilon_-'(\eta)$ מסמן את הנגזרת משמאל של פונקציה זו).

לכן קיים עקום קשור ומונוטוני לא-עולה $\eta(K)$, המודגם בציור 3.8, וניתן לפרש את $(-K)$ כשפוע של $\varepsilon(\eta)$ בנקודות $\eta \in I_K$. בנוסף $\bigcup_{(\eta_m, \infty)} D(\eta) \subseteq \bigcup_{[0,\infty)} \xi(K) \subseteq \bigcup_{[\eta_m, \infty)} D(\eta)$ ולכן (עד כדי ערך הקצה $\eta = \eta_m$).

אוסף הפתרונות של שתי הבעיות מזדהה.



<u>ציור מס' 3.8</u>: עקום אופייני של $\eta(K)$ עבור בעית הקירוב הכללית.

<u>Fig. 3.8</u>: Typical curve of $\eta(K)$ for this approximation problem.

(ד) <u>תנאים ליחידות</u>

כאשר N האברים שיש לקרב, אינם בתתי-המרחב המקרבים שלהם, הסמי-נורמות המגדירות את שגיאות הקירוב בכל תתי-המרחב הן קמורות-ממש, והשקלול הסופי שלהן מונוטוני ביחס לכל אחד מערכי שגיאות אלו, הרי $\varepsilon(\eta)$ קמורה ממש, $\varepsilon'(\eta)$ מונוטונית עולה ממש וקיים וקטור מקדמים אופטימלי יחיד בכל קבוצה $D(\eta)$.
יתר על כן, במקרה זה I_K מכילה נקודה בודדת (כלומר $\eta(K)$ זו העתקה חד-ערכית), והקבוצה $\xi(K)$ גם כן מכילה וקטור מקדמים אופטימלי יחיד עבור כל K חיובי.

(ה) <u>תוצאות עבור מרחבי העתקות בלבד:</u>

ראשית מוצגים תנאים על ההעתקות השונות המעורבות בבעית הקירוב, עבורם לכל וקטור ששייך ל-$D(\eta)$ גם הצמוד הקומפלקסי שלו שייך ל-$D(\eta)$. כשתכונה זו מתקיימת, מוכח שלכל $\eta_M \le \eta \le \eta_m$ ישנו וקטור מקדמים ממשי (לפחות אחד) ב-$D(\eta)$.

כשכל הסמי-נורמות המודדות את שגיאות הקירוב הן מונוטוניות ביחס לערך המוחלט של ההעתקות, לכל (N+1) ההעתקות הרצויות יש אותה פזה (עד כדי סימן), ותת-מרחבי ההעתקות שבהם נעשה הקירוב ניתנים לתאור על ידי זוגות של העתקות הצמודות סביב הפזה הנתונה, אזי קיים ב-$D(\eta)$ וקטור מקדמים (אחד לפחות) עבורו כל ההעתקות המקרבות הן בעלות הפזה הרצויה. זו כמובן ההכללה של התנאי לפזה לינארית של המסננים האופטימליים שהוצג בסעיף 3.1.

בנוסף מוצג משפט סימטריה המתאר תנאים מסוימים בהם קיימת בעית אב-טיפוס של קירוב העתקה אחת שמתוך הפתרון האופטימלי שלה ניתן לקבל בפשטות את הפתרון המלא של בעית הקירוב הסימולטני. מקרה פרטי של משפט סימטריה זה המתאים למערכי מסננים אחידים נידון בהרחבה בסעיף הבא של העבודה.

פרק 4 : שיטות לתכנון מערכי מסננים אחידים

4.1 תכונות כלליות של מערכי מסננים אחידים אופטימליים

כמתואר במבוא לעבודה (סעיף 1.1), למערכי מסננים אחידים חשיבות רבה בעבוד אותות ספרתיים לאור היכולת לממשם ביעילות בעזרת ה-FFT [6] ושימושיהם המגוונים (למשל [7-10]).

כפי שתואר בסקר המקורות (סעיף 2.2), ישנן מספר שיטות לתכנון מערכי מסננים אחידים (למשל: [39,40]), כשבכולן מוגבל התכנון מראש למערכי מסננים שנוצרים על ידי הזזות בתדר של מסנן אב-טיפוס, וממילא עוסקות שיטות התכנון השונות בתכנון מסנן אב-טיפוס זה. נשאלת השאלה מתי (עבור אילו ספציפיקציות), קיים מערך מסננים מצורה זו שהוא אכן מערך המסננים האופטימלי.

בנספח ג', סעיף 1 מוצגת בפרוט התשובה לשאלה זו. נתאר כאן את עיקרי התוצאות. נגדיר בתור PTFB (Prototype Translated Filter Bank), כל מערך מסננים בן N מסננים, שתגובת התדר של המסנן ה-i-י בו, נוצרת על ידי הזזה ב-(i/N) של תגובת תדר כלשהי $H_0(f)$ ב"ת ב-i. נתייחס אזי ל-$H_0(f)$ כאל תגובת התדר של מסנן האב-טיפוס של המערך. מערכי PTFB הם בדיוק המערכים שניתן לממש ביעילות על ידי DFT.

נגדיר בתור CUFB (Complex Uniform Filter Bank) את מערכי המסננים בני N מסננים, המאופיינים על ידי חמשת הדרישות התכנוניות הבאות:

(א) תגובת התדר הרצויה של המסנן ה-i -י, היא הזזה ב-(i/N) של תגובת תדר רצויה בסיסית כלשהי שתסומן על ידי $D_0(f)$.

(ב) התגובה הכוללת הרצויה של המערך היא אינווריאנטית להזזת תדר של (1/N).

(ג) נורמות השקלול של השגיאות בתגובות התדר של המסננים במערך, הן הזזות ב-(i/N) של אותה נורמת שקלול בסיסית שתסומן על ידי $\|\cdot\|_0$.

(ד) נורמת השקלול של שגיאת התגובה הכוללת של המערך היא אינווריאנטית להזזת תדר של (1/N).

(ה) שקלול העוותים הנוצרים ב-N מסנני המערך, הוא אינווריאנטי לפרמוטציה שלהם (קרי, אין דרישה של קירוב טוב במיוחד במסנן מסויים).

ניתן להוכיח (ראה בנספח ג', סעיף 1) שעבור מערכי מסנני FIR שחולזיותיהם הבסיסיות הם אלמנטי השהייה (כבציור 3.2), לכל בעית תכנון של מערך CUFB ישנו פתרון אופטימלי אחד לפחות שהוא מערך PTFB.

תוצאה זו מתארת תנאים מספיקים לאופטימליות של מערכי PTFB ולא תנאים הכרחיים. במקרים מסוימים לבעיות תכנון שאינן CUFB, קיים פתרון אופטימלי שהוא מערך PTFB, אך בדרך כלל הפתרון האופטימלי לא יהא מערך PTFB.

כפועל יוצא מהמשפט שצוטט לעיל, ניתן להגדיר את מסנן אב-הטיפוס האופטימלי כמסנן בעל תגובת התדר $H_0(f)$ שהוא קירוב אופטימלי תחת $\|\cdot\|_0$ של התגובה הרצויה $D_0(f)$, בכפוף לאילוצי התגובה הכוללת של המערך שמבוטאים בתלות במקדמי מסנן האב-טיפוס בלבד.

בעית הקירוב האקויולנטית מתוארת בפרוט בנספח ג', סעיף 1, והיא שימושית לתכנון PTFB אופטימלי מאחר ומורידה בפקטור N את מספר הנעלמים בתכנון.

עבור תכנון בשיטת WMMSE של מערכי מסננים עם תגובה כוללת מוכתבת פיתחנו בסעיף 3.1 (ובנספח א'), תנאים מספיקים לממשיות של מקדמי מערך המסננים האופטימלי, ולפזה לינארית של המסננים בו. תנאים אלה התבססו על הנוסחאות המפורשות שמתארות את המערך האופטימלי על פי קריטריון זה.

עבור קריטריוני תכנון כלליים, לכאורה לא ברור כלל האם קיים מערך מסננים אופטימלי, ועל אחת כמה וכמה תנאים לממשיות המקדמים ופזה לינארית של המסננים שבו. בעיה זו היא שהוותה מוטיבציה לניתוח המתמטי המתואר בנספח ב'. כפי שכבר תואר בסעיף 3.2, ניתן להסיק מניתוח זה תנאים מספיקים לקיום מערך מסננים אופטימלי ולבדוק מספר תכונות שלו כגון: ממשיות המקדמים, ופזה לינארית של המסננים. תוצאות אלה תופסות כמובן גם למקרה של מערך מסננים אחיד ולכן לא יתוארו כאן בשנית.

ראוי לציין שבנספח ב', מפותח משפט סימטריה (משפט 6), שמהווה הכללה של המשפט שתואר לעיל (קרי, PTFB <= CUFB), לבעיות קירוב כלליות, במרחבים וקטוריים מסויימים.

4.2 תכנון מערכי מסננים אחידים אופטימליים בקריטריון WMMSE

בסעיף 3.1 (ובפרוט בנספח א'), תוארה שיטת WMMSE לתכנון מערכי מסננים עם תגובה כוללת מוכתבת. השיטה פותחה עבור מערכי מסננים כלליים, ולמרות הניצול של תכונות בעית האופטימיזציה על מנת להוריד את סיבוכיות התכנון ככל האפשר, הרי התוצאה הסופית, במקרה הכללי, היא בעלת סיבוכיות לא זניחה כלל ועיקר.

לאור המשפטים שתוארו בסעיף הקודם, עבור בעיות תכנון של מערכי CUFB, ניתן לחסוך בפקטור N (ואף יותר) על ידי תכנון מסנן האב-טיפוס האופטימלי של מערך ה-PTFB, במקום תכנון ישיר של כל N מסנני המערך.

חשוב להדגיש שחסכון זה בסיבוכיות אינו גורע מביצועי המערך המתקבל.

בנספח ג', סעיף II, מתואר האפיון של מערך CUFB עבור קריטריון WMMSE (מקרה פרטי של ההגדרה הכללית של מערך CUFB שתוארה לעיל), ומפותח אלגוריתם לתכנון מסנן אב-הטיפוס האופטימלי למערך כזה.

כמו במקרה הכללי (שתואר בהרחבה בסעיף 3.1), מפותחת תחילה שיטת התכנון עבור המקרה שבו שגיאת התגובה הכוללת מוכפלת בקבוע משקול K^2 ונוספת למדד השגיאה של WMMSE עבור מסנן אב-הטיפוס.

כאשר שגיאת התגובה הכוללת צריכה להוות קטנה מגודל מסוים (η^2) המוכתב על ידי המתכנן, הרי האלגוריתם לקביעת $K^2(\eta^2)$ זהה לזה שתואר בסעיף 3.1, למעט הקטנת ממדי כל הוקטורים והמטריצות בפקטור של N ולכן לא יתואר כאן בשנית.

גם אלגוריתם התכנון למקרה שבו השגיאה בתגובה הכוללת מהווה חלק ממדד השגיאה, מפותח על פי אותם עקרונות כמו האלגוריתם הכללי שתואר בסעיף 3.1, ולכן לא יתואר כאן בפרוט. עם זאת תאור מדויק ומפורט שלו שמאפשר מימוש במחשב ניתן בנספח ג', סעיף II.

נפרט כאן את החסכון המושג בסיבוכיות התכנון:

עבור מערך מסננים אחיד המכיל N מסנני FIR, שכל אחד מהם בעל M מקדמים . אלגוריתם התכנון הכללי מחייב היפוך של N מטריצות טואפליץ ממימד M ולכסון משותף של שתי מטריצות הרמיטיות ממימד M. לכן סיבוכיותו היא של $O(NM^2 + \alpha M^3)$ פעולות (כש-$\alpha > 1$, אך אינו גדול בהרבה מ-1).

מאידך, האלגוריתם עבור מערכי מסננים אחידים שמפותח בנספח ג', מחייב היפוך מטריצת טואפליץ אחת ממימד M ולכסון משותף של שתי מטריצות הרמיטיות ממימד (M/N). לכן סיבוכיותו היא של $O(M^2 + \alpha(M/N)^3)$ פעולות.

החסכון האופייני הוא בפקטור של Max(M,N), שערכו הטיפוסי הוא כ-100.

עד כה דנו במערכי CUFB. בשימושים רבים נדרש שמוצא מערך המסננים יכיל אותות ממשיים, ובו זמנית מעוניינים לשמור על הצורה של מערכי PTFB שניתנים למימוש יעיל. הדרך המקובלת להבטיח תכונות אלו היא על ידי סיכום זוגות מתאימים של יציאות של ה-PTFB (ראה [6]).

בדומה לאפיון של בעית תכנון CUFB נאפיין בעית תכנון RUFB (Real Uniform Filter Bank) כדלקמן:

(א) תגובת התדר הרצויה של המסנן ה-i-י היא סיכום של הזזות ב-$\pm(i/2N)$ של תגובת תדר רצויה שתסומן על ידי $D_0(f)$.

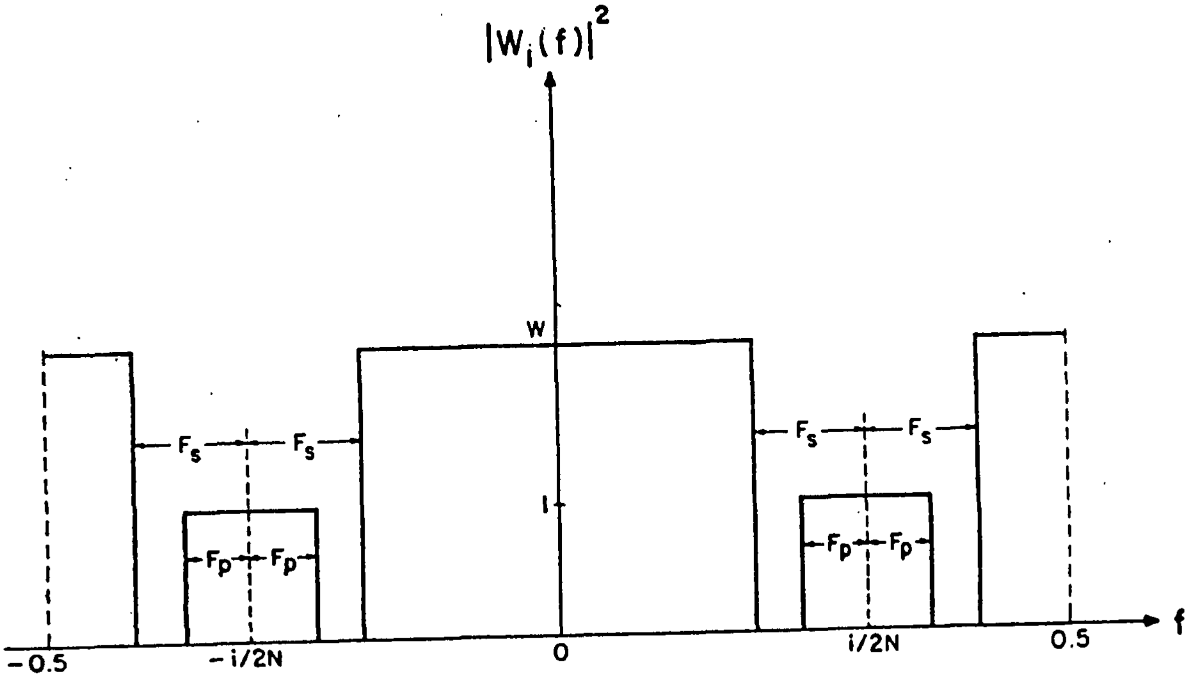
(ב), (ד) ו-(ה) זהים לאפיון של CUFB.

(ג) נורמות השקלול של השגיאות בתגובות התדר של המסננים במערך, מקבלות ערך זהה עבור הזזת תגובות התדר ב- (i/2N)+ וב-(i/2N)-.

בניגוד למקרה של CUFB, הפתרון האופטימלי של בעית תכנון RUFB לא יהא בדרך כלל מהצורה של PTFB. לפיכך הדרישה שמערך מסננים אחיד לבעית RUFB יהא PTFB משמעה תכנון תת-אופטימלי. עם זאת לאור המימוש היעיל של מערכים כאלה, יתכן שזו תהא דרישה תכנונית.

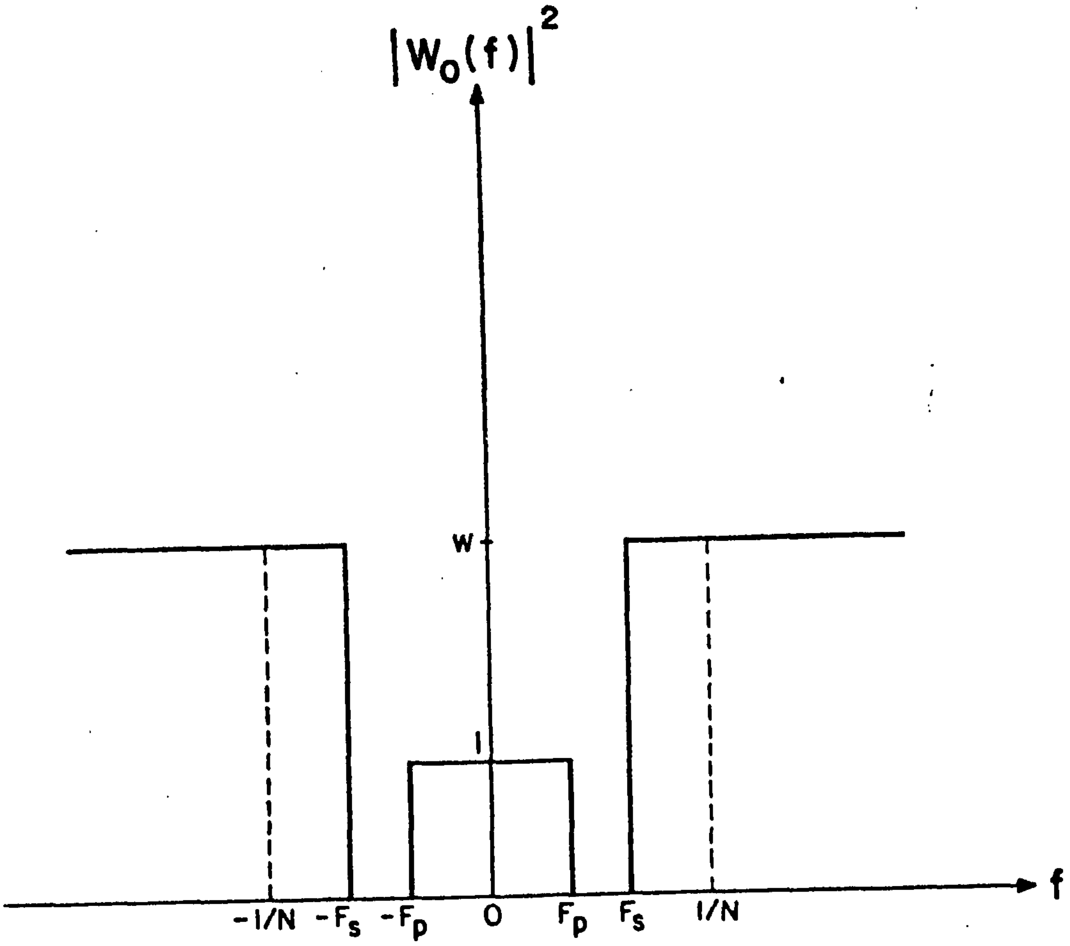
עבור מקרה זה פיתחנו אלגוריתם לתכנון מסנן אב-הטיפוס האופטימלי בקריטריון WMMSE, המתואר בפרוט בנספח ג', סעיף III. אלגוריתם זה זהה במבנהו לאלגוריתם לתכנון מסנן אב-הטיפוס האופטימלי לבעית תכנון CUFB אופטימלי בקריטריון WMMSE.

מאחר ומדדי השגיאה של RUFB ו-CUFB בדרך כלל שונים זה מזה, האברים של מספר מטריצות ווקטורים המופיעים באלגוריתם לתכנון המסנן האופטימלי שונים במקרה של RUFB. שינויים אלו מתוארים בפרוט בנספח ג' סעיף III עבור פונקציות המשקל הטיפוסיות המתוארות בציור 4.1.



ציור מס' 4.1a: פונקצית משקל טיפוסית עבור המסנן ה-i במערך מסננים ממשי אחיד (RUFB).

Fig. 4.1a: A typical weight function of the i-th filter of an RUFB.



ציור מס' 4.1b: פונקצית המשקל המתאימה למסנן אב-הטיפוס של מערך מסננים אחיד, בן N מסננים.

Fig. 4.1b: The weight function of the CUFB prototype corresponding to Fig.1a.

4.3 תכנון מערכי מסננים אחידים אופטימליים בקריטריון Min-Max

תכנון מערכי מסננים אופטימליים בקריטריון Min-Max, תחת אילוץ תגובה בוללת מוכתבת מחייב שימוש בטכניקות של תכנות לינארי, כפי שתואר בסקר המקורות (סעיף 2.2).

גם בשיטות שפותחו במיוחד לתכנון מסנני FIR (ראה [26]), בעית התכנון של מערך מסננים טיפוסי היא בעלת סיבוכיות נכבדה מאד, וזמן המחשב הדרוש הוא בסדרי גודל של שעות CPU על מחשב דוגמת מחשב IBM-370.

עבור בעיות תכנון של מערכי מסננים אחידים, קרי מערכי CUFB, ניתן להשתמש במשפטים שתוארו בסעיף 4.1, על מנת לתכנן את המערך האופטימלי בקריטריון Min-Max, בעזרת תכנון מסנן אב-טיפוס אופטימלי ושימוש בו ליצירת PTFB אופטימלי. בדומה לנאמר בסעיף 4.2, גישה זו חוסכת בפקטור N לפחות בסיבוכיות התכנון ללא פגיעה בביצועי המערך המתקבל.

בנספח ג', סעיף IV מנוסחת במדויק בעית התכנון המתאימה למסנן אב-הטיפוס האופטימלי בקריטריון Min-Max. גם בעיה זו מחייבת שימוש בתכנות לינארי אך נעשית רדוקציה של מס' הנעלמים והאילוצים ביחס של N.

במקרה הפרטי של תגובה כוללת מוכתבת שהיא תגובת יחידה, מוראה שם שבעית התכנון של המסנן אב-הטיפוס האופטימלי, זהה לזו שהוצגה ב-[40] עבור תכנון מסנני אינטרפולציה. במקרה הכללי יותר נוצרת בעית תכנון אחרת, אך בעלת סיבוכיות דומה.

עבור מסננים קצרים (באורך טיפוסי של מאה מקדמים או פחות), גישת התכנות הלינארי מאפשרת קבלת פתרון אופטימלי תוך זמן סביר (ראה [26,40]), אך עבור מסננים ארוכים מתעוררות עדיין בעיות עקב סיבוכיות התכנון.

4.4 שיטת ה"חלון" המוכללת ושימוש ב"חלון" אופטימלי לתכנון המערכים

בסעיף 2.1 תוארה בקצרה שיטת ה"חלון" לתכנון מסנן FIR ספרתי יחיד (פירוט ראה ב-[16,21,22]), ובסעיף 2.2 מתואר השימוש בה לתכנון מערכי מסננים בעלי תגובה כוללת שהיא תגובת יחידה (בהתבסס על [5,37]).

יתרונה העיקרי של שיטה זו הוא בפשטותה, אך חסרונה, הוא בכך שבדרך כלל התגובה המתקבלת אינה אופטימלית על פי קריטריון Min-Max.

נשאלת השאלה האם ניתן בעזרת שיטת ה"חלון" לתכנן את מסנן האב-טיפוס האופטימלי (בקריטריון Min-Max), עבור בעיות תכנון CUFB. כלומר, האם קיים "חלון" אופטימלי שבשימוש בו מתקבל בשיטת ה"חלון" המסנן אב-הטיפוס האופטימלי.

לא לכל בעית תכנון CUFB התשובה לשאלה זו חיובית. אולם, כאשר נדרשת תגובה כוללת שהיא תגובת יחידה, ותגובת התדר הרצויה של מסנן אב הטיפוס $D_0(f)$ היא של LPF אידיאלי בעל תדר קטעון (1/2N), הרי ניתן להוכיח שקיים "חלון" אופטימלי כאמור לעיל. ההוכחה המפורטת מופיעה בנספח ג', סעיף V (משפט 3, שם).

מאחר ומקרה זה הוא הנפוץ ביותר בתכנון מערכי מסננים אחידים נצטמצם לדיון בו ונפתח שיטה פשוטה לתכנון ה"חלון" האופטימלי.

לאור משפט 3 בנספח, ניתן להציג בעית אופטימיזציה אקויולנטית שפתרונה הוא ה"חלון" האופטימלי. הצגה זו עדיין אינה מקטינה את סיבוכיות התכנון, שכן תכנון ה"חלון" האופטימלי מחייב שימוש באותן טכניקות של תכנות לינארי.

על מנת לפשט את סיבוכיות התכנון חיפשנו קירוב של ה"חלון" האופטימלי, שאותו ניתן לתכנן ללא צורך בתכנות לינארי. לשם כך ביטאנו את הפונקציה שיש למזער (שגיאת תגובת התדר של מסנן אב-הטיפוס המתקבל שמסומן על ידי δ_ω), בתלות בערכי הפונקציה $J_\omega(f)$ (שהיא האינטגרל של תגובת-התדר של מקדמי ה"חלון" במספר תדרים).

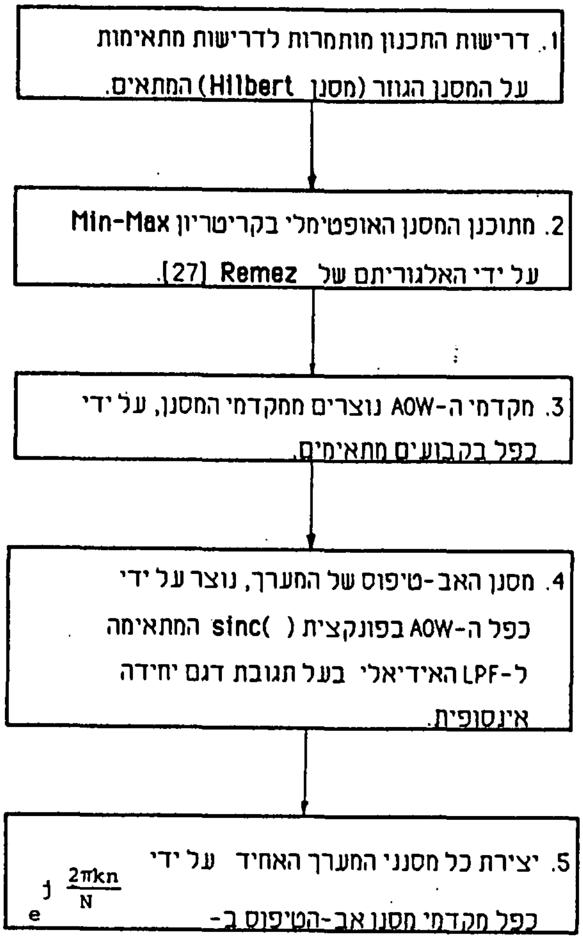
תכונותיה של $J_\omega(f)$ נחקרו על-ידינו בפירוט רב ב-[22], ועל פיהן הצענו קירוב של δ_ω , על ידי שקלול אחד של ערכי $J_\omega(f)$ בתדרים הנדונים (נסמנו להלן ב- $\hat{\delta}_\omega$).
הביטויים המדוייקים של δ_ω ו- $\hat{\delta}_\omega$ ופיתוחם מתוארים בנספח ג', סעיף ץ.
נעיר כאן מספר הערות על טיב הקירוב של δ_ω על ידי $\hat{\delta}_\omega$, כדלקמן:

(א) ניתן להוכיח שלכל "חלון" $\delta_\omega \leq 2\hat{\delta}_\omega$ (זה נובע מיידית, מצורת הקירוב).

(ב) קירוב דומה שימש אותנו ב-[22] למטרה אחרת, ושם קיבלנו (נסיונית), שעבור כל ה"חלונות" המקובלים מתקיים ש- $\hat{\delta}_\omega \geq \delta_\omega$. ניתן גם לתת הנמקה היוריסטית לתכונה זו, אם כי אין לה הוכחה אנליטית.

נגדיר את ה-AOW (Approximate Optimal Window) כסדרת המקדמים שמביאה למינימום של $\hat{\delta}_\omega$, על פני כל הסדרות בעלות האורך M הנתון. לאור תכונות (א') ו-(ב') לעיל, נצפה ש- δ_ω עבור ה-AOW יהא קטן מפעמיים הערך המינימלי של δ_ω , כלומר שביצועי ה-AOW שונים מביצועי ה"חלון" האופטימלי ב-3dB לכל היותר. מאחר וערכים אופייניים של δ_ω הם -40dB - -80dB הרי שדגרדציה של 3dB בביצועים נחשבת בדרך כלל זניחה. לפיכך ה-AOW מהווה קירוב טוב של ה"חלון" האופטימלי.
נותר לבדוק שקיימת דרך פשוטה לתכנן את ה-AOW ללא שימוש בתכנות לינארי.

לשם כך רשמנו במפורש את בעית המינימיזציה של $\hat{\delta}_\omega$ בתלות במקדמי ה"חלון" והצגנו בעיה זו כך שניתן לפותרה באמצעות האלגוריתם של Remez. יתר על כן, הבאנו אותה לצורה המאפשרת שימוש בתוכנית המחשב הנפוצה שב-[27], ולא מצריכה תוכנה מיוחדת. פיתוח זה, וניסוח אלגוריתם התכנון, מופיעים בנספח ג', סעיף ץ, ואילו תאור סכמטי שלו (דיאגרמת בלוקים) מופיע בציור 4.2.



<u>ציור מס' 4.2</u>: תאור סכמטי של האלגוריתם לתכנון מערכי CUFB על ידי ה-AOW.

<u>Fig. 4.2</u>: Schematic description of the algorithm for design of CUFB filter banks by the AOW.

ניתן להשתמש ב-AOW גם כשיטה לתכנון מערכי מסננים כלליים (לאו דוקא CUFB) עם תגובה כוללת שהיא תגובת יחידה. במקרה הכללי זהו קירוב עם הנמקה היוריסטית למערך האופטימלי בקריטריון Min-Max, אך ללא הקשרים האנליטיים שהוצגו כאן עבור מערכי CUFB. יתר על כן, מאחר ותכנון מסנן אב-הטיפוס למערך CUFB אקויולנטי (מבחינה מתמטית) למספר בעיות תכנון של מסנני אינטרפולציה לשימושי תקשורת ספרתית שהוצגו ב-[40], הרי שה-AOW יכול לשמש כפתרון מקורב גם עבור בעיות תכנון אלו. מאחר והכללה זו פשוטה יחסית, ואינה קשורה לנושא העבודה לא נפרט בנושא זה מעבר לאמור לעיל.

את ביצועי התכנון בעזרת ה-AOW השווינו לשיטות המקובלות (שנסקרו בסעיף 2.2) על ידי דוגמאות תכנון טיפוסית.

מאחר ולא היתה בידינו התכנה של האלגוריתמים שתוארו ב-[38] וב-[40] השווינו את ה-AOW לשתי שיטות אלו רק על בסיס אותן דוגמאות תכנון שמתוארות שם.

על-פי הגדרת שיטת ה-AOW היא מביאה לתכנון מסנן אב-טיפוס שמקרב את המסנן האופטימלי בקריטריון Min-Max, אותו ניתן לתכנן רק על ידי תכנות לינארי כפי שמוצג ב-[40]. לפיכך ביצועי המסננים המתוארים שם מייצגים את הערך המינימלי של δ_ω (שגיאת תגובת התדר של המסנן האופטימלי), ועבור ה-AOW יתקבל תמיד ערך גדול מערך זה.

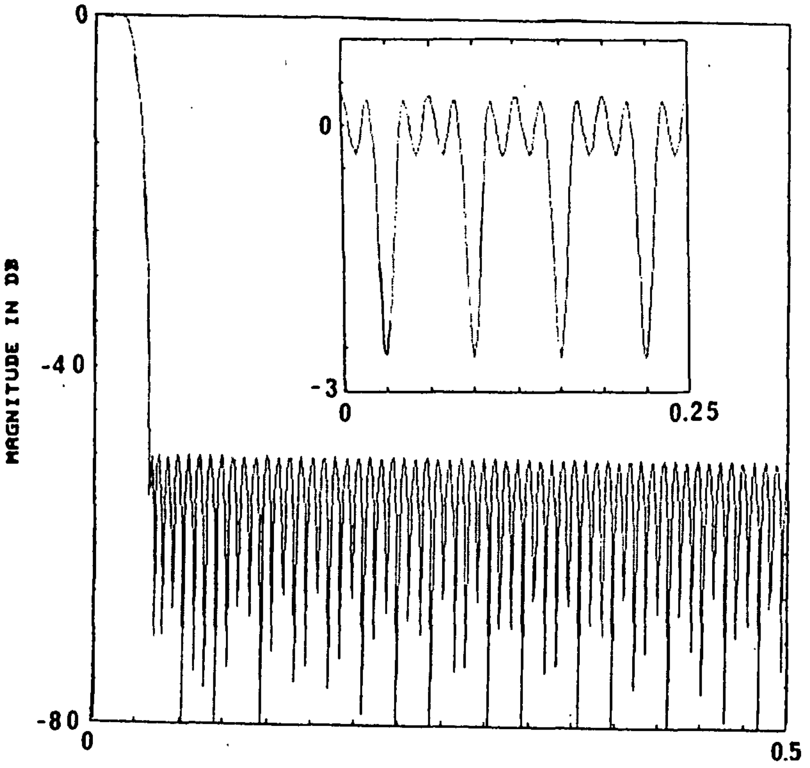
בהשוואה לגבי דוגמה טיפוסית (של מסננים באורך 39 מקדמים) התקבל ניחות של 32dB ב-AOW לעומת ניחות של 33dB במסנן האופטימלי וכצפוי הדגרדציה זניחה. תוצאות דומות התקבלו בהשוואה לדוגמאות שמתוארות ב-[38], קרי - ביצועי ה-AOW נחותים מאלו המדווחים שם, אך באופן לא משמעותי (ראוי לציין שאלו דוגמאות של מערכי מסננים לא-אחידים).

הדוגמאות המתוארות ב-[38,40] הן של מסננים קצרים בלבד, כנראה לאור הסיבוכיות של אלגוריתמים אלו.

סיבוכיות התכנון בשיטת ה-AOW, שקולה לזו של התכנה ב-[27], וידוע שניתן בקושי לא רב לתכנן בעזרתה מסננים באורך של כ-512 מקדמים.

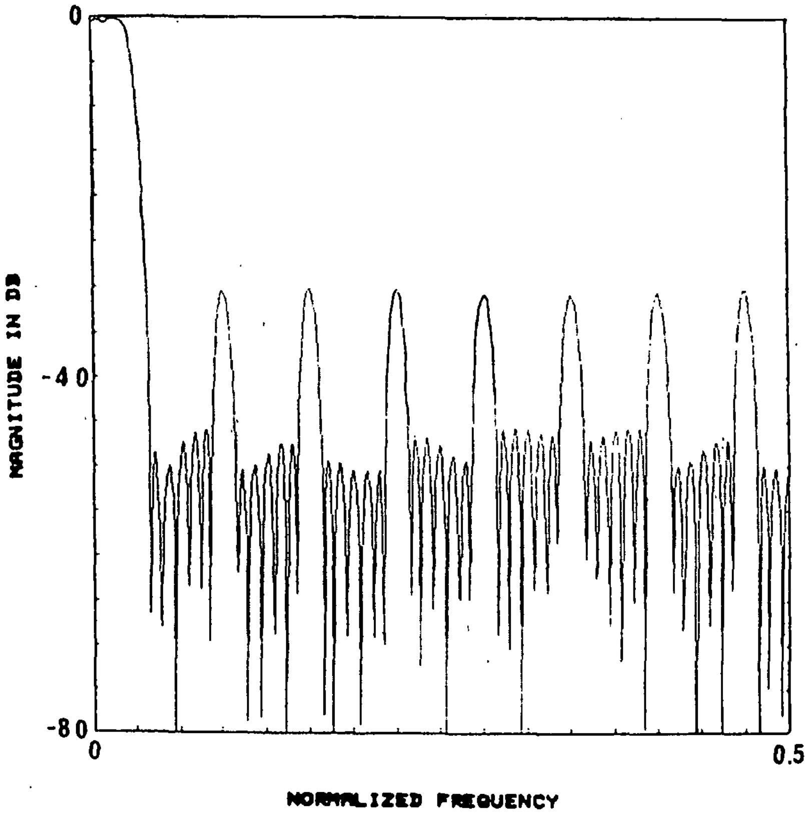
עבור דוגמה טיפוסית של מערך 16 מסננים, שכל אחד מהם ארכו 123 מקדמים השווינו את התכנון על ה-AOW לשיטות הלא-אופטימליות שתוארו ב-[37,39], ולתכנון ללא הכתבת התגובה-הכוללת.

ציורים 4.3 - 4.6 מתארים את תגובות התדר של ארבעת מסנני אב-הטיפוס המתקבלים.



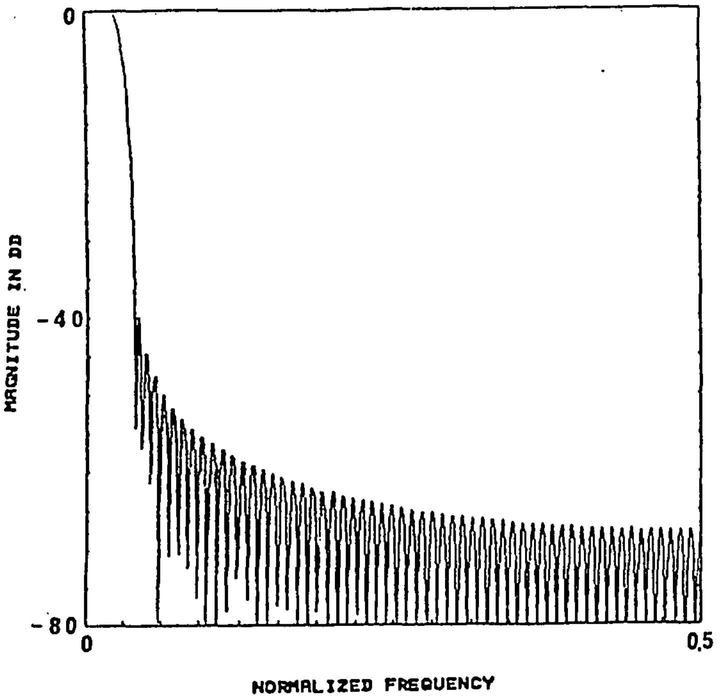
ציור 4.3: תגובת התדר של מסנן אב-טיפוס אופטימלי שתוכנן ללא דרישות על התגובה הכוללת, והתגובה הכוללת של המערך האחיד המתקבל.

Fig. 4.3: Frequency response of the optimal lowpass prototype filter, and the resulting composite frequency response when there is no constraint of the composite response.



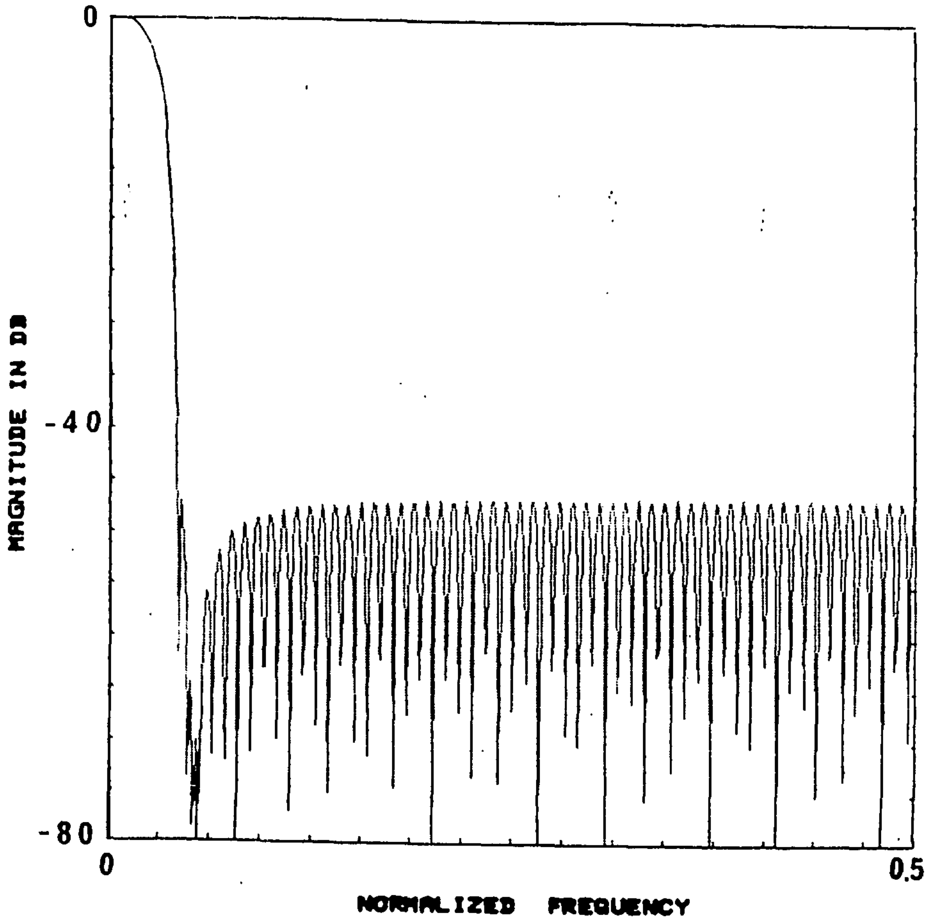
ציור 4.4: תגובת התדר של מסנן אב-טיפוס שתוכנן בשיטה המתוארת ב-[39], עבור אותן דרישות כמו בציור 4.3 למעט אילוץ תגובה כוללת שהיא תגובת יחידה.

Fig. 4.4: Frequency response of the lowpass prototype filter using the method of [39], for the same specifications as in fig. 4.3 with additional constraint of unity composite response.



ציור מס' 4.5: תגובת התדר של המסנן שתוכנן בשיטה המתוארת ב-[37], עבור אותן הדרישות כמו בציור 4.4.

Fig. 4.5: Frequency response of the lowpass prototype filter using the method of [37] for these specifications.



ציור מס' 4.6: תגובת התדר של המסנן שתוכנן בשיטת AOW המוצעת כאן, עבור אותן דרישות.

Fig. 4.6: Frequency response of the lowpass prototype filter using the AOW method for these specifications.

בבירור ה-AOW מביא למסנן טוב יותר מאלו המתקבלים בשיטות של [37,39]. המסנן המתקבל ללא הכתבת התגובה הכוללת הוא בעל ניחות של 50dB בעוד ה-AOW מביא לניחות של 46dB בלבד, אך כמוראה בציור 4.3 התשלום על הניחות הנוסף של 4dB הוא בתגובה כוללת גרועה של המערך (סטיות של ±3dB מהערך הרצוי).

תאור מפורט של דוגמאת התכנון מופיע בנספח ג', סעיף V, והשוואת התוצאות רוכזה שם, בטבלה C.1.

פרק 5 : תכנון מערכות אנליזה-סינתזה אופטימליות הכוללות כימות (Quantization)

5.1 תאור המודל הסטטיסטי ומדדי השגיאה

כל השיטות לתכנון מערכות אנליזה-סינתזה (A/S) שנסקרו בסעיף 2.3, מבוססות על הפרדה בין תכנון המסננים שבמערכת ובין המודיפיקציה הנעשית בתוכה. בדרך זו המסננים המתקבלים הם אוניברסליים במובן זה שהם תלויים רק בפרמטרי המערכת (הגדלים של M,R, וארכי המסננים) ולא בסוג המודיפיקציה שנעשית בתוכה. אולם אין כל ודאות שמסננים שתוכננו תחת קריטריון אופטימיזציה כלשהו בהזנחת המודיפיקציה, יהיו אכן אופטימליים בתוך המערכת המכילה מודיפיקציה.

לפיכך נקטנו בגישה שונה לתכנון המסננים במערכות A/S והיא לתכנן מסננים אופטימליים עבור מודיפיקציה נתונה. בפרט התמקדנו במודיפיקציות של כימות (קוונטיזציה) שבהן, משיקולי קצב השידור המוגבל בערוץ, נדרש שימוש ב-R = M וכן לטיב המסננים יש השפעה רבה על ביצועי המערכת.

על מנת לתכנן מסננים אופטימליים למערכת A/S המכילה כימות, נדרש ראשית לאפיין באמצעות מודל סטטיסטי את פעולת הכימות, ושנית להגדיר את מדד השגיאה שמעונינים למזער.

הנחת היסוד שלנו היא שמערכת ה-A/S פועלת כמקודד צורת-גל ולכן מדד השגיאה יהא וריאציה כלשהי (שתתואר בהמשך) של מדד MSE. הנחה זו סבירה בהחלט שכן השימוש העיקרי של מערכות A/S הוא בקידוד בקצב בינוני (9.6-32Kbps) ומטרתן להשיג איכות דיבור טובה, תוך שמירה על קצב נמוך בכל האפשר.

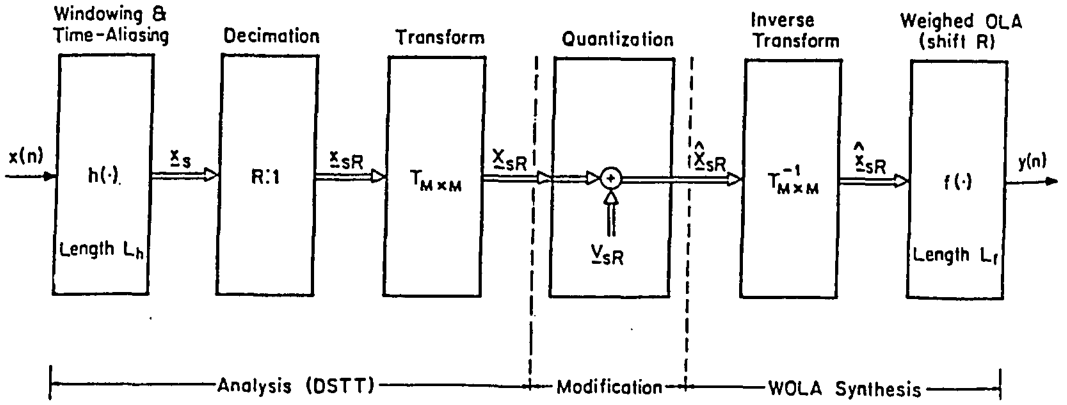
מערכות A/S תוארו כבר בקצרה בסעיף 1.1, ולכן לא נרחיב כאן בתאורן. נציין שבנספח ד', סעיף 1 ניתנות המשוואות המתמטיות שמתארות את פעולת האנליזה ואת פעולת הסינתזה בשיטת WOLA שתוארה בסעיפים 1.1 ו-2.3 (לפרוט נוסף, ראה ב-[6]).

נציין בנוסף שעל מנת לכלול באותה מסגרת את מערכות ה-A/S הפועלות עם טרנספורמים שונים (למשל: DFT, DCT, טרנספורם הדמרד), הכללנו שם את תאור מערכת ה-A/S המקובלת (שבה הטרנספורם הוא DFT או GDFT כמתואר ב-[6]), כפי שכבר תואר בציור 1.7.

שני מודלים שונים שימשו אותנו עבור פעולת הכימות.

המודל הראשון הקרוי כימות עדין (FQ) מתאים לכימות בנפרד של כל דגם בוקטור המוצא של שלב האנליזה (וקטור ה-DSTT הכולל M דגמים). כשהכימות נעשה עם מספר מספיק גדול של סיביות לדגם (באופן טיפוסי כ-4 סיביות ויותר). מקרה זה מתאים

למשל למערכות A/S המכילות בתוכן כימות מסוג RCM (Pulse Code Modulation) או DPCM (Differential Pulse Code Modulation) עם הקצאת סיביות (ורמות החלטה וייצוג) קבועה או משתנה לאט. במקרה זה (המתאים למערכות A/S לקצב של 16Kbps ומעלה) מקובל לתאר את פעולת הכימות על ידי תוספת של רעש אדיטיבי בלתי-תלוי סטטיסטית באות (ראה למשל ב-[66]). מודל זה בו השתמשנו עבור FQ, מתואר סכמטית בציור מס' 5.1 כאשר ההנחה היא שהן האות בכניסה והן הרעש הם ת"א סטציונריים במובן-הרחב (למשך זמן קצר מספיק), בעלי ממוצע אפס וסדרת קווריאנס ידועה (מדודה).



ציור מס' 5.1: מערכת אנליזה-סינתזה עם כימות עדין (FQ).

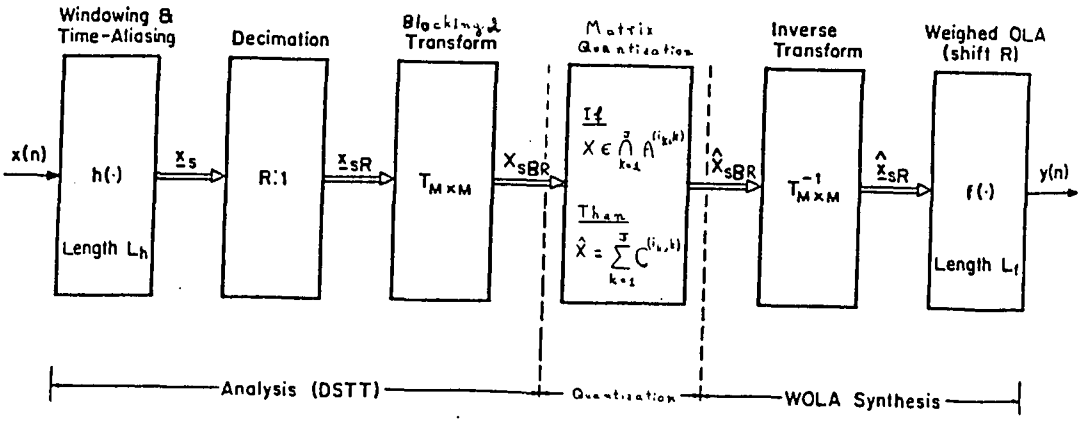
Fig. 5.1: Analysis/synthesis system with fine quantization (FQ).

המודל השני הוא מודל של כימות מטריצי (MQ) הנעשה כדלקמן: בלוק של $B \geq 1$ וקטורי DSTT עוקבים מאוגד למטריצה בת $M \times B$ מספרים קומפלקסיים (ממשיים). את מרחב כל המטריצות ממימד $M \times B$, מחלקים ל-L תחומים הקרויים אזורי החלטה. אם המטריצה הנוכחית שייכת לאזור ההחלטה ה-i-י, משדרים את האינדקס i ובמקלט משתמשים לצורך הסינתזה במטריצה C_i במימד $M \times B$, המהווה את האבר המייצג של אזור ההחלטה ה-i-י. הסינתזה בשיטת WOLA נעשית על סדרה משורשרת של מטריצות כאלה, בהתאם לאות במוצא שלב האנליזה. גישה זו סבירה עבור L קטן מספיק, כשבאופן טיפוסי $L \leq 1024$ ו-$B = 1$ עבור הכימות הוקטורי (VQ) המקובל.

כאשר נדרש ערך גדול של L, על מנת להשיג שחזור מאיכות גבוהה, הרי ישנם קשיים בבנית ספר-הקוד (קרי בחירת אזורי ההחלטה והאברים המייצגים), ומקובל אזי להשתמש בקוד מכפלה (Product Code) שהוא אמנם תת-אופטימלי אך פשוט יותר למימוש.

קידוד בקוד-מכפלה נעשה על ידי יצירת J ספרי-קוד נפרדים באורכים $\{L_k\}_{k=1}^{J}$, כך שלכל ספר-קוד יש את אזורי ההחלטה והאברים המייצגים שלו. בהתאם למטריצה הנוכחית שבמוצא שלב האנליזה בוחרים אינדקס בכל אחד ואחד מספרי-הקוד באופן בלתי-תלוי ומשדרים את וקטור J האינדקסים $\underline{i}$. במקלט שולפים את המטריצה המייצגת המתאימה בכל ספר-קוד, על פני האינדקס ששודר עבור ספר הקוד הזה, ולצורך הסינתזה משתמשים במטריצה $C^{(\underline{i})}$ שהיא הסכום של J מטריצות אלה. לפיכך בעוד שקצב השידור (ומספר המטריצות השונות $C^{(\underline{i})}$) מתאים ל-$L = \prod_{k=1}^{J} L_k$, הרי סיבוכיות הקידוד (סכום גדלי ספרי-הקוד) היא $\hat{L} = \sum_{k=1}^{J} L_k$ ובדרך כלל $\hat{L} \ll L$.

מערכת קידוד-פענוח זו מתוארת סכמתית בציור מס' 5.2.



<u>ציור מס' 5.2</u>: מערכת אנליזה-סינתזה עם כימות מטריצי (MQ).

<u>Fig. 5.2</u>: Analysis-synthesis system with Matrix Quantization (MQ).

שיטות קידוד רבות ניתנות לניסוח כמקרים פרטיים של מודל זה. נציין כאן ארבע שיטות:

(א) כימות וקטורי אנכי - במקרה זה $J = B = 1$ ומשתמשים בספר-קוד עבור וקטורים של DSTT (למשל, ראה ב-[64]).

(ב) כימות סקלרי עם הקצאת סיביות קבועה - זהו מקרה של קוד-מכפלה לכימות וקטורי אנכי, $B = 1$ ו-$J = M-1$ (או $J = M/2$ עבור ה-DFT, אך עם מילון קומפלקסי), כשישנו ספר-קוד נפרד לכל דגם ב-DSTT (ספר-קוד זה מייצג את המכמת הסקלרי של דגם זה), דוגמא למערכת כזו ראה ב-[65].

(ג) כימות וקטורי אופקי - במקרה זה $J = M$ (או $J = M/2$ עבור ה-DFT), כש-$B > 1$ וכל ספר קוד מתאים לוקטורים שנוצרים על ידי לקיחת אותו רכיב מסוים בכל אחד מ-B וקטורי DSTT עוקבים, דוגמא למערכת כזו ראה ב-[70].

(ד) כימות וקטורי בקסקדה - זהו מקרה של קוד-מכפלה לכימות וקטורי אנכי, עם $J \geq 2$, $B=1$. משתמשים ב-J ספרי-קוד. ספר-קוד "גס" בגודל L_1 מופעל על וקטורי ה-DSTT. לאחר מכן מתוכנן ספר קוד "עדין" בגודל L_2 שמוזן על ידי וקטור השגיאה שמתקבל על ידי החסרת הוקטור המייצג המתאים מהספר הראשון מוקטור ה-DSTT, וחוזר חלילה.

עד כאן תאור שיטת הקידוד המטריצית. המודל הסטטיסטי שמניחים עבורה הוא כדלקמן: מערכת הסינתזה מוזנת על ידי מקור בעל אלף-בית סופי בן L אותיות. כל אות מקור היא J-יה סדורה, כשברכיב ה-k-י שלה יתכנו L_k סמלים שונים. מניחים שידועה (נמדדה) שכיחות ההופעה של כל אחד מ-$\hat{L}$ הסמלים השונים המרכיבים את אותיות המקור, וכן ידועות (נמדדו) שכיחויות ההופעה של זוגות סמלים בכל מרווח-זמן $d \cdot B \cdot R$ נתון (למעשה דרישות השכיחויות רק במרווחי זמן $|d \cdot B \cdot R| \leq L_f$ על מנת לתכנן את מסנן הסינתזה האופטימלי באורך L_f דגמים).

בנוסף לצורך הסינתזה מניחים שידועה (מדידה) התוחלת המותנה של דגמי-הכניסה בזמן $s \cdot B \cdot R + d$, בהנתן שבזמן $s \cdot B \cdot R$ שודר סמל מסוים (מתוך $\hat{L}$ הסמלים השונים) על ידי מנגנון הקידוד (גדלים אלה דרושים רק עבור $|d| \leq L_f$).

מתוך שכיחויות ותוחלות אלה (שבדרך כלל מושגות על בסיס אותה סדרת לימוד שמשמשת לבנית ספרי-הקוד) יחושבו מקדמי מסנן הסינתזה האופטימלי.

תאור מתמטי מפורט של המודלים הסטטיסטיים של שתי שיטות הכימות שתוארו לעיל מופיע בנספח ד', סעיף 1.

מכיון שמערכת A/S היא מערכת משתנה-בזמן (ועבור המקרה של MQ, היא אף הופכת למערכת לא-לינארית), הרי שהשגיאה בין דגמי-המוצא ודגמי הכניסה (שתסומן על ידי $\varepsilon(n)$) היא תהליך לא-סטציונרי. לפיכך, נדרשה הכללה של קריטריון ה-MSE המקובל לתהליכים לא-סטציונריים.

מסתבר שתחת ההנחה שהמכמת הוא בלתי-מוטה (unbiased) הרי $E[\varepsilon(n)] = 0$.

בנוסף עבור שתי שיטות הכימות שתוארו לעיל, הרי $\varphi(d,m) \triangleq E[\varepsilon(m)\ \varepsilon(m+d)]$ היא סדרה מחזורית ב-m לכל ערך של d ועם מחזור סופי $N = RBM/gcd(RB,M)$ (כש-$gcd(\cdot)$ הוא סימון למחלק המשותף הגדול ביותר), ובהנחות לא מגבילות על תהליך הכניסה, הרי $\varphi(d,m)$ חסומה (במידה אחידה ב-d ו-m). הביטויים המפורשים של $\varphi(d,m)$ והוכחת התכונות שצוטטו לעיל מוצגים בנספח ד', סעיף V.

עבור סדרות אקראיות $\varepsilon(n)$ בעלות תוחלת אפס ומומנט שני $\varphi(d,m)$ מחזורי ב-m וחסום, פיתחנו קריטריון טיב שהוא הרחבה של קריטריון ה-MSE המקובל. הקריטריון מוגדר כדלקמן: נחשב את תוחלת הערך המוחלט בריבוע של התמרת-פורייה של קטע סופי $(\ell-r) \leq n \leq (\ell+r)$ של הסדרה $\varepsilon(n)$ וננרמל אותה ביחס לאורך הקטע על ידי חלוקה ב-$(2r+1)$. נכנה גודל זה ה"ספקטרום לזמן סופי" $s_{\ell,r}(f)$, ונחשב את הממוצע המשוקלל (בתדר) שלו, ביחס לפונקצית משקל נתונה $G(f)$ בעלת פירוק לטור-פוריה המתכנס-בהחלט. נסמן גודל זה $\left(\int_{-0.5}^{0.5} s_{\ell,r}(f)G(f)df\right)$ בתור $u_{\ell,r}$ שיהא מדד הטיב "לזמן סופי" של "חלון" המאופיין על ידי הפרמטרים r, ℓ. עתה ניתן להוכיח שלכל ℓ קיים הגבול (הסופי) של $u_{\ell,r}$ כש-$r \to \infty$ וגבול זה (שנסמנו ב-U) בלתי תלוי ב-ℓ. משפט זה, וביטוי מפורש (שניתן לחישוב ישיר) לערך של U מנוסחים בנספח ד', סעיף II, בעוד שהוכחת המשפט מוצגת בנספח ד', סעיף V. במקרה הפרטי שבו אורך המחזור של $\varphi(d,m)$ הוא $N=1$ ולתהליך $\varepsilon(n)$ יש ספקטרום $S(f)$ שהוא פונקציה ב-$[-0.5, 0.5]$-L_2, הרי מדד הטיב המתקבל יהא בדיוק $\int_{-0.5}^{0.5} S(f)\ G(f)df$ (כפי שמוכח שם), שהוא מדד טיב מקובל לת"א סטציונריים במובן הרחב.

גישה אלטרנטיבית להגדרת מדד טיב המערכת מבוססת על חישוב השגיאה בין סדרות ה-DSTT של התהליך בכניסה ושל התהליך ביציאה (זו למעשה הגישה שנקוטה ב-[16,17] אם כי ב-"לבוש דטרמיניסטי", בניגוד ל-"לבוש הסטטיסטי" שנקוט על ידינו כאן). תהא $\underline{E}_{SR}$ סדרת וקטורי השגיאה הנוצרת, אזי $E[\underline{E}_{SR}] = 0$ עבור מכמתים בלתי מוטים. בנוסף, כשהסדרה $\varphi(d,m)$ מחזורית במחזור N במשתנה m, הרי גם סדרת הוריאנסים $v_m = E[||\underline{E}_m||^2]$ מחזורית עם אותו מחזור. עובדות אלו מוכחות בנספח ד', סעיף V. שם, בסעיפים II, V, מוראה גם שמדד הטיב V, שהוא הממוצע ע"פ זמן מחזור אחד (N דגמים) של הסדרה v_m, הוא מקרה פרטי של מדד הטיב U, עבור כל טרנספורם לינארי T שהוא הרכב של טרנספורם יוניטרי וטרנספורם ציקלי (Circulant), (ובפרט זה מתקיים עבור ה-DFT, DCT וטרנספורם הדמרד שהם כולם טרנספורמים יוניטריים).

יתר על כן המדד V מתקבל מתוך U על ידי בחירת פונקצית משקל G(f) מסוימת הקשורה לתגובת התדר של מערך מסנני האנליזה שביטוי מפורש שלה ניתן בנספח ד', סעיפים II, V.

לאור האמור לעיל נמשיך מעתה ואילך את פתרון בעית התכנון על בסיס מדד הטיב U, כשהמדד V מהווה מקרה פרטי שלו.

בתכנון מסנן-אנליזה חשובה גם הפרדת התדר של מערך מסנני האנליזה שנובע ממנו, ולצורך זה ביצענו מודיפיקציה של המדד U, על ידי הוספת ה-MSE (המשוקלל בתדר על ידי פונקציה $W(f) \geq 0$) של שגיאת תגובת התדר של מסנן האנליזה ביחס למסנן רצוי (בדרך כלל LPF אידיאלי) למדד U. את המדד המתקבל נכנה $\hat{U}$ והוא ישמש הן לתכנון מסנן-סינתזה אופטימלי, בהנתן מסנן אנליזה, והן לתכנון של מערכת אנליזה-סינתזה אופטימלית.

5.2 __מסנני סינתזה אופטימליים עבור כימות עדין, ועבור כימות באמצעות ספרי קוד__

קריטריון האופטימיזציה הוא המזעור של המדד $\hat{U}$ ביחס למקדמי מסנן-הסינתזה שאינם ידועים. מאחר והקריטריון $\hat{U}$ הוא תבנית ריבועית מוגדרת אי-שלילית במקדמי מסנן הסינתזה, הרי מסנן הסינתזה האופטימלי מוגדר היטב והוא פתרון של מערכת משוואות לינאריות שמימדה L_f. בנספח ד', סעיפים III, VI מוצגים הבטויים המפורשים עבור מערכת משוואות זו בתלות בפרמטרי מערכת ה-A/S הידועים.

נדון בתכונות הפתרון עבור שתי שיטות הכימות שהצגנו ותחילה המקרה של FQ.

__כימות עדין (FQ)__

במקרה זה מטריצת המקדמים של מערכת המשוואות ניתנת לפרוק לסכום של שתי מטריצות שיסומנו להלן ב-Q_h ו-Q_v. המטריצה Q_v אינה תלויה בסדרת הקווריאנס של האות בכניסה, או במקדמי מסנן-האנליזה, כי אם רק בסדרת הקווריאנס של רעש הכימות. בפרט כשאין כלל כימות הרי $Q_v = 0$. המטריצה Q_h היא בלתי-תלויה ברעש הכימות, אך תלויה במקדמי מסנן האנליזה ובסדרת הקווריאנס של הכניסה.

כפי שמוכח בנספח ד' סעיף IV, הרי כאשר $Q_v = 0$, לכל מערכת יחידה (מערכת המקיימת את תנאי פורטנוף, שמנוסחים במדויק שם ושמשמעותם כבר תוארה בסעיפים 1.1 ו-2.3), אם קיימת כזו (ראה טבלה 2.2), מסנן הסינתזה הוא מסנן סינתזה אופטימלי על פי הגדרתנו. מאידך כאשר $Q_v \neq 0$, תכונה זו בדרך כלל לא תתקיים.

כאשר $R = M$, $L_f, L_h > M-1$ (שהוא המקרה הטיפוסי בשימושים של קידוד), הרי לא קיימת מערכת יחידה בשיטת WOBA. בנספח ד' סעיף VI מוכח שבמקרה זה אם $G(f) = 1$, תהליך הכניסה הוא חסר-קורלציה, ואף אחד מה-polyphases של מסנן האנליזה הוא זהותית אפס (ראה הסבר מדויק שם), מובטחת יחידות של מסנן הסינתזה האופטימלי גם כשאין כלל כימות ($Q_v = 0$).

כשקיים כימות, מובטחת יחידות של מסנן הסינתזה האופטימלי עבור $G(f) = 1$, ובלבד שרעש הכימות הוא תהליך אקראי לא-מנוון (קרי, שלא ניתן בשערוך לינארי ממימד סופי של דגמי התהליך על סמך הדגמים הקודמים שלו להשיג שגיאת שערוך בעלת שונות אפס). הגדרה מתמטית מדויקת של תכונה זו והוכחת היחידות, ראה בנספח ד' סעיפים III, VI.

במקרה של $R > M$, שבו קיימות אינסוף מערכות יחידה (וכן אינסוף מסנני סינתזה אופטימליים למסנן אנליזה נתון כשאין כימות), ניתן גם ללא-כימות (או ליתר דיוק ברעש כימות השואף לאפס) לבחור מסנן סינתזה אופטימלי לסטטיסטיקת רעש אופיינית בתהליך הבא:

נסמן את סדרת (מטריצות) הקווריאנס האופיינית לרעש הכימות (והידועה לנו) על ידי $\hat{\psi}(d)$. לכל רמת רעש $\varepsilon > 0$ קטנה כרצוננו (קרי, לסדרת הקווריאנס של רעש הכימות שהיא $\varepsilon\hat{\psi}(d)$) נתבונן במסנן הסינתזה האופטימלי, שנסמן על ידי $\underline{f}_\varepsilon$. קל להוכיח שתחת תנאים לא מגבילים על $\psi(d)$, קיים מסנן סינתזה אופטימלי יחיד לכל $\varepsilon > 0$. בנספח ד' סעיף IV מוכח שקיים הגבול (הסופי): $\lim_{\varepsilon \to 0} \underline{f}_\varepsilon$ ומוצג בטוי מפורש (שניתן לחישוב ישיר) עבורו. מסנן הסינתזה המתקבל בשיטה זו מבטיח כמובן מערכת יחידה והוא קשור בהפוך המוכלל [67] של המטריצה Q_h כפי שמוראה שם.

נדגיש כאן שמסנן הסינתזה ה"גבולי" שתואר לעיל, תלוי מפורשות בסטטיסטיקת הכניסה ובסטטיסטיקה האופיינית של רעש הכימות.

בשני מקרים פרטיים ניתן לקבל ביטוי מפורש למסנן הסינתזה האופטימלי המתקבל ונתארם כאן בקצרה. תאור מפורט שלהם מופיע בנספח ד', סעיפים III, VI.

(א) $G(f) = 1$, $L_f = L_h = M$, ותהליך הכניסה ורעש הכימות הם תהליכים אקראיים חסרי-קורלציה, אזי מקדמי מסנן הסינתזה האופטימלי ניתנים על ידי:

$$f(n) = \frac{h(M-n)}{\sum_{k=-\infty}^{\infty} h(M-n-kR)^2 + (\frac{\sigma_v}{\sigma_x})^2} \qquad 0 \le n \le M-1 \qquad (5.1)$$

כאשר $\{h(1),...,h(M)\}$ הם מקדמי מסנן האנליזה ו- $(\frac{\sigma_v}{\sigma_x})^2$ הוא היחס בין וריאנס רעש הכימות לווריאנס האות בכניסת מערכת A/S (שקשור למספר הסיביות שבכימות דגמי ה-DSTT, כמתואר למשל, ב-[66]). תוצאה זו היא מעין הכללה של התוצאות ב-[16,17] שמתאימות לנוסחא (5.1) עבור $\sigma_v = 0$.

(ב) $G(f) = 1$, $R = M$. במקרה זה מערכת המשוואות הלינאריות הקובעות את L_f מקדמי מסנן הסינתזה האופטימלי, מתפרקת ל-M תת-מערכות משוואות. כל תת-מערכת משוואות קובעת את המקדמים של polyphase אחר של מסנן הסינתזה, בתלות ב-polyphase המתאים של מסנן האנליזה. בניגוד למקרה (א) לעיל לתת-מערכות משוואות אלה אין פתרון אנליטי (פרט למקרה של $L_f = M$), אך עבור $L_f = \infty$ יש להן אינטרפרטציה בתחום התדר שמזכירה את מסנן Wiener הדיסקרטי ומתוארת בנספח ד', סעיף III.

הסיבוכיות של פתרון מערכת המשוואות הלינארית במקרים שונים והאינפורמציה הסטטיסטית הדרושה (מספר הדגמים של סדרות הקווריאנס של רעש-הכימות ותהליך הכניסה המשפיעים על מקדמי מסנן הסינתזה האופטימלי) מתוארים בפרוט בנספח ד', סעיף VI ומתומצתים בטבלה D.1.

<u>כימות מטריצי (MQ)</u>

במקרה זה עקב האי-לינאריות של תהליך הכימות, הרי מטריצת המקדמים במערכת המשוואות הלינאריות אינה ניתנת לפרוק מהצורה $Q_v + Q_h$ כבמקרה של FQ. יתר על כן, אין בטוי מפורש לתלות של מסנן הסינתזה האופטימלי במקדמי מסנן האנליזה הנתון, ומאחר ויש רק L מטריצות שונות $C^{(i)}$ שבהן מוזן שלב הסינתזה, ברור שמערכת יחידה לא קיימת ללא קשר לערבי M, R, L_f, L_h.

התנאי ליחידות של מסנן הסינתזה האופטימלי הוא שלכל $0 \le m \le M-1$, סדרת הוקטורים $\hat{x}_{sR}(m)$ שעליה מבוצעת פעולת ה-WOLA תהיה תהליך אקראי לא-מנוון (כמתואר בנספח ד', סעיף III). תנאי זה מוכח בנספח ד', סעיף VI, והוא מתקיים בדרך כלל עבור ערך לא יותר מדי קטן של L ($L \ge 10$).

המקרה היחיד בו קיים פתרון אנליטי עבור מסנן הסינתזה האופטימלי הוא המקרה של $M = R \triangleq L_f$ ו-$G(f)$, שעבורו התוצאה המפורשת ניתנת בנספח ד', סעיף III. באותם תנאים כש-$M < L_f$ מפותחות בנספח ד', סעיף VI, משוואות מפושטות עבור מסנן הסינתזה האופטימלי וסיבוכיות פתרונן מתוארת שם.

מתקבל שמסנן הסינתזה האופטימלי מושפע מ-$\hat{L}(L_f+B)$ תוחלות מותנות של דגמי הכניסה ומ-$\hat{L}^2(L_f/MB)$ ערכי שכיחויות של זוגות סמלים, כך שעבור ערכים קטנים מספיק של $\hat{L}$ גישת תכנון זו ניתנת לביצוע. סיבוכיות החישוב של מסנן הסינתזה האופטימלי היא בקירוב $O(\hat{L}^2 BL_f)$, כפי שמתואר שם.

5.3 __מערכות אנליזה-סינתזה אופטימליות עבור כימות עדין__

במקרה של FQ, מדד השגיאה $\hat{U}$ ניתן כפונקציה מפורשת של מקדמי מסנן (חלון) האנליזה $h(\cdot)$. לפיכך במקרה זה (ובמקרה זה בלבד!) ניתן לתכנן את מסנן האנליזה האופטימלי בהנתן מסנן סינתזה.

רישום מפורש של $\hat{U}$ מעלה שזו תבנית ריבועית מוגדרת אי-שלילית במקדמי מסנן האנליזה $h(\cdot)$, כפי שמתואר בפרוט בנספח ד', סעיפים IV, VII. לכן מסנן האנליזה האופטימלי עבור מסנן סינתזה נתון מתקבל על ידי פתרון מערכת משוואות לינאריות (כמובן אחרת מזו שתוארה בסעיף 5.2).

יתר על כן, יחידות של מסנן האנליזה האופטימלי מובטחת ובלבד שהפונקציה $w(f)$ שהוזכרה בסעיף 5.1 שונה מאפס על קבוצה שמידתה אינה אפס (כמוכח בנספח ד' סעיף VII). כמוכן, מסנן האנליזה האופטימלי כלל אינו מושפע על ידי סטטיסטיקת רעש הכימות (או עצם קיומו של רעש כזה), ובדרך כלל הוא לא יתאים למערכת יחידה, כי אם לפשרה בין דרישת מערכת היחידה שמתבטאת על ידי המזעור של U, ולבין דרישת הפרדת התדר שהוספנו לה.

בהתאם לאמור לעיל ניתן לפרק את מטריצת המקדמים של מערכת משוואות זו לסכום $\tilde{Q}_D + \tilde{Q}_f$, כש-$\tilde{Q}_D$ משקפת את דרישת הפרדת התדר (והיא בלתי-תלויה במסנן הסינתזה) ואילו $\tilde{Q}_f$ משקפת את דרישת מערכת היחידה.

מאחר ותיארנו לעיל ובסעיף 5.2 שיטות פשוטות לתכנון אופטימלי של אחד ממסנני מערכת ה-A/S בהנתן המסנן השני, מתבקש מאליו האלגוריתם האיטרטיבי הבא לתכנון מערכת A/S אופטימלית:

(א) יהא $r = 0$ (מספר האיטרציה), ונניח שנתון צמד מסננים התחלתי $(\underline{h}^{(o)}, \underline{f}^{(o)})$.

(ב) יהא $\underline{h}^{(r+1)}$ מסנן אנליזה אופטימלי כלשהו עבור מסנן הסינתזה הנתון $\underline{f}^{(r)}$, כאשר אם $\underline{h}^{(r)}$ הוא בעצמו מסנן אנליזה אופטימלי, נבחר אותו כמסנן $\underline{h}^{(r+1)}$.

(ג) יהא $\underline{f}^{(r+1)}$ מסנן סינתזה אופטימלי כלשהו עבור מסנן האנליזה הנתון $\underline{h}^{(r+1)}$, כאשר אם $\underline{f}^{(r)}$ בעצמו הוא מסנן סינתזה אופטימלי, נבחר אותו כמסנן $\underline{f}^{(r+1)}$.

(ד) אם צמד המסננים $(\underline{h}^{(r)}, \underline{f}^{(r)})$ זהה לצמד המסננים $(\underline{h}^{(r+1)}, \underline{f}^{(r+1)})$. אזי סיימנו.
אחרת, הגדל את r באחד וחזור לשלב (ב).

עבור המקרה הפרטי בו: $R = M$, $G(f) = 1$, תהליך הכניסה חסר-קורלציה, ואין רעש כימות, הרי מדד השגיאה $\hat{U}$, מזדהה[3] עם מדד השגיאה הדטרמיניסטי ששימש ב-[54] וכן האלגוריתם האיטרטיבי שנוסח לעיל.

בהנחה הלא-מגבילה שפונקצית המשקל $W(f)$ היא חיובית על קבוצה בעלת מידה שאינה אפס (ולפיכך קיים מסנן אנליזה אופטימלי יחיד עבור כל מסנן סינתזה נתון), ניסחנו בנספח ד', סעיף VI (והוכחנו בנספח ד', סעיף VII) משפט המסכם את תכונות ההתכנסות של האלגוריתם האיטרטיבי שתואר לעיל. להלן תמציתו:

(א) מדד השגיאה $\hat{U}$ יורד מונוטונית מאיטרציה לאיטרציה, אלא אם כן האלגוריתם עוצר בנקודת שבת שלו.

(ב) תהא Γ קבוצת נקודות השבת של האלגוריתם, אזי זו בדיוק קבוצת הפתרונות של המשוואה $\nabla\hat{U} = \underline{0}$, וכל הנקודות היציבות בתוך Γ הן מינימומים לוקליים של $\hat{U}$.

---

3. ההבדל שנותר בין גישתנו ובין [54], הוא ששם היה אילוץ לינארי על הגבר מערכת ה-A/S שחייב שימוש בכופל לגרנז' וסיבך את האיטרציות לעומת אלו שמתוארות כאן. אצלנו אילוץ זה משולב באופן אוטומטי בתוך מדד השגיאה. גם האלגוריתם האיטרטיבי שנוסח שם שונה מעט משלנו, מסיבה זו.

(ג) כשהמטריצה Q_v (המשקפת את רעש-הכימות) היא לא סינגולרית, אזי ל-$\hat{U}$ יש מינימום גלובלי שהוא נקודת שבת של האלגוריתם האיטרטיבי, לכל סדרה המתקבלת על ידי האלגוריתם יש לפחות נקודת גבול אחת, וכל נקודות הגבול הללו הן נקודות שבת של האלגוריתם (קרי שייכות ל-Γ).

(ד) גם כאשר המטריצה Q_v היא סינגולרית (למשל, כשכלל אין רעש כימות) הרי כל נקודות הגבול של סדרות המתקבלות על ידי האלגוריתם הן נקודות שבת שלו (קרי, ב-Γ). אולם, במקרה זה לא מובטח של-$\hat{U}$ יהא מינימום גלובלי, ולא לכל סדרה יש נקודת גבול.

בנוסף למשפט דלעיל מחושב שם קצב ההתכנסות של האלגוריתם. מוכח שם שהאלגוריתם הנ"ל מתכנס בקצב לינארי לכל נקודה ב-Γ, שבה לשתי מערכות המשוואות הלינאריות שהוזכרו לעיל יש פתרון יחיד. כמובן מחושב שם במפורש קצב ההתכנסות שקשור לע"ע האקסטרמליים של המטריצות:

$$\left(\frac{\partial^2\hat{U}}{\partial h_i \partial h_j}\right) \quad , \quad \left(\frac{\partial^2\hat{U}}{\partial f_i \partial f_j}\right) \quad \text{-ו} \quad \left(\frac{\partial^2\hat{U}}{\partial f_i \partial h_j}\right)$$

פרק 6 : מערכי מסננים לסינתזה אופטימלית של אותות לאחר אנליזה ומודיפיקציה

6.1 תנאים לקיום מערכות אנליזה-סינתזה שהן מערכות יחידה ללא מודיפיקציה

התנאים לקיום מערכות אנליזה-סינתזה שהן מערכות יחידה ללא מודיפיקציה, נחקרו כבר באופן חלקי (למשל ב-[12,15]). כאן נציג הכללה של תוצאות אלה. התוצאות של [15] הן רק לגבי סינתזה על ידי מערך מסננים אחיד (שיטת WOLA לסינתזה), ולא ברור מהן ההנחות המדויקות לגבי תנאי ההתחלה. התוצאות של [12] הן עבור סינתזה לינארית כלשהי, אך בהנחה שמסנן האנליזה הוא מסנן FIR, וכן כשתנאי ההתחלה אינם ידועים. בהמשך ובנספח ה', סעיף I אנו מגדירים שלושה סוגי מערכות יחידה. אנו מניחים שהאנליזה נעשית על ידי מסנן סיבתי (לאורך כל הדיון לא נבחין בין מסנן FIR למסנן IIR), שתגובת דגם יחידה שלו היא $\{h(n)\}_{n=0}^{\infty}$, כאשר $h(0) = 0$ (נוח למטרות "סינכרון" של אינדקסים שונים ולפישוט התוצאות, להזיז את המסנן בדגם אחד). סידרת הכניסה $\{x(n)\}_{n=0}^{\infty}$ היא סדרה חצי-אינסופית וכתוצאה מתהליך האנליזה נוצרת סדרה חצי-אינסופית של התמרות לזמן-קצר $\{Y_{sR}\}_{s=1}^{\infty}$.

כל מערכות היחידה חייבות לקיים שכל דגם בפלט שלהן מופק תוך זמן סופי, ואין שום הנחות א-פריורי על סדרות הקלט (קרי, אלו יכולות למשל להיות סדרות לא-חסומות, וכד'), או על צורת הסינתזה (קרי, הדיון תקף גם לסינתזה לא-לינארית). להלן הגדרת שלושת סוגי מערכות היחידה:

(א) US - מערכת המשחזרת את סדרת הדגימות המקורית ללא שגיאה, מתוך סדרת ההתמרה לזמן-קצר שלה.

(ב) f.US - מערכת כנ"ל, שיש לה את התכונה הנוספת, שניתן לשחזר כל רישא סופית של ILR דגימות מסדרת הדגימות המקורית, מתוך IL הוקטורים הראשונים בסדרת ההתמרה לזמן-קצר (DSTT), וזאת לכל $L \geq 1$. כאשר R הוא קצב הדצימציה ואילו $I \triangleq M/gcd(M,R)$, כש-M הוא מימד הטרנספורם ו-$gcd(\cdot,\cdot)$ זהו המחלק המשותף הגדול ביותר. כמו כן נסמן על ידי $J \triangleq R/gcd(M,R)$

(ג) DUS - מערכת המתאימה לכל סדרת התמרה לזמן-קצר (גם כזו שעברו מודיפיקציה כלשהי לאחר האנליזה), סדרת דגימות בעלת אותה התמרה לזמן-קצר

נציג להלן את התוצאות העיקריות שפיתוחן מצוי בנספח ה', סעיף II. לפני כן, נעיר רק שאנו מניחים ש- $R \leq M$, שכן זהו המקרה המעניין מבחינה מעשית

(1) כל התוצאות שיוצגו להלן לגבי קיום מערכות יחידה, הן בלתי תלויות בטרנספורם הספציפי שבמערכת, כי אם רק בפרמטרים R, M ובמסנן האנליזה

(2) קיים DUS אם ורק אם $M = R$ ו-$h(n) \neq 0$ עבור $1 \leq n \leq M$ (משפט 1, שם).

(3) כל מערכת US היא מערכת לינארית. בנוסף, אם לא קיימת מערכת כזו, אזי גם הוספת מספר סופי כלשהו של דגימות מהסדרה המקורית כקלט לא תספיק לשחזור ללא שגיאה. (משפטים 2, 4-1, שם).

(4) קיימת fUS (והיא בודאי גם US) אם ורק אם $h(n) \neq 0$ עבור $1 \leq n \leq R$ (משפט 3, שם).

(5) קיימת US רק אם $\{h(p+Rk)\}_{k=0}^{\infty}$ אינה זהותית אפס, לכל $1 \leq p \leq R$. זהו גם תנאי מספיק לקיום US, כאשר $gcd(R,M) = R$ (משפט 5, שם).

(6) תנאי מספיק והכרחי, לקיום US הוא שלכל $1 \leq p \leq R$ ישנה רישא סופית מסוימת של ה-DSTT שמתוכה ניתן לשחזור את $\{x(Mr-p)\}_{r=1}^{J}$ (טענה 7, שם).

במקצת מתוצאות אלו נשתמש בסעיפים הבאים.

6.2 סינתזה אופטימלית בקריטריון WMMSE עבור אות סופי

מאחר ולא תמיד קיימת DUS, הרי כשהאנליזה מלווה במודיפיקציה יתכנו סדרות של MDSTT שאינן DSTT של אף סדרת דגימות. במקרה זה מתעוררת בעית הסינתזה האופטימלית של סידרת דגימות מתוך MDSTT. גם כש- $M = R$ וקיימת DUS, הרי קל לודא שלמשל עבור מסנן אנליזה שהוא FIR בעל פזה לינארית, מערכת ה-DUS מכילה תמיד מסנני IIR לא יציבים, ולכן מבחינה מעשית אינה מועילה.

נושא הסינתזה האופטימלית כבר נדון ב-[17, 16, 12] בהקשר של שינוי ציר הזמן, השמטת הפזה ו/או ערבול האות במישור הטרנספורם. בניגוד למודיפיקצית הכימות שנדונה בפרק הקודם, ההנחה באפליקציות הנ"ל היא שהמודיפיקציה רצויה ולכן על הסינתזה ליצור סדרת דגימות שאינה בהכרח קרובה לקלט המקורי של שלב האנליזה, כי אם בעלת DSTT שקרוב לזה הנתון.

קריטריון האופטימיזציה שהוצג ב-[12,16,17] הוא המזעור של ה-WMMSE בין ה-MDSTT הנתון לבין ה-DSTT של מוצא הסינתזה. אנו אימצנו קריטריון זה, כשביצענו את ההכללות הבאות:

(א) הטרנספורם T הוא טרנספורם לינארי, רגולרי כלשהו (לאו דוקא ה-DFT).

(ב) על מנת שבעית האופטימיזציה תהא מוגדרת היטב, הגבלנו עצמנו לסינתזה מתוך MDSTT סופי (לעת עתה), והנחנו שמסנן האנליזה מקיים את התנאי ההכרחי והמספיק לקיום ftUS, כך שאם לא נעשית מודיפיקציה אין שגיאה בשחזור.

(ג) בנספח ה', בסעיף III אנו מראים שבעית הסינתזה (הכללית) שהוצגה כאן מזדהה (מבחינה מתמטית) עם בעית הסינתזה של סדרה זמנית אופטימלית עבור מודיפיקציה לינארית ידועה שבה קריטריון האופטימיזציה מחושב ביחס לאות לאחר אנליזה שניה ומודיפיקציה הפוכה.

בסעיף III של נספח ה', אנו מוכיחים שבעית הסינתזה האופטימלית היא בעלת פתרון יחיד לכל MDSTT ושהמערכת המממשת פתרון זה היא מערכת ftUS. יתר על כן הסינתזה האופטימלית היא פתרון מערכת משוואות לינאריות, כאשר וקטור האברים החופשיים נוצר על ידי סינתזה בשיטת WOLA (עם מסנן סינתזה לא-סיבתי, שהוא שיקוף בזמן של מסנן האנליזה שמגדיר את ה-DSTT), ואילו מטריצת המערכת היא ב"ת ב-MDSTT הנתון.

הבעיתיות המרכזית של הסינתזה האופטימלית, היא הצורך בהיפוך מטריצת המקדמים הנ"ל שממדיה האופיניים הם 10,000 × 10,000 וברור שדרוש פתרון אנליטי של בעית ההיפוך (גם אם כאמור לעיל, ניתן לממש אותו מראש כי הוא ב"ת ב-MDSTT הנתון). תופעה זו אובחנה גם ב-[12].

הדיון עד כאן נעשה ללא הגבלת מסנן האנליזה להיות מסנן FIR, אולם לצורך פיתוח שיטות להיפוך יעיל של מטריצת המקדמים (שתסומן ב-S), נניח שזהו מסנן FIR באורך של $L_h = RIK_h$ מקדמים.

בסעיפים IV, V של נספח ה', נחקרות התכונות של המטריצה S, במקרה זה. התוצאה הראשונה המעניינת היא שעבור טרנספורם יוניטרי (שהוא המקרה הנפוץ) ומשקול אחיד של השגיאות באברי הטרנספורם השונים, הרי המטריצה S פריקה ל-M תת-מטריצות זרות זו-לזו, כשכל תת-מטריצה אחראית לשחזור תת-סדרה זרה של דגימות מתוך המוצא של polyphase אחר של הסינתזה בשיטת WOLA. אם בנוסף $K_h = 1$, הרי כל תת-מטריצה כזו היא אלכסונית בבלוקים ממימד J כל אחד, ובעית ההיפוך שלה הופכת לטריויאלית.

התוצאה הסופית המתקבלת היא סינתזת WOLA עם תיקון על ידי קבועי משקל מתאימים שהם (פרט לקצוות) בלתי-תלויים באורך ה-MDSTT. עבור $J = 1$ זו בדיוק התוצאה שפותחה כבר ב-[16,17], אולם מאחר ו-$IR = MJ-1$ הרי $J = 1$ מחייב $M \leq L_h$ וביצועי מסנן האנליזה במקרה זה אינם טובים (ראה למשל ב-[6]). ההכללה שלנו ל-$J > 1$ היא בעלת משמעות מעשית מעניינת, כדלקמן: כאשר R אינו מוכתב על ידי האפליקציה הנדונה, כדאי לבחור את R כמספר ראשוני, ואזי $R = J$ כך שניתן להשתמש בסינתזה WOLA המוכללת עבור מסנני אנליזה באורך של עד MR דגימות!

במקרה הכללי (של טרנספורם לא-יוניטרי), או כש-1 < K_h, פתרון בעית ההיפוך של S נעשה מורכב יותר. יהא ILR אורך סדרת הדגימות שיש לחשב (קרי, המימד של S), אזי סיבוכיות ההיפוך של S באופן נומרי היא $O((ILR)^3)$ פעולות. בסעיף IV אנו מציגים פתרון המבוסס על פירוק S לבלוקים וניצול היותה Banded (בעלת מספר קטן של אלכסונים שונים מאפס), שמחייב רק $O((IK_h R)^2(ILR))$ פעולות. המשמעות של פתרון זה הוא שמספר הפעולות לדגם פלט קבוע, ומשתמע גם שניתן לבצעו בזכרון ביניים של (IK_hR) ערכים לכל דגם פלט, קרי זכרון וסיבוכיות לינאריים באורך הקלט.

לפתרון זה יש עדיין שני חסרונות מעשיים והם:

(א) מחייב להמתין עד לסוף ה-MDSTT לפני ביצוע הסינתזה, דהיינו אין אפשרות לפעולה On-line.

(ב) כאשר $L \to \infty$, עלולות להתעורר בעיות של יציבות נומרית של הפתרון (כפי שנראה בסעיף הבא).

ננסה להתגבר עליהן על ידי פיתוח סכמת סינתזה ל"מצב-מתמיד" עבור אות אינסופי כמתואר בסעיף הבא.

6.3 הסינתזה האופטימלית עבור אות אין-סופי (פתרון במצב מתמיד)

שוב, למטרות הפשטות, נגביל עצמנו למקרה של טרנספורם יוניטרי, אולם עם $K_h > 1$. כאמור בסעיף הקודם, המטריצה S פריקה ל-M תת-מטריצות $\{s^{(p)}\}_{p=1}^{M}$ שכל אחת מהן מתאימה למוצא של polyphase [6] אחר של סינתזת WOLA. בסעיף V אנו מראים שפרט לזנב ממימד קבוע (ושערכו קבוע) המטריצות $s^{(p)}$ הן מטריצות Block Toeplitz, עם בלוקים ממימד $J \times J$ והן Banded לרוחב של K_h בלוקים (טענה 8, שם).

ניתן לכן לפתח את הסינתזה האופטימלית בעקבות הגישה האלגברית של [12,16]. אנו העדפנו לנסח את תוצאותינו במושגים מתורת המערכות הלינאריות ולכן נקטנו בדרך שונה מעט. עם זאת, למטרות השלמות, נבהיר במקצת את הגישה האלגברית לבעיה זו, ונציין כבר עתה שהתוצאות בשתי הדרכים מתלכדות וההבדל הוא למעשה רק בצורת הצגתן, ובהנמקתן.

הגישה האלגברית שפותחה עבור המקרה של $J = R = 1$ ב-[12] מבוססת על מציאת מטריצה שהיא אסימפטוטית (כש-$L \to \infty$) אקוויולנטית ל- $S^{(p)}$ (במובן שהוגדר ב-[73]), כך שחישוב ההפיך שלה, פשוט יחסית. במקרה הפרטי של $J = R = 1$, $S^{(p)}$ הן מטריצות טואפליץ והמטריצות האקוויולנטיות להן הן מטריצות Circulant [73]. ההפכי של הללו ניתן לחישוב ביעילות על ידי שימוש ב-DFT [67] ומוביל שוב למטריצות Circulant. ב-[12] מוראה שמשיקולי אנרגיה ניתן לקטום את מטריצות ה-Circulant המתקבלות ולהפכן למסנני FIR סופיים שפועלים כמסנני תיקון אחרי הסינתזה בשיטת WOLA. ההכללה למקרה הכללי אינה מסובכת ונציג כאן את עיקריה:

(א) ניתן להראות שהמטריצות $S^{(p)}$ הן אסימפטוטית אקויולנטיות למטריצות Block-Circulant עם בלוקים ממימד $J \times J$.

(ב) בתנאי רגולריות מסוימים (לא קשים במיוחד) המטריצות הללו הפיכות. במקרה זה ניתן להפכן ביעילות על ידי DFT שמופעל על הבלוקים + היפוכי מטריצות ממימד $J \times J$ במישור התדר + DFT הפוך על הבלוקים שנוצרו בתדר (זו הרחבה מיידית של השיטה המתוארת ב-[67]).

(ג) לאחר קטימה משיקולים דומים לאלו שהופעלו ב-[12], מתקבלת סינתזה על ידי כפל במטריצה שהיא Banded-Block-Toeplitz שמהווה את ההרחבה של מסנני התיקון של [12], למקרה של $J > 1$. הסיבוכיות היא צנועה, שכן ממדי הבלוקים הם $J \times J$ והמימוש הוא כמובן On-line (בדומה לזה של [12]).

נציג כעת את עיקרי הפיתוח שלנו (שמוצג במלואו בסעיף V של נספח ה'), עבור $L \to \infty$, ה"זנב" של המטריצה $S^{(p)}$ נעלם. נסמן ב- $\{B_k\}_{k=-(K_h-1)}^{(K_h-1)}$ את $(2K_h - 1)$ הבלוקים השונים מאפס במטריצה זו (כאשר: $B_{-k} = B_k^*$) ונארגן את מוצא ה-polyphase המתאים של סינתזת WOLA, ותת-סידרת הסינתזה האופטימלית הדרושה, כשתי סדרות חצי אינסופיות של וקטורים באורך J, $\{\underline{v}_n\}_{n=0}^{\infty}$ ו- $\{\underline{u}_n\}_{n=0}^{\infty}$ בהתאמה. בעית הסינתזה האופטימלית הופכת לבעית דה-קונוולוציה מהצורה הבאה -> מצא את הסדרה $\{\underline{u}_n\}_{n=0}^{\infty}$ כך שהוא תקיים $\sum_{k=-(K_h-1)}^{(K_h-1)} B_k \underline{u}_{n-k} = \underline{v}_n$ $n \geq 0$. קל לראות שללא ידיעת $\underline{v}_n$ עבור $n < 0$, זו בעיה מרובת פתרונות

פתרון קרוי יציב כאשר הוא אינו רגיש לתנאי ההתחלה של $\underline{v}_n$ (במובן שעבור כל סידרה $\underline{v}_n$ שמתאפסת לכל $n \geq n_o$, הרי $\lim_{n\to\infty} \|\underline{u}_n\|_2 = 0$). אנו מראים בסעיף V של נספח ה' (טענה 9, שם) שקיים פתרון יציב יחיד שניתן על ידי הנוסחא:

$$\tilde{\underline{u}}_n = \sum_{k=-\infty}^{\infty} Y_k \underline{v}_{n-k} \qquad (6.1)$$

כאשר $\{Y_k\}$ זו סידרת מטריצות המוגדרת מתוך הסידרה $\{B_k\}$ על ידי התמרת Z והיפוכה (תאור מדויק ראה במשפט 7, שם). יתר על כן, מוראה שם (במשפט 7), שלכל ε קטן כרצוננו, קיים T_ε כך שקטימת הקונוולוציה ב-(6.2) לתחום $|k| < T_\varepsilon$, אינה משנה את ערכי $\underline{u}_n$ ביותר מ-$\pm\varepsilon$ (וזאת כש-T_ε ב"ת ב-n, או בסדרה $\underline{v}_n$ הנתונה).

תנאי הרגולריות הדרוש לשם קיום הפתרון הנ"ל (וזהו גם תנאי הרגולריות הדרוש לצורך פיתוח הגישה האלגברית) הוא שכל פתרונות המשוואה:

$$\det \left\{ \sum_{k=-(K_h-1)}^{(K_h-1)} B_k Z^{-k} \right\} = 0 \qquad (6.3)$$

(ויש לכל היותר $J(2K_h - 1)$ כאלה, כי אגף שמאל הוא פולינום ממעלה סופית ב-Z), אינם על מעגל היחידה. לא קשה לבדוק תנאי זה נומרית כי $J(2K_h - 1)$ הוא בדרך כלל קטן מ-10.

בסעיף V של נספח ה', מוראה גם שניתן לחשב על ידי DFT מספיק גדול את ערכי Y_k בדיוק שיספיק על מנת להבטיח טעות ב-$\tilde{\underline{u}}_n$ החסומה על ידי $\pm\varepsilon$ (טענה 10, שם). שיטת החישוב של Y_k היא לפיכך:

(א) בחר Q_ε מספיק גדול (לכל הפחות: $Q_\varepsilon \geq (2T_\varepsilon - 1)$ ו-$Q_\varepsilon \geq (2K_h - 1)$), ובצע לסדרת המטריצות $\{B_k\}_{k=0}^{Q_\varepsilon - 1}$ DFT ממימד Q_ε (שמשמעותו J^2 DFT-ים).

(ב) את סדרת Q_ε המטריצות המתקבלות, הפוך אחת אחת (Matrix-Inversion), פעולה שניתנת לביצוע, לאור משוואה (6.3) ותנאי הרגולריות שציטטנו.

(ג) לסדרה המתקבלת בצע IDFT ממימד Q_ε (שוב זהו למעשה בצוע J^2 DFT-ים) ו-$(2T_\varepsilon - 1)$ האברים המקיימים $|k \bmod Q_\varepsilon| \geq (T_\varepsilon - 1)$ הם בדיוק סדרת המטריצות $\{Y_k\}$ הדרושה.

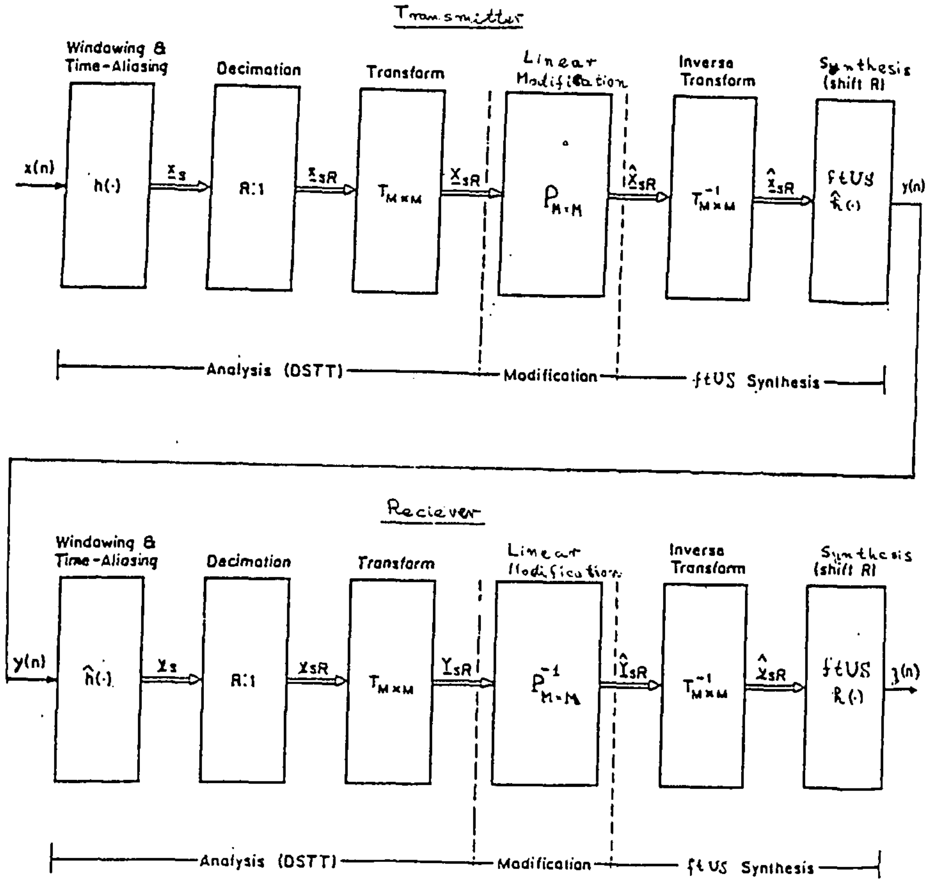
6.4 תנאים לקיום מערכות יחידה המכילות מודיפיקציה לינארית

בציור 1.8 מתוארת מערכת המבוססת על שתי מערכות A/S, הראשונה מכילה מודיפיקציה ( )m והשניה מכילה את המודיפיקציה ההפוכה.

סכמה כזו מתאימה במיוחד לשינוי ציר הזמן של אות-הדיבור ולעירבול שלו (ראה למשל [12] ליתר פירוט).

ננתח כעת את המקרה של מודיפיקציה לינארית (וזהו סוג המודיפיקציה המקובל למשל למטרות ערבול), המאופיינת על ידי מטריצה רגולרית M×M ממדית P.

סכמת הצפנה / פענוח אופיינית למקרה זה מתוארת בציור 6.1.



ציור 6.1: סכמת הצפנה / פענוח על ידי מערכת A/S ומודיפיקציה לינארית

Fig. 6.1: Scrambling /De-Scrambling System using A/S systems with Linear Modification.

כאשר לסדרה חצי-אינסופית $x(\cdot)$ ישנה סדרה $y(\cdot)$ המקיימת שה-DSTT שלה (במסנן $h(\cdot)$) מזדהה עם ה-MDSTT הנתון על ידי $\{\hat{x}_{sR}\}_{s=1}^{\infty}$, הרי מובטח (כפי שקל לראות מציור 6.1) שמוצא המפענח $z(\cdot)$ מזדהה עם כניסת המצפין $x(\cdot)$.

מודיפיקציה P שהיא בעלת תכונה זו, עבור <u>כל סדרה</u> x(•) חצי-אינסופית, קרויה מודיפיקציה "חוקית" (LM) ביחס למסנן האנליזה h(•) (כשהמסנן $\hat{h}(\cdot)$ נבחר כך שיהא אבן שחזור ללא-טעות, בהגבלות שזהו מסנן סיבתי והוא ב"ת בסדרה הספציפית x(•)).

עבור מודיפיקציה "חוקית" המערכת המתוארת בציור 6.1 היא מערכת יחידה. במהלך חקירת התנאים שעל P לקיים על מנת שתייצג LM, התברר שמקצת התנאים תלויים במסנן האנליזה h(•) הספציפי. לפיכך נגדיר מודיפיקציה "חוקית" אוניברסלית (ULM) כמודיפיקציה P שהיא LM ביחס לכל מסנן אנליזה h(•) שעבורו קיימת tfUS.

ברור לכן שכשמשתמשים ב-ULM מובטחת מערכת יחידה (כולל המודיפיקציה) לכל מסנן אנליזה h(•).

פיתוח התנאים שעל P לקיים על מנת להיות LM (ULM) מובא בנספח ה', סעיף VI. נתאר כאן את התוצאות המרכזיות בנושא זה.

(א) תכונת ה-LM נקבעת רק על פי מיקום האברים השונים מאפס במטריצה $\hat{P} = (T^{-1}PT)$, ולא על פי ערכם. יתר על כן, בהנתן $\hat{P}$ תכונת ה-LM תלויה רק במסנן האנליזה h(•) וב-R, M ולא בטרנספורם T שבשימוש.

(ב) מטריצה P היא LM ביחס למסנן אנליזה h(•) איזשהו, רק אם $\hat{P}$ היא Block-Diagonal עם בלוקים ממימד $g \times g$ כל אחד (כש-$g = gcd(R,M)$, זו טענה 13, בנספח).

(ג) כש-$\hat{P}$ היא LM ביחס ל-h(•) שיש לו ftUS, הרי גם למסנן $\hat{h}(\cdot)$ תהא ftUS (טענה 12, שם).

(ד) עבור $R = M$, כל מטריצה רגולרית P היא ULM, כשמשתמשים ב-$h = \hat{h}$ (משפט 8, שם).

(ה) עבור $R < M$, קיים מסנן h(•) שביחס אליו כל מטריצה רגולרית $\hat{P}$ שהיא Block-Diagonal עם בלוקים ממימד $g \times g$ כל אחד, היא LM, כש-$h = \hat{h}$ (משפט 9, שם).

(ו) עבור $R < M$, המטריצות $\hat{P}$ שהן ULM מאופיינות ככפל מטריצה אלכסונית לא-סינגולרית כלשהי, במטריצה Block-Diagonal עם בלוקים זהים שהם מטריצת פרמוטציה כלשהי ממימד $g \times g$ (משפט 10, שם). יתר על כן, ניתן לבנות מסנן אנליזה h(•) שהוא FIR מאורך 2R, שביחס אליו אלו כל ה-LM.

<u>הערה:</u>

למטרות ערבול מקובל להשתמש בטרנספורם ה-DFT, ובמודיפיקציה המיוצגת על ידי מטריצת פרמוטציה P. באופן כללי לפרמוטציה P לא תהא תכונת ה-LM ביחס למסנן האנליזה $h(\cdot)$ (כפי שדווח נסיונית ב-[12]), אולם קל לראות שלפרמוטציות ציקליות (מהצורה $\pi(i) = (i+s)\mathrm{mod}M$, $0 \le s \le M-1$) יש את תכונת ה-ULM, ולכן מבטיחות הצפנה / פענוח ללא טעות. אם זאת, הניתוח שמתואר לעיל מאכזב מעט מבחינה מעשית כי נאלצנו לצמצם את מרחב הערבולים האפשריים מ-$M!$ פרמוטציות ל-M פרמוטציות ציקליות בלבד.

פרק 7 : סיכום ותאור בעיות פתוחות

7.1 סיכום

בעבודה זו התמקדנו בתכנון מערכי מסננים ספרתיים המשולבים במערכות אנליזה וסינתזה. טיפלנו בשני סוגים שונים של מערכות, כדלקמן - מערכות שבהן נעשית אנליזה בלבד ומערכות שבהן האנליזה מלווה בסינתזה (כשבדרך כלל ביניהן נעשית מודיפיקציה של האות). בעוד שבפרקים 3 ו-4 עסקנו במערכות מהסוג הראשון, הרי שפרקים 5 ו-6 יוחדו לסוג השני של מערכות. באופן כללי הוצגו בעבודה זו ארבע שיטות שונות לתכנון מערכי מסננים, כשבכל אחד מהפרקים 3-6 הוצגה שיטת תכנון עקרונית אחת, יחד עם נושאים תאורטיים הקשורים אליה.

(א) מערכות לאנליזה בלבד (מערכי מסננים)

בפרק 3, הצגנו את האלגוריתם (האנליטי) לתכנון המערך האופטימלי בקריטריון שגיאה-ריבועית משוקללת, תחת אילוץ תגובה כוללת מוכתבת (סעיף 3.1). בפרק 4, הצגנו את האלגוריתם המפושט לתכנון מערך אחיד אופטימלי בקריטריון L_{∞} כשהתגובה הכוללת מוכתבת (סעיף 4.3), וקירוב שלו על ידי הסבת שיטת ה"חלון" (סעיף 4.4).

בנוסף לפיתוח אלגוריתמים אלו ישנן בעבודה גם תרומות תאורטיות, המתייחסות למכלול רחב יותר של בעיות תכנון מסננים (ולבעיות קירוב בכלל). תרומות אלו הוצגו לחוד בצורה הכללית ביותר (בסעיף 3.2) ובמקרה הפרטי של מערכי מסננים אחידים (בסעיף 4.1). הן כוללות פיתוח תנאים לקיום ויחידות של מערך מסננים אופטימלי תחת אילוצי תגובה כוללת, ותנאים לממשיות מקדמי המערך, לפזה לינארית של המסננים שבו, ולסימטריה המאפשרת גזירת כל המסננים במערך מתוך מסנן אב-טיפוס. כמובן, הוצג הקשר שבין "טיב" המסננים המרכיבים את המערך לבין הדרישות על התגובה הכוללת שלו.

(ב) מערכות אנליזה-סינתזה

בפרק 5 ניתחנו מערכות אנליזה-סינתזה המשמשות לקידוד ומכילות בתוכן כימות (קוונטיזציה). מאחר ותהליך הכימות הוא בלתי-הפיך, הרי המערכות האופטימליות אינן בהכרח מערכות יחידה. ככל שהכימות גס-יותר (קרי, המערכת מיועדת לקידוד בקצב נמוך של סיביות לשניה), הרי על ידי תכנון מערכת אנליזה-סינתזה תוך התחשבות בצורת הכימות הספציפית שבה, ניתן להשיג שיפור משמעותי יותר ביחס לתכנון אוניברסלי (בהזנחת אפקט הכימות).

הצגנו אלגוריתם לתכנון מערך אחיד אופטימלי של מסנני-סינתזה עבור שני סוגים מקובלים של כימות (קרי, כימות עדין, וכימות על ידי ספר-קוד), תוך התחשבות במאפיינים הסטטיסטיים של הכימות (סעיף 5.2).

במקרה של כימות עדין (ש"ממודל" על ידי מקור רעש אדיטיבי), הוצגה התלות המפורשת של קריטריון הטיב במסנני האנליזה, וזו אפשרה פיתוח אלגוריתם איטרטיבי לתכנון מערכות אנליזה-סינתזה אופטימליות (סעיף 5.3). אלגוריתם זה הוא הרחבה של האלגוריתם הדטרמיניסטי שאינו מתחשב בכימות שתואר ב-[54], וכאן הוכחו לראשונה גם תכונות ההתכנסות שלו (שמרביתן תופסות גם לגבי האלגוריתם שהוצג ב-[54]).

בתהליך הניתוח של מערכות אלו הושגו גם מספר תוצאות תאורטיות שיכולות להיות מיושמות לבעיות אחרות. למשל: פיתחנו הרחבה של קריטריון השגיאה הספקטרלית המשוקללת. עבור תהליכים המשתנים בזמן באופן מחזורי (סעיף 5.1), וכן פיתחנו שיטה לבחירת פתרון מסוים עבור מערכות משוואות מרובות פתרונות המאופיינות על ידי מטריצת מקדמים סימטרית, על ידי פרטורבציה של מטריצת המקדמים (סעיף 5.2).

בפרק 6 התייחסנו לבעית הסינתזה האופטימלית של סדרה זמנית מתוך האות המתקבל לאחר אנליזה ומודיפיקציה כלשהי (לא ידועה).

בעיה זו טופלה ב-[16,17] (ראה בפרק 2), ונפתרה שם באופן חלקי עבור מערכי מסנני-אנליזה אחידים שאורך תגובת דגם היחידה של כל אחד מהם קטן מ-M (מספר המסננים שבהם). הפתרון הכללי מוצג כאן לראשונה הן עבור אות סופי (סעיף 6.2) והן עבור אות אינסופי (סעיף 6.3). המקרה האחרון מתייחס לפתרון במצב-מתמיד של מערכת משוואות אינסופית. זו הכללה של הפתרונות החלקיים שתוארו שם ([13,16,17]), והיא מתייחסת בדרך כלל למערכת סינתזה אופטימלית שאינה מערכת OLA המקובלת.

פתרון בעיה זו מוצג במסגרת הכללית יותר של מערכות עם טרנספורם לינארי רגולרי כללי ואינו מוגבל למערכות המבוססות על STFT בלבד. במסגרת הכללה זו הצגנו גם מספר הרחבות של התנאים לקיום מערכות יחידה (שהוצגו לראשונה ב-[15]), ובעיקר הצגנו בצורה מפורשת תנאים מספיקים והכרחיים.

מערכות אנליזה-סינתזה הוצעו לאחרונה ככלי לערבול של אותות דיבור. במסגרת זו נעשות מודיפיקציות לינאריות על האות לאחר האנליזה, ובדרך כלל מודיפיקציות אלו גורמות לעוות רציני באות המשוחזר לאחר הפענוח (ראה [12]). בסעיף 6.4 הצגנו אפיון מלא ומפורש של מחלקת המודיפיקציות שעבורן ניתן להשיג שחזור ללא עוות של האות על ידי שתי מערכות אנליזה-סינתזה המחוברות בטור (כשהראשונה משמשת לערבול והשניה לפענוח). לאפיון זה חשיבות רבה בתכנון מערכות לערבול דיבור (פרוט ראה ב-[3,12]).

7.2 בעיות פתוחות

נותרו עדיין מספר בעיות פתוחות בנושאים שתוארו לעיל.

ראשית, על מנת לפתח אלגוריתם מהיר לתכנון מערך המסננים האופטימלי בקריטריון L_∞ כאשר התגובה הכוללת מוכתבת, נדרש אפיון שלו (בדומה לאפיון של Chebyshev עבור מסנן FIR אופטימלי בקריטריון L_∞). אפיון כזה אינו ידוע.

ישנם כיוונים רבים להרחבה אפשרית של העבודה התאורטית שתוארה בסעיף 3.2 ובעיות פתוחות רבות (אם כי מרביתן בעיות מתמטיות ולא בעיות בעלות אופי הנדסי). למשל: אחת הבעיות הפתוחות היא להרחיב את התוצאות למקרה של מספר אילוצים שונים על מערך המסננים. בעיה שניה היא למצוא פתרון לבעית הקרוב שמוצגת שם במרחב אינסוף-ממדי (קרי, כשמספר דרגות החופש הלא ידועות הוא אינסופי).

בנושא של מערכות אנליזה-סינתזה עם כימות המתואר בפרק 5, נותרה כבעיה פתוחה הרחבת שיטת התכנון למערכות עם כימות מסתגל (אדפטיבי) שבהן הסטטיסטיקה האפיינית של הכימות משתנה בזמן. בנוסף, קיים נושא למחקר נוסף והוא הרחבת כל התוצאות למערכות המבוססות על מערכי מסננים לא אחידים. זו בעיה לא פשוטה שכן במקרה זה גם קצבי הדצימציה שונים ממסנן למסנן.

בנושא הסינתזה-האופטימלית המתואר בפרק 6, נותרה כבעיה פתוחה הסינתזה האופטימלית כאשר האנליזה היא על ידי מערך מסננים לא-אחיד (בדומה לבעיה שתוארה לעיל), וכן הרחבת הייצוג של מחלקת המודיפיקציות הלינאריות מסעיף 6.4 למודיפיקציות הלא-לינאריות המאפשרות שחזור האות ללא עוות.

בנוסף, ניתן להרחיב את מרבית הבעיות והתוצאות למקרה הדו-ממדי, על מנת להשתמש בטכניקה של אנליזה-סינתזה למטרות עבוד תמונות. הרחבה כזו אינה קשה, אולם בנקודות מסוימות היא דורשת מחקר נוסף.

REFERENCES

[1] B. A. Dautrich, L. R. Rabiner and T. M. Martin, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-31, No. 4, August 1983, pp. 793-807.

[2] J. S. Lim and A. V. Oppenheim, "Enchancement and Bandwidth Compression of Noisy Speech", Proc. IEEE, Dec. 1979, pp. 1586-1605.

[3] L. S. Lee, G. C. Chon and C. S. Chang, "A New Frequency Domain Speech Scrambling System which Does not Require Frame Synchronization", IEEE Trans. Communications, April 1984, pp. 444-457.

[4] J. M. Tribolet and R. E. Crochiere, "Frequency Domain Coding of Speech", IEEE Trans. on ASSP - 27, No. 5, Oct. 1979, pp. 512-530.

[5] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice Hall, N. J., 1978 (chapter 6).

[6] R. E. Crochiere and L. R. Rabiner, Multirate Digital Signal Processing, Prentice Hall Signal Processing series, 1983.

[7] M. G. Bellanger and J. L. Daguet, "TDM-FDM Transmultiplexer - Digital Polyphase and FFT", IEEE Trans. Communications, vol. COM - 22, pp. 1199-1204, September 1974.

[8] M. G. Bellanger, G. Bonnerot, and M. Coudreuse, "Digital Filtering By Polyphase Network: Application to Sample Rate Alternation and Filter Banks", IEEE Trans. on ASSP, Vol. ASSP-24, No. 2, pp. 109-114, April, 1976.

[9] N. R. Dixon and H. F. Silverman, "A Description of a Parametrically Controlled Modular Structure for Speech Processing", IEEE Trans. on ASSP, Vol. ASSP-23, pp. 87-91, February 1975.

[10] D. Malah and J. L. Flanagan, "Frequency Scaling of Speech Signals by Transform Techniques", The BSTJ, vol. 60, No. 9, November 1981, pp. 2107-2156.

[11] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short Time Spectral Amplitude Estimator", IEEE Trans. on ASSP, Vol. ASSP-32, No. 6, pp. 1109-1122, December 1984.

[12] Z. Shpiro, Analog Speech Scrambling by Means of Discrete Short Time Fourier Transform, M.Sc. Thesis, Technion-Israel Institute of Technology, November 1983.

[13] R. E. Crochiere, "A Weighted Overlapp-Add Method of Short Time Fourier Analysis/Synthesis", IEEE Trans. on ASSP, vol. ASSP-28, No. 1, pp. 99-102, February 1980.

[14] R. V. Cox and R. E. Crochiere, "Real Time Simulation of Adaptive Transform Coding", IEEE Trans. on ASSP, pp. 147-154, April 1981.

[15] M. R. Portnoff, "Time Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis", IEEE Trans. on ASSP., vol. ASSP-28, No. 1, pp. 55-69, February 1980.

[16] Z. Shpiro and D. Malah, "An Algebraic Approach to Discrete Short Time Fourier Transform Analysis and Synthesis", Proc. ICASSP, pp. 2.3.1-2.3.4, 1984.

[17] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short Time Fourier Transform", IEEE Trans. on ASSP, vol. ASSP-32, No. 2, pp. 236-243, April 1984.

[18] L. R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1975 (Chapter 3).

[19] K. M. M Prabhu and V. U. Reddy, "Data Windows in Digital Signal Processing - A Review", J. Inst, Electron. and Telecommunications Eng., India, January 1980, pp. 69-76.

[20] J. F. Kaiser, "Non-Recursive Digital Filter Design Using the I_o-sinh Window Function", Proc. 1974 IEEE Int. Symp. Circuit Theory, pp. 20-23.

[21] A. Dembo and D. Malah, "Generalization of the Window Method for FIR Digital Filter Design", IEEE Trans on ASSP, October 1984, pp. 1081-1083.

[22] A. Dembo and D. Malah, "Generalization of the Window Method for FIR Digital Filter Design", EE. Pub. 456, Technion, I.I.T August, 1983.

[23] A. Papoulis and M. S. Bertram, "Digital Filtering and Prolate Functions", IEEE Trans. Circuit Theory, November 1972, pp. 674-681.

[24] A. H. Nuttall, "Some Windows with Very Good Sidelobe Behavior", IEEE Trnas. on Assp, Vol. ASSP-29, No. 1, pp. 84-91, February 1981.

[25] L. R. Rabiner, "The Design of Finite Impulse Response Digital Filters Using Linear Programming Techniques", BSTJ, July 1972, pp. 1177-1198.

[26] Y, C. Lim, "Efficient Special Purpose Linear Programming for FIR Filter Design", IEEE Trans. on ASSP, Vol. ASSP 31, No. 4, November 1983.

[27] J, H, McClellan, T. W. Parks and L. R. Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters", IEEE Trans. Audio and Electroacoustics, December 1973, pp. 506-525.

[28] A. Antoniou, "Accelerated Procedure for the Design of Equiripple Nonrecursive Digital Filters", IEE Proc., February 1982, pp. 1-9.

[29] G. Cortelazzo, M. R. Lighter, W. K. Jenkins, "Frequency Domain Design of Multiband Finite Impulse Response Digital Filters Based on the Minimax Criterion", IEEE Int. Symp. on Circuits and Systems, 27-29 April, 1981, pp. 532-535.

[30] F. Grenez, "Design of Linear or Minimum-Phase FIR Filters by Constrianed Chebyshev Approximation", Signal Processing, vol. 5, No. 4, July 1983, pp. 325-332.

[31] S. A. Tretter, Introduction to Discrete-Time Signal Processing, John Wiley and Sons Inc. 1976.

[32] D. C. Farden and L. L. Scharf, "Statistical Design of Nonrecursive Digital Filters", IEEE Trans. on ASSP, June 1974, pp. 188-196.

[33] H. Clergeot and L. L. Scharf, "Connections Between Classical and Statistical Methods of FIR Digital Filter Design", IEEE Trans on ASSP, October 1978, pp. 463-465.

[34] R. Kumar, "A Fast Algorithm for Solving a Toeplitz System of Equations", IEEE Trans on ASSP, vol. ASSP-33, No. 1, pp. 254-268, February 1985.

[35] J. F. Kaiser and R. W. Hamming, " Sharpening the Response of a Symmetric Non-Recursive Filter by Multiple Use of the Same Filter", IEEE Trans. on ASSP, vol. ASSP-25, No. 4, October 1977, pp. 415-422.

[36] S. Nakamura and S. K. Mitra, "Design of FIR Digital Filters Using Tapped Cascaded FIR Subfilters", Circuits Systems and Signal Processing, vol. 1, No. 1, 1982, pp. 43-56.

[37] R. W. Schafer, L. R. Rabiner and O. Herrman, "FIR Digital Filter Banks for Speech Analysis", BSTJ,vol. 54, No. 3, March 1975, pp. 531-544.

[38] J. T. Rubinstein and H. F. Silverman, "Some Comments on the Design and Implementation of FIR Filter Banks for Speech Recognition", ICASSP 83, pp. 812-815.

[39] F. Mintzer, "On Half-Band, Third-Band and N-th Band FIR Filters and Their Design", IEEE Trans. on ASSP, vol. ASSP-30, No. 5, October 1982, pp. 734-738.

[40] J. K. Liang, R.J.P de Figueiredo, and F. C. Lu, "Design of Optimal Nyquist, Partial Response, N-th Band, and Non-uniform Top Spacing FIR Digital Filters Using Linear Programming Techniques", IEEE Transactions on Circuits and Systems, CAS-32, No. 4, April 1985, pp. 386-391.

[41] J. Allen, "Cochlear Modeling", IEEE ASSP Magazine, vol.2, No. 1, January 1985, pp. 3-30.

[42] J. L. Flanagan, Speech Analysis, Synthesis and Perception, 1972, Springer-Verlag, New York.

[43] D. Esteban and C. Galand, "Application of Quadrature Mirror Filters to Split Band Voice Coding Schemes", Proc. ICASSP, 1977, pp. 191-195.

[44] J. D. Johnston, "A Filter Family Designed for Use in Quadrature Mirror Filter Banks", Proc. ICASSP, 1980, pp. 291-294.

[45] V. K. Jain and R. E. Crochiere, "Quadrature Mirror Filter Design in the Time Domain", IEEE Trans. ASSP, Vol. ASSP-32, No. 2, April 1984, pp. 353-362.

[46] C. R. Galand and H. J. Nussbaumer, "New Quadrature Mirror Filter Structures", IEEE Trans. ASSP, vol. ASSP-32, No. 3, June 1984, pp. 522-532.

[47] A. Croisier, D. Esteban and C. Galand, "Perfect Channel Splitting by Use of Interpolation, Decimation, Tree Decomposition Techniques", in Proc. Int. Conf. Inform. Sci. Syst., Petras, Greece, August 1976.

[48] D. Esteban and C. Galand, "Direct Approach to Quasi-Perfect Decomposition of Speech in Sub-bands", in Proc. Int. Cong. Acoust., Madrid, Spain, July 1977.

[49] H. J. Nussbaumer, "Pseudo QMF Filter Bank", IBM Tech. Disclousûre Bull., vol. 24, No. 6, pp. 3081-3087, November 1981.

[50] J. H. Rothweiler, "Polyphase Quadrature Filters, A New Subband Coding Technique", in Proc. ICASSP, Boston, 1983, pp. 1980-1983.

[51] R. W. Schafer and L. R. Rabiner, "Design and Simulation of a Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis", IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No. 3, pp. 165-174, 1973.

[52] J. B. Allen, "Short Term Spectral Analysis, Synthesis and Modification by Discrete Fourier Transform", IEEE Trans. on ASSP, Vol. ASSP-25, No. 3, pp. 235-238, June 1977.

[53] J. B. Allen and L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Syntesis", Proc. of the IEEE, vol. 65, No. 11, pp. 1558-1564, November 1977.

[54] V. K. Jain and R. E. Crochiere, "A Novel Approach to the Design of Analysis/Synthesis Filter Banks", Proc. ICASSP-83, Boston, pp. 228-231.

[55] L. R. Rabiner, Private Communication.

[56] M. Avriel, Nonlinear Programming: Analysis and Methods, Prentice Hall, 1976 (Chapters 4, 5).

[57] A. Dembo and D. Malah, "Design of Digital Filter Banks with Flat Composite Response", Conference on Digital Signal Processing, Florence, Italy, 1984, pp. 22-26.

[58] A. Dembo and D. Malah, "Design of Filter Banks with Specified Composiet Response and Maximum Output SNR", Forth Int. Conf. on Digital Processing of Signals in Communications, Loughborough, England, April 1985.

[59] W. Ledermann and S. Vajda, Ed., Handbook of Applicable Mathematics, Volume I (Algebra), John Wiley and Sons Ltd., 1980, (chapter 7.12, pp. 230-232).

[60] C. L. Lawson and R. J. Hanson, Solving Least Squares Problems, Problems, Prentice-Hall, Inc., Englewood Cliffs, New-Jersey, 1974, (Chapters 18 and 19).

[61] A. Dembo and D. Malah, "WMMSE Design of Digital Filter Banks with Specified Composite Response", Submitted for publication to the IEEE Trans on ASSP, also summarized in Appendix A.

[62] R. T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1970.

[63] W. Rudin, Principles of Mathematical Analysis, McGraw Hill, 1953.

[64] H. Abut and S. A. Luse, "Vector Quantizers for Sub-Band Waveforms", Proc. ICASSP, pp. 10.6.1-10.6.4, 1984.

[65] R. E. Crochiere, "On the Design of Sub-Band Coders for Low Bit Rate Speech Communication", BSTJ, pp. 741-771, May-July 1977.

[66] N. S. Jayant and P. Noll, Digital Coding of Waveforms, Prentice Hall, N.J., 1984, (Section 4.7).

[67] P. J. Davis, Circulant Matrices, John Wiley and Sons, NY, 1979 (chapters 2,3).

[68] D. G. Luenberger, Introduction to Linear and Non-Linear Programming, Reading MA: Addison Wesley, 1973 (Chapter 6.5).

[69] C. G. Broyden, Basic Matrices, The Macmillan Press, 1975 (Chapters 3, 5).

[70] A. Gersho, T. Ramstad and I. Versvik, "Fully Vector-Quantized Subband Coding with Adaptive Codebook Allocation", ICASSP 1984, pp. 10.7.1-10.7.4.

[71] Y. Katznelson, An Introduction to Harmonic Analysis, John Wiley and Sons, NY, 1968 (chapter 1).

[72] P. Lancaster, Theory of Matrices, 1969 Edition of Academic Press (Chapter 3).

[73] R.M. Gray, "Toeplitz and Circulant Matrices: II", Revisited version of Report No. 6502-1, Stanford University Information System Laboratory.

נספח א': תכנון מערכי מסננים אופטימליים בקריטריון WMMSE עם תגובה כוללת מוכתבת

1. The WMMSE Design of Filter Banks

The formulation of the design problem is as follows:

(A). The filter bank is composed of N individual digital filters. The i-th filter is a linear combination of M_i basic components having frequency responses $E_{ik}(f)$ $k=1,2,\cdots,M_i$ (for a conventional FIR filter, the basic components are delays, and thus $E_{ik}(f)$ takes the form $E_{ik}(f)=e^{-j2\pi(k+l_i)f}$).

All the results to be derived in this section are for complex filter banks. Thus, the coefficients of the linear combinations (denoted by a_{ik}, $k=1,\cdots,M_i$) are assumed to be complex numbers. In the next section we state and prove sufficient conditions for the realness of these coefficients.

The frequency response of the i-th filter is therefore:

$$H_i(f) \triangleq \sum_{k=1}^{M_i} a_{ik} E_{ik}(f) \qquad (1)$$

(B). The desired frequency response of the i-th filter is denoted by $D_i(f)$, $i=1,\cdots,N$. The error between this desired frequency response and the frequency response of the corresponding filter is weighted according to a specified (real) weight function $W_i(f)^2$.

We use the Mean Square Error (MSE), as the error norm, and therefore the i-th filter response error is defined as:

$$\delta_i^2 \triangleq \int_{-0.5}^{0.5} W_i(f)^2 |D_i(f)-H_i(f)|^2 df \qquad (2)$$

(C). The composite response of the filter bank is the sum of the responses of the individual filters. Let the composite frequency response be denoted by $H_{N+1}(f)$. Thus, $H_{N+1}(f)=\sum_{i=1}^{N} H_i(f)$. The specifications on the composite response are given by a desired composite frequency response denoted by $D_{N+1}(f)$, and by a (real) weight function $W_{N+1}(f)^2$ related to the MSE norm of the composite response. Therefore, the composite response error is given by:

$$\delta_{N+1}^2 = \int_{-0.5}^{0.5} W_{N+1}(f)^2 |D_{N+1}(f) - H_{N+1}(f)|^2 df \qquad (3)$$

E.g., if a flat composite response is specified $|D_{N+1}(f)| = 1$, and for equal error weighting in frequency $W_{N+1}(f) = 1$ as well.

(D). The performance of the filter bank is measured in terms of a weighted combination of the individual filters response *errors* . The i-th coefficient of this combination, denoted by K_i^2, reflect the relative importance of the i-th filter specification. Thus, $K_i = 0$ means that the frequency response of the i-th filter can be set arbitrarily (but subject to fulfilling the composite response specifications), whereas $K_i \to \infty$ means that the frequency response of the i-th filter should be as close as possible to its desired frequency response, regardless of the composite response specifications.

The overall weighted MSE is denoted by ε^2, and is thus given by:

$$\varepsilon^2 \triangleq \sum_{i=1}^{N} K_i^2 \delta_i^2 \qquad (4)$$

(E). Two kinds of composite response specifications are possible. The first is a tolerance specification, stated by the constraint $\delta_{N+1}^2 \le \eta^2$, and the second is an indirect specification, by incorporating the composite response error δ_{N+1}^2 into the weighted MSE:

$$\varepsilon_t^2 \triangleq \varepsilon^2 + K_{N+1}^2 \delta_{N+1}^2 \qquad (5)$$

(F). The design problem is to find the optimal set of $M_a = \sum_{i=1}^{N} M_i$ coefficients $\left\{a_{ik}\right\}_{k=1,\, i=1}^{M_i,\, N}$. The optimization criterion is minimization of ε_t^2, or minimization of ε^2 subject to the composite response constraint.

Therefore, two different optimization problems can be stated:

$$I. \quad \min_{\{a_{ik}\}_{i,k}} \{\varepsilon_t^2\} \qquad (6a)$$

$$II. \quad \min_{\{a_{ik}\}_{i,k},\ \delta_{N+1}^2 \leq \eta^2} \{\varepsilon^2\} \qquad \text{(6b)}$$

(C). Both ε^2 and δ_{N+1}^2 are convex functions of the unknown variables. From the theory of convex programming [56, Sec. 4.5], it follows immediately that the two optimization problems are equivalent. We therefore start by deriving the solution of the first optimization problem, and then describe how $K_{N+1}(\eta)$ is found, in order to use this solution for the second optimization problem.

Before we derive the optimal filter bank solution, we focus on two extreme composite response specifications:

(1). As $\eta \to \infty$ ($K_{N+1} \to 0$), an arbitrary composite response is allowed. Therefore in order to minimize $\varepsilon^2(\varepsilon_l^2)$, we can minimize the N individual filter errors $\{\delta_i\}_{i=1}^N$ separately. The original optimization problem is thus converted into N simpler optimization problems. For conventional FIR structure the solution of each of the N optimization problems is the Wiener filter derived in [32,33].

(2). As $K_{N+1} \to \infty$ (η approaches its minimal possible value), the optimal composite response is obtained. If the desired composite response can be met by any filter bank of the prescribed structure (i.e. if there is at least one set of $\{a_{ik}\}_{i,k}$ for which $\delta_{N+1}^2 = 0$), then it is guaranteed that this composite response is achieved by the proposed design method. If this desired composite response is not feasible, the resulting filter bank will have a composite response that is its best possible approximation in the MSE sense.

Solution of the First Optimization Problem

ε_l^2 is clearly a p.s.d. quadratic form of the unknown variables $\left\{a_{ik}\right\}_{i=1,k=1}^{M_i,N}$.

Thus the optimal set of coefficients is given by a solution of a set of M_a linear equations. However, in most practical applications, the basic components of all the N individual filters are taken out of a se' of only $M_{N+1} << M_a$ distinct

elements (e.g., for conventional FIR structures $E_{ik}(f)$ represents delays and $M_{N+1} \triangleq \max_{i=1,\cdots,N}\{M_i\}$ which is the largest delay in the filter bank).

In this case the size of the set of linear equations is reduced to M_{N+1}, thus reducing dramatically the complexity of the design, as elaborated further in Section IV. In order to exploit this property we introduce the following notation:

The composite frequency response $H_{N+1}(f)$ is a linear combination of the frequency responses of all the M_{N+1} distinct basic components. We order these M_{N+1} basic components arbitrarily and denote them by $E_{(N+1)k}(f)$ $k=1,\cdots,M_{N+1}$. We denote the coefficients of the linear combination by $a_{(N+1)k}$ and thus:

$$H_{N+1}(f) = \sum_{k=1}^{M_{N+1}} a_{(N+1)k} E_{(N+1)k}(f) \tag{7}$$

Let the vector $\underline{a}_i \in \mathbb{C}^{M_i}$ represent the coefficients of the i-th filter for $i=1,\ldots,N+1$ (where $\underline{a}_{N+1}$ represents the above coefficients of the composite response).

In what follows, an augmented version of any vector $\underline{v}_i \in \mathbb{C}^{M_i}$ is a vector in $\mathbb{C}^{M_{N+1}}$ denoted by $(\underline{v}_i)^{aug}$ as defined below:

The k-th element of the vector $(\underline{v}_i)^{aug}$ is zero if $E_{(N+1)k}(f)$ is not a basic component of the i-th filter. Otherwise, if $E_{(N+1)k}(f)=E_{im}(f)$ for some m, $1 \le m \le M_i$, then the k-th element of the vector $(\underline{v}_i)^{aug}$ is the m-th element of $\underline{v}_i$.

From this definition, and equation (7) it follows:

$$\underline{a}_{N+1} = \sum_{i=1}^{N} (\underline{a}_i)^{aug} \tag{8}$$

It is easily verified that the augmentation operation is a one to one mapping of $\mathbb{C}^{M_i}$ into $\mathbb{C}^{M_{N+1}}$. The reduced version of a vector $\underline{v} \in \mathbb{C}^{M_{N+1}}$, denoted by $\underline{v}]_{M_i} \in \mathbb{C}^{M_i}$, is defined as follows: If $\underline{v}$ is in the range of the augmentation

operation, i.e. $\underline{v}=(\underline{v}_i)^{aug}$ then $\underline{v}]_{M_i}=\underline{v}_i$. Otherwise, the vector $\underline{v}$ is projected into the range of the augmentation by replacing the appropriate $(M_{N+1}-M_i)$ elements by zeros and is then reduced to $\mathbb{C}^{M_i}$ as defined above. Augmentation (reduction) of square matrices is done by augmenting (reducing) both the columns and the rows.

Note that augmentation (reduction) from $\mathbb{C}^{M_{N+1}}$ to itself is an identity operation, hence in the sequel we use augmentation symbols for matrices and vectors in $\mathbb{C}^{M_{N+1}}$ as well if it is convenient for the presentation.

In the sequel a super-bar denotes complex conjugation and $\underline{v}^H$ denotes conjugate transposition of $\underline{v}$.

Substituting (1-4) and (7) in equation (5) and rearranging the expression of ε_i^2 in terms of the coefficient vectors $\{\underline{a}_i\}_{i=1}^{N+1}$ we obtain the following alternative expression for the optimization problem in (6a):

$$\min_{\{\underline{a}_i\}_{i=1}^{N}} \sum_{i=1}^{N+1} K_i^2[(\underline{a}_i-\underline{a}_i^o)^H \mathbf{R}_i(\underline{a}_i-\underline{a}_i^o)+\hat{\delta}_i^2] \qquad (9)$$

Where $\underline{a}_{N+1}$ is given in (8), and $\underline{a}_i^o \triangleq \mathbf{R}_i^{-1}\underline{d}_i$ is a vector in $\mathbb{C}^{M_i}$.

The elements of the square matrix $\mathbf{R}_i$ are:

$$R_i(m,k) = \int_{-0.5}^{0.5} W_i(f)^2\bar{E}_{im}(f)E_{ik}(f)df \quad i=1,\cdots,N+1\ ;\ m,k=1,\cdots,M_i \qquad (10)$$

The elements of $\underline{d}_i \in \mathbb{C}^{M_i}$ are:

$$d_i(m) = \int_{-0.5}^{0.5} W_i(f)^2\bar{E}_{im}(f)D_i(f)df \quad i=1,\cdots,N+1\ ;\ m=1,\cdots,M_i \qquad (11)$$

The value of $\hat{\delta}_i^2$ is:

$$\hat{\delta}_i^2 = \int_{-0.5}^{0.5} W_i(f)^2|D_i(f)|^2df - \underline{d}_i^H\mathbf{R}_i^{-1}\underline{d}_i \quad i=1,\cdots,N+1\ ; \qquad (12)$$

Note that the WMMSE approach in [33] leads to the solution $\underline{a}_i=\underline{a}_i^o$ $i=1,\cdots,N$, for which $\delta_i^2=\hat{\delta}_i^2$ is minimal for $i \le N$. However, the optimal set of coefficients of the composite response, which is $\underline{a}_{N+1}^o$, is in general not equal to

the augmented sum of these filters. Therefore, this filter bank is not necessarily the solution of (9).

The optimization problem stated in (9) is the minimization of a p.s.d. quadratic form. Its analytical solution (obtained by differentiation with respect to the unknown variables) is:

$$\underline{a}_i = \underline{a}_i^o + \frac{1}{K_i^2} \mathbf{R}_i^{-1} \underline{q}\rfloor_{M_i} \quad i=1,\cdots,N \tag{13}$$

Where $\underline{q} \in \mathbb{C}^{M_{N+1}}$ is a correction vector due to the composite response specification and is given by solving the following set of linear equations:

$$\left[\sum_{i=1}^{N+1} \frac{1}{K_i^2} (\mathbf{R}_i^{-1})^{aug}\right] \underline{q} = \underline{p} \tag{14}$$

The vector $\underline{p}$ is the difference between the optimal set of coefficients of the composite response and the augmented sum of the optimal individual filters, i.e.:

$$\underline{p} = \underline{a}_{N+1}^o - \sum_{i=1}^{N} (\underline{a}_i^o)^{aug} \tag{15}$$

The resulting errors are:

$$\delta_i^2 = \hat{\delta}_i^2 + \frac{1}{K_i^4} \underline{q}\rfloor_{M_i}^H \mathbf{R}_i^{-1} \underline{q}\rfloor_{M_i} \quad i=1,\cdots,N+1 \tag{16}$$

This completes the solution of the first optimization problem under the following two restrictions:

(a). We assume that each of the weight factors K_i is neither zero nor approach infinity, i.e. $0<K_i<\infty \quad i=1,\cdots,N+1$.

(b). We assume that all the p.s.d. matrices that appear in equations (9)-(14), are regular matrices (i.e. p.d.) and thus their inverses exist.

We will now extend the results for weight values which are either zero or approach infinity.

If $K_j \to \infty$, the j-th filter desired response over-rides all other specifications, thus forcing the j-th filter to have the minimal error $\hat{\delta}_i^2$ (i.e. in (13) we get in the limit, as $K_j \to \infty$, that $\underline{a}_j = \underline{a}_j^0$). This affects the correction vector $\underline{q}$ by omitting $(R_j^{-1})^{aug}/K_j^2$ from (14), since this value approaches zero as $K_j \to \infty$. However, as shown in the sequel, increasing the value of K_j results in an increase of the overall error of the remaining filters.

In particular, $K_{N+1}^2 \to \infty$ corresponds to a constraint on the composite response, and in this case R_{N+1}^{-1}/K_{N+1}^2 is omitted in (14), thus increasing the overall error of the individual filters. The solution then coincides with an earlier result we presented in [58]. As mentioned earlier, the composite response error δ_{N+1}^2 is minimized in that case. However its minimal value $\hat{\delta}_{N+1}^2$ is not necessarily zero.

When $K_j=0$ equations (13) and (14) become singular. Returning to the original problem statement in (9), we observe that for $K_j=0$ the value of $\underline{a}_j$ affects the error measure only through the composite response. Taking the derivative of ε_t^2 in (9) with respect to $\underline{a}_j$, we find that the optimal value of $\underline{a}_j$ is the one that guarantees $\underline{q}]_{M_j} \triangleq K_{N+1}^2 R_{N+1}(\underline{a}_{N+1} - \underline{a}_{N+1}^0)]_{M_j} = \underline{0} \in \mathbb{C}^{M_j}$. Therefore, only the $(M_{N+1}-M_j)$ elements of $\underline{q}$ which are not in $\underline{q}]_{M_j}$ have to be evaluated via equation (14). The terms $\frac{1}{K_i^2}(R_i^{-1})^{aug}\underline{q}$ in equation (14) were substituted there instead of $(\underline{a}_i - \underline{a}_i^0)^{aug}$, according to equation (13). The latter equation is valid now only for $i \neq j$ (since $K_j=0$). Therefore equation (14) is modified to:

$$\left[\sum_{\substack{i=1 \\ i\neq j}}^{N+1} \frac{1}{K_i^2}(R_i^{-1})^{aug}\right]\underline{q} = \underline{p} - (\underline{a}_j - \underline{a}_j^0)^{aug} \tag{17}$$

Substituting $\underline{q}]_{M_j} = \underline{0}$ in equation (17), gives the following reduced set of $(M_{N+1}-M_j)$ linear equations whose solution is $\underline{q}]_{M_{N+1}-M_j}$ (i.e. the elements of $\underline{q}$ which are not in $\underline{q}]_{M_j}$):

$$\left[\sum_{\substack{i=1\\i\neq j}}^{N+1} \frac{1}{K_i^2}(\mathbf{R}_i^{-1})^{aug}\right]_{M_{N+1}-M_j} \underline{q}]_{M_{N+1}-M_j} = \underline{p}]_{M_{N+1}-M_j} \tag{18}$$

Note that matrix reduction means omitting both columns and rows.

We summarize below the modified design process in this case:

(a). Find the $(M_{N+1}-M_j)$ elements of $\underline{q}$ which are not in $\underline{q}]_{M_j}$ by using equation (18).

(b). Complete the correction vector $\underline{q}$ using $\underline{q}]_{M_j}=\underline{0}$.

(c). The coefficients of all the filters, except for the j-th filter, are given by equation (13), and the resulting errors by equation (16).

(d). The j-th filter coefficients are now evaluated from the constraint: $\underline{q}]_{M_j} = \underline{0}$.

The above discussion considered in detail the case in which $K_j=0$ for some $j \leq N$. A more important case is the one in which $K_{N+1}=0$ (i.e. an unconstrained composite response). Following the same arguments it is easy to verify that now the optimal filter bank has a zero correction vector, and thus the optimal filter bank is based on the optimal individual filters $\{\underline{a}_i^o\}_{i=1}^N$. Thus, in this case, the new method coincides with previous design methods [32,33], which are applicable only if the composite response is unconstrained.

So far we considered the situation in which one of the K_i's is either zero or tends to infinity. However, the modifications of the basic algorithms that were derived for these two situations can easily be extended to the general case in which some of the K_i's are zero and some others are approaching infinity. Since, for each value of K_i that approaches infinity the corresponding term $\frac{1}{K_i^2}(\mathbf{R}_i^{-1})^{aug}$ is omitted from equation (14), the system of linear equations defined there may become singular.

We refer now to this issue. By definition, the $(N+1)$ matrices $\mathbf{R}_i$ are hermitian ($\mathbf{R}_i^H=\mathbf{R}_i$) p.s.d. matrices. The matrix $\mathbf{R}_i$ is singular iff there exists a non-zero

set of the i-th filter coefficients for which $\int_{-0.5}^{0.5} W_i(f)^2 |H_i(f)|^2 df = 0$, and hence the i-th filter error does not have a unique global minima. A sufficient condition for $\mathbf{R}_i$ to be a non-singular matrix is that $W_i(f)^2 > 0$ on a set of frequency points of non-zero measure, and that the functions $E_{ik}(f)$ are linearly independent functions on this set of points. For example $W_i(f)^2 > 0$ in an interval and $E_{ik}(f) = e^{-j2\pi f(k+l_i)}$ obey this condition. Hence, for all practical purposes we can assume that $\mathbf{R}_i$ are non-singular matrices. It is easy to verify that therefore $\mathbf{R}_i^{-1}$ are also p.d. and hence the matrix which appears in equation (14) is p.s.d., being an augmented positive linear combination of p.d. matrices. This matrix is singular iff there is at least one function $E_(N+1k)(f)$ which is not a component of any of the filters with $K_i < \infty$, and in this case it has at least one zero row and column. If this occurs, one simply has to omit the irrelevant basic components from the design, since the corresponding coefficients do not affect the performance of the filter bank. There is one exception to this rule, which is when both K_{N+1} and at least one of the K_i values approach infinity. In this case the element of $\underline{p}$ corresponding to the zero row of the matrix may be non-zero, reflecting a contradiction in the design specifications which have to be modified. Thus we have seen that if the design problem is well defined all the matrices in the design process are non-singular.

Solution of the Second Optimization Problem

The second optimization problem, stated in (6b), can be solved by converting it to the problem in (6a) which we just solved. This is done by finding the weighting factor K_{N+1}, which incorporates the composite response specification into (6a), from the given tolerance η on the composite response error. We describe now an algorithm for computing $K_{N+1}(\eta)$.

Re-writing equation (14), we obtain the following relation between $\underline{q}$ and the value of K_{N+1}:

$$\left[\frac{1}{K_{N+1}^2}\mathbf{R}_{N+1}^{-1} + \mathbf{T}\right]\underline{q} = \underline{p} \tag{19}$$

where $\mathbf{T} \triangleq \sum_{i=1}^{N} \frac{1}{K_i^2}(\mathbf{R}_i^{-1})^{aug}$ is an $M_{N+1} \times M_{N+1}$ matrix which is independent of K_{N+1}. Equations (16) and (19) give an implicit relation between the value of δ_{N+1}^2 and K_{N+1}. In order to find an explicit relation, we make a change of basis in equation (19) so that both $\mathbf{R}_{N+1}^{-1}$ and $\mathbf{T}$ become diagonal in the new basis of $\mathbb{C}^{M_{N+1}}$. For that purpose we use the following lemma which is easily derived from theorem 7.12.2 in [59]:

Lemma 1: For any two hermitian matrices $\mathbf{A}$ and $\mathbf{B}$, with $\mathbf{A}$ being a p.d. matrix, there exists a non-singular matrix $\mathbf{V}$ such that $\mathbf{V}^H\mathbf{A}\mathbf{V}=\mathbf{I}$ and $\mathbf{V}^H\mathbf{B}\mathbf{V}=\mathbf{D}$ where $\mathbf{D}$ is a diagonal matrix.

The matrix $\mathbf{R}_{N+1}$ is clearly a p.s.d. hermitian matrix, and for well defined design problems it is non-singular. Thus, $\mathbf{R}_{N+1}^{-1}$ exists and is a p.d. hermitian matrix. The matrix $\mathbf{T}$ is also a p.s.d. hermitian matrix (being an augmented sum of p.d. hermitian matrices $\mathbf{R}_i^{-1}$). Thus, Lemma 1 holds for the pair of matrices $\mathbf{R}_{N+1}^{-1}$ and $\mathbf{T}$ and there is a non-singular matrix $\mathbf{V}$ of dimension $M_{N+1} \times M_{N+1}$ so that:

$$\mathbf{V}^H\mathbf{R}_{N+1}^{-1}\mathbf{V} = \mathbf{I} \tag{20a}$$

$$\mathbf{V}^H\mathbf{T}\mathbf{V} = diag(d_1, \cdots, d_{M_{N+1}}) \tag{20b}$$

We can express d_n in terms of $\underline{v}_n$, the n-th column of $\mathbf{V}$, as follows: $d_n = \underline{v}_n^H\mathbf{T}\underline{v}_n$. Since $\mathbf{T}$ is p.s.d., this expression implies that $d_n \geq 0$. The matrix $\mathbf{V}$ is non-singular, and therefore we can perform a change of variables from $\underline{p}$, $\underline{q}$ to $\hat{\underline{p}}$, $\hat{\underline{q}}$ as follows:

$$\underline{q} = \mathbf{V}\hat{\underline{q}} \tag{21a}$$

$$\hat{\underline{p}} = \mathbf{V}^H\underline{p} \tag{21b}$$

Multiplying equation (19) by $\mathbf{V}^H$ and using (20,21) we obtain M_{N+1} scalar equations:

$$\hat{q}_n \triangleq \frac{K_{N+1}^2 \hat{p}_n}{1+K_{N+1}^2 d_n} \quad n=1,\cdots,M_{N+1} \tag{22}$$

Substituting (22) and (21a) in (16) we obtain the following relation between K_{N+1}^2 and δ_{N+1}^2.

$$\delta_{N+1}^2 = \hat{\delta}_{N+1}^2 + \sum_{n=1}^{M_{N+1}} \frac{|\hat{p}_n|^2}{(1+K_{N+1}^2 d_n)^2} \tag{23}$$

From (16) and the definition of the filter bank error ε^2 in (4) we obtain:

$$\varepsilon^2 = \hat{\varepsilon}^2 + \underline{q}^H \mathbf{T} \underline{q} \tag{24}$$

where $\hat{\varepsilon}^2 = \sum_{i=1}^{N} \hat{\delta}_i^2$ is the error of the optimal filter bank with unspecified composite response ($\eta \to \infty$). Substituting (21a), (20b) and (22) in (24) we obtain:

$$\varepsilon^2 = \hat{\varepsilon}^2 + \sum_{n=1}^{M_{N+1}} d_n \left[\frac{K_{N+1}^2 |\hat{p}_n|}{1+K_{N+1}^2 d_n} \right]^2 \tag{25}$$

Using (20) and (21b) we can evaluate the values of $\lim_{K_{N+1}^2 \to 0} (\delta_{N+1}^2)$ and $\lim_{K_{N+1}^2 \to \infty} (\varepsilon^2)$ from (23) and (25) and obtain:

$$\tilde{\delta}_{N+1}^2 \triangleq \lim_{K_{N+1}^2 \to 0} (\delta_{N+1}^2) = \hat{\delta}_{N+1}^2 + \underline{p}^H \mathbf{R}_{N+1} \underline{p} \tag{26a}$$

$$\tilde{\varepsilon}^2 \triangleq \lim_{K_{N+1}^2 \to \infty} (\varepsilon^2) = \hat{\varepsilon}^2 + \underline{p}^H \mathbf{T}^{-1} \underline{p} \tag{26b}$$

It is easily verified from (23) and (25) that δ_{N+1}^2 is a monotonically decreasing function of K_{N+1}^2, and ε^2 is a monotonically increasing function of K_{N+1}^2. Upper and lower limits of these two functions are $\tilde{\delta}_{N+1}^2$ ($\tilde{\varepsilon}^2$) and $\hat{\delta}_{N+1}^2$ ($\hat{\varepsilon}^2$), respectively. The following important property is obtained by evaluating $d(\delta_{N+1}^2)/d(K_{N+1}^2)$ and $d(\varepsilon^2)/d(K_{N+1}^2)$ from (23) and (25), respectively:

$$\frac{d(\varepsilon^2)}{d(\delta_{N+1}^2)} = -K_{N+1}^2 \tag{27}$$

From (27) it follows that the design curve of ε^2 as function of δ_{N+1}^2 is a monotonically decreasing convex curve, and K_{N+1}^2 has the geometric

interpretation as the slope of the curve at the point $(\delta_{N+1}^2, \varepsilon^2)$. Fig. 3.1a illustrates a typical design curve.

Three different ranges of the value of the tolerance specification η^2 should now be considered:

The first range is $\eta^2 < \hat{\delta}_{N+1}^2$, in which the tolerance specification is actually irrelevant since there exists no filter bank of the given structure that can fulfill this specification. Values in the second range, defined by $\eta^2 \geq \tilde{\delta}_{N+1}^2$, are actually fulfilled by the optimal filter bank which ignores the composite response specifications (the Wiener solution $\underline{a}_i^o$). The third range is $\tilde{\delta}_{N+1}^2 > \eta^2 \geq \hat{\delta}_{N+1}^2$, and in this case, since ε^2 is a monotonically decreasing function of δ_{N+1}^2, $\delta_{N+1}^2 = \eta^2$. Thus, for the latter situation we have to solve the non-linear scalar equation derived from (23), namely $\eta^2 = \delta_{N+1}^2(K_{N+1}^2)$, in order to evaluate K_{N+1}^2. An alternative approach is to draw first the design curve $\varepsilon^2(\delta_{N+1}^2)$ using (23) and (25), then choose the desired point on this curve, and find K_{N+1}^2 geometrically as illustrated in Fig. 3.1b.

If we want to solve the non-linear equation $\eta^2 = \delta_{N+1}^2(K_{N+1}^2)$ by numerical methods we can take advantage of the fact that δ_{N+1}^2 is a monotonically decreasing and convex function of K_{N+1}^2, whose higher order derivatives can be evaluated analytically a-priori. In order to reduce the number of iterations needed in evaluating $K_{N+1}^2(\eta^2)$ by numerical methods, we derive upper and lower bounds on $K_{N+1}^2(\eta^2)$ which are easily computed. We denote by d_o the minimal value among $\{d_n\}_{n=1}^{M_{N+1}}$ and by d_∞ the maximal value in this set. Now the following inequalities hold for all n:

$$1 + K_{N+1}^2 d_o \leq 1 + K_{N+1}^2 d_n \leq 1 + K_{N+1}^2 d_\infty \tag{28a}$$

$$(d_n / d_\infty) + K_{N+1}^2 d_n \leq 1 + K_{N+1}^2 d_n \leq (d_n / d_o) + K_{N+1}^2 d_n \tag{28b}$$

Substituting these inequalities in equation (23) and using (26) and the monotonicity of $K_{N+1}^2(\eta^2)$ we obtain the following upper and lower bounds:

$$\max[\sqrt{B(\eta)}-1/d_0,\ (\sqrt{A(\eta)}-1)/d_\infty] \le K_{N+1}^2 \le \min[\sqrt{B(\eta)}-1/d_\infty,\ (\sqrt{A(\eta)}-1)/d_0] \tag{29}$$

where

$$A(\eta) \triangleq \frac{\underline{p}^H R_{N+1} \underline{p}}{\eta^2-\hat{\delta}_{N+1}^2} = \frac{\tilde{\delta}_{N+1}^2-\hat{\delta}_{N+1}^2}{\eta^2-\hat{\delta}_{N+1}^2} \tag{30a}$$

$$B(\eta) \triangleq \frac{\underline{p}^H T^{-1} R_{N+1} T^{-1} \underline{p}}{\eta^2-\hat{\delta}_{N+1}^2} = \frac{\sum_{n=1}^{M_{N+1}} |\hat{p}_n|^2/d_n^2}{\eta^2-\hat{\delta}_{N+1}^2} \tag{30b}$$

Note: The design curve of $\varepsilon_i^2-K_j^2\delta_j^2$ as a function of K_j^2 has the same properties as the design curve of ε^2 as a function of K_{N+1}^2. Therefore, our remark that an increase in the value of K_j^2 implies an increase in the overall error of the remaining filters (i.e., $\varepsilon_i^2-K_j^2\delta_j^2$), follows as a consequence of (25).

II. Phase Linearity and Realness of Optimal Filter Banks.

In the general model presented in the previous section, we assumed that the designed filter bank has complex coefficients $\left\{a_{ik}\right\}_{i=1,k=1}^{N,M_i}$, and have not considered the issue of phase linearity of the resulting filters. We discuss the subjects of realness and phase linearity in this section.

Two theorems are presented, the first provides a sufficient condition for realness of the optimal filter bank coefficients, and the second provides a sufficient condition for zero phase error in the responses of all N individual filters in the bank. These two theorems are derived for the general structure defined in Section I. However, their interpretation for the important class of FIR filter banks is given by means of corollaries following the relevant theorem. Theorem 1 provides sufficient conditions for the realness of the coefficients of the optimal filter bank.

Theorem I: If $W_i(f)^2 = W_i(-f)^2 \quad i=1,\cdots,N+1$, and a function $\vartheta(f)$ exists such that the following conditions are satisfied:

$$(a).\ D_i(f) = \bar{D}_i(-f)e^{j\vartheta(f)} \qquad i=1,\cdots,N+1$$
$$(b).\ E_{im}(f) = \bar{E}_{im}(-f)e^{j\vartheta(f)}\ i=1,\ldots,N+1;m=1,\cdots,M_i$$

then all the filters in the *optimal* filter bank have real coefficients. Furthermore, the matrices $\{\mathbf{R}_i\}_{i=1}^{N+1}$ and the vectors $\{\underline{d}_i,\underline{a}_i^o\}_{i=1}^{N+1}$, $\underline{p}$, and $\underline{q}$, are all real. Thus, the design process involves then only operations on real numbers.

Proof: It is easily verified from (10), (11), that conditions (a)-(c) are sufficient conditions for realness of the matrices $\mathbf{R}_i$ and the vectors $\underline{d}_i$. Since $\{\underline{a}_i^o\}_{i=1}^{N+1}$, $\underline{p}$, and $\underline{q}$ are given in terms of these values, they all are real vectors and so are the coefficients of the individual filters of the optimal filter bank.

Corollary I: For a filter bank composed of conventional FIR filters (i.e. $E_{ik}(f)=e^{-j2\pi f(k+l_i)}$), the individual filters in the optimal filter bank have real coefficients provided that all the impulse responses related to the desired frequency responses $D_i(f)$ and weight functions $W_i(f)^2$ are real sequences.

Proof: For conventional FIR filters, condition (c) is satisfied with $\vartheta(f)=0$. The corollary thus restates conditions (a) and (b) for $\vartheta(f)=0$, in a slightly different manner.

Theorem II below provides sufficient conditions for exact fulfillment of the desired phase response specifications (up to an integer multiple of π due to sign inversions).

Theorem II: Under the following two conditions:

(a). All the filters in the bank have the *same* desired phase response (up to an integer multiple of π) $\psi(f)$, i.e.:

$D_i(f)=\hat{D}_i(f)e^{j\psi(f)}\ i=1,\cdots,N+1$; with $\hat{D}_i(f)$ being real functions.

(b). The basic components of each filter can be divided into distinct pairs such that in every pair the frequency response of one component is the complex conjugate of the frequency response of the other component multiplied by $e^{j2\psi(f)}$.

Written formally: For every i, $i = 1, \ldots, N+1$, there exists a permutation π_i such that $(\forall\, k)\ (\ E_{ik}(f) = \bar{E}_{i\pi_i(k)}(f)e^{j2\psi(f)}\ )$.

The phase response of each filter in the optimal filter bank is exactly the desired phase response (up to a integer multiple of π) i.e. $H_i(f) = \hat{H}_i(f)e^{j\psi(f)}$ $i=1, \cdots, N+1$; with $\hat{H}_i(f)$ being real functions.

Proof: It is easily verified that conditions (a) and (b) are sufficient for the following results:

(1). $d_i(m) = \bar{d}_i(\pi_i(m))$ for all m and i.

(2). $R_i(m,k) = \bar{R}_i(\pi_i(m),\pi_i(k))$ for all m,k and i.

It is easily proven that from (2) follows:

(3). $R_i^{-1}(m,k) = \overline{R_i^{-1}}(\pi_i(m),\pi_i(k))$ for all m,k and i.

From (1) and (3) follows directly that:

(4). $a_i^o(m) = \bar{a}_i^o(\pi_i(m))$ for all m and i.

Since (3) and (4) hold for $i=N+1$, the augmentation of $\underline{a}_i^o$ and $\mathbf{R}_i^{-1}$ only reorders the pairs of elements according to $\pi_{N+1}(\cdot)$ instead of $\pi_i(\cdot)$, and therefore:

(5). $p(m) = \bar{p}(\pi_{N+1}(m))$ for all m.

(6). $T(m,k) = \bar{T}(\pi_{N+1}(m),\pi_{N+1}(k))$ for all m,k.

Thus:

(7). $q(m) = \bar{q}(\pi_{N+1}(m))$ for all m.

The reduction of $\underline{q}$ to $\underline{q}]_{M_i}$ reorders the pairs of elements according to the permutation $\pi_i(\cdot)$ and therefore:

(8). $q]_{M_i}(m) = \bar{q}]_{M_i}(\pi_i(m))$ for all m and i.

And the final result from (3),(4) and (8) is that the optimal coefficients satisfy:

(9). $a_i(m) = \bar{a}_i(\pi_i(m))$ for all m and i.

Combining (9) above with condition (b) one easily obtains the result that $H_i(f) = \bar{H}_i(f)e^{j2\psi(f)}$, which means that $H_i(f)=\hat{H}_i(f)e^{j\psi(f)}$ with $\hat{H}_i(f)$ being real functions.

Note: Similar results holds for $\hat{D}_i(f)$ being pure imaginary functions, with $\hat{H}_i(f)$ being pure imaginary, and a (-) sign in (1)-(9).

Corollary II: For filter banks composed of conventional FIR filters (i.e. $E_{ik}(f)=e^{-j2\pi f(k+l_i)}$), with the additional delay values being $l_i = \frac{1}{2}(M_{N+1}-M_i)-1$, so that all N filters have the same delay, and desired frequency responses $D_i(f)$ which have the same linear-phase response $\psi(f) = -\pi f(M_{N+1}-1)$, the optimal filters are also linear-phase filters.

Proof: For $l_i = \frac{1}{2}(M_{N+1}-M_i)-1$, and $\psi(f) = -\pi f(M_{N+1}-1)$ it is easily verified that condition (b) of Theorem II holds for $\pi_i(n)=(M_i+1-n)$. Condition (a) was satisfied in the corollary statement and thus the result follows from Theorem II.

It should be noted that the phase-linearity of the optimal filter bank is not obtained when odd-length and even-length filters are mixed together in the same filter bank, since then some of the additional delays of the individual filters (l_i values) involve half-sample delays, which are difficult to realize.

III. Statistical Interpretation of the WMMSE Method

The statistical interpretation of the WMMSE criterion for the design of a single FIR filter was presented in [32]. We present here its extension to the design of filter banks, composed of FIR filters, with a specified composite response. This interpretation is useful for applications in which the input process has a statistical characterization.

Since each filter in the filter bank is usually designed to pass a different frequency-band of the common input signal, we may define differently the so called signal and noise components for each filter in the bank. The convention taken here is to consider all the frequency components of the common input which are in the passband of the i-th filter as its input signal s_i and all the components in the stopband as noise n_i. Because we deal with a filter design problem, components in the transition bands of each individual filter are ignored. Thus we view each filter as having its own input denoted by $x_i = s_i + n_i$, for the i-th filter in the bank. Note that according to the above convention the inputs $x_i, i=1, \cdots, N$ are not identical, unless all the transition bands are eliminated (i.e. set to have zero bandwidth). For the mathematical development it is convenient to apply the following vector notation:

The impulse response of the i-th filter, which is of length M_i, is denoted by the vector $\underline{a}_i$. The input vector, which comprises of M_i consecutive samples of the random process x_i, is denoted by $\underline{X}_i(k)$, i.e. $\underline{X}_i(k) = [x_i(k), \cdots, x_i(k-(M_i-1))]^T$. Thus the corresponding output is $y_i(k) = \underline{a}_i^T \underline{X}_i(k)$. As explained above we regard input samples as being the sum of signal samples and noise samples, and we assume that they are samples of two wide-sense stationary continuous random processes. The *desired* signal at the i-th output at time k is defined to be the delayed version of the input signal, i.e. $y_i^d(k) = s_i(k-\rho_i)$. We divert now from the usual convention of assuming that the

signal component at the output is the response of the filter to the signal component at the input and instead set the signal component at the i-th filer *output* to be the desired response $y_i^d(k)$, which is independent of the filter $\underline{a}_i$. This way, the noise component at the output of the i-th filter contains both the filtered input noise, and the distortions of the input signal introduced by the i-th filter. With these assumptions the signal power at the output of the i-th filter is given by:

$$S_{oi} = E[\,|y_i^d(k)|^2] = \sigma_{s_i}^2 \tag{31}$$

and the corresponding noise power is:

$$N_{oi} = E[\,|y_i(k)-y_i^d(k)|^2] = (\underline{a}_i-\underline{a}_i^o)^H R_i(\underline{a}_i-\underline{a}_i^o)+\hat{\delta}_i^2 \tag{32}$$

where:

$$\underline{a}_i^o \triangleq R_i^{-1}\underline{d}_i \tag{33}$$

and:

$$\hat{\delta}_i^2 \triangleq \sigma_{s_i}^2-\underline{d}_i^H R_i^{-1}\underline{d}_i \tag{34}$$

R_i is a $M_i \times M_i$ autocorrelation (Toeplitz) matrix defined by:

$$R_i \triangleq E[\,\underline{\bar{X}}_i(k)\underline{X}_i(k)^T] \tag{35}$$

with $\underline{d}_i \in \mathbb{C}^{M_i}$ defined by:

$$\underline{d}_i \triangleq E[\,s_i(k-\rho_i)\underline{\bar{X}}_i(k)\,] \tag{36}$$

$\underline{a}_i^o$ is exactly the Wiener filter coefficient vector which minimizes the output noise power of the i-th filter [32]. This filter also maximizes the output SNR of the i-th filter since S_{oi} is independent of the filter $\underline{a}_i$. Independent designs of the individual filters, using the Wiener filters, may result however in a poor composite response. To solve this problem a composite response specification is now incorporated into the design process. The desired composite response is specified as the frequency response of a desired FIR filter $\underline{a}_{N+1}^o$ (e.g., $\underline{a}_{N+1}^o$ which is a unit vector represents a flat composite response). Since each filter has a different length and delay, the composite response of the filter bank is an aug-

mented sum of the coefficients of the individual filters, i.e.: $\underline{a}_{N+1} \triangleq \sum_{i=1}^{N} (\underline{a}_i)^{aug}$. The augmentation operation takes care of the different lengths as well as the additional delays needed. The composite response error measure is a weighted MSE between the desired response $\underline{a}_{N+1}^o$ and the actual response $\underline{a}_{N+1}$, i.e. the composite response "noise" is:

$$\mathrm{N}_{o(N+1)} = (\underline{a}_{N+1}-\underline{a}_{N+1}^o)^H \mathbf{R}_{N+1}(\underline{a}_{N+1}-\underline{a}_{N+1}^o) \tag{37}$$

where $\mathbf{R}_{N+1}$ is an $M_{N+1} \times M_{N+1}$ p.d. matrix (with $M_{N+1}=\max_i(M_i)$). This specific error measure is used in order to obtain a statistical interpretation to the WMMSE criteria. If $\mathbf{R}_{N+1}$ is a Toeplitz matrix, one can interpret $\mathrm{N}_{o(N+1)}$ as the weighted L_2 norm of the composite frequency response error. Using Parseval's theorem, the frequency weight function is given by:

$$W_{N+1}(f)^2 = F\{R_{N+1}(k+d,k)\} \tag{38}$$

With $F\{\cdot\}$ representing the Fourier transform of the autocorrelation sequence with respect to the variable d.

The optimal filter bank is the filter bank with the minimal weighted sum of output noise powers, among all the filter banks having composite response noise power which is below η^2. Written formally:

$$\min_{\{\underline{a}_i\}_{i=1}^N,\ \mathrm{N}_{o(N+1)} \le \eta^2} \sum_{i=1}^{N} K_i^2((\underline{a}_i-\underline{a}_i^o)^H \mathbf{R}_i(\underline{a}_i-\underline{a}_i^o)+\hat{\delta}_i^2) \tag{39}$$

This is exactly the second optimization problem presented in Section I, for the special case of FIR filters (see (9) for comparison).

Since the output signal powers of the individual filters are independent of the coefficients of the filters, it follows that the above defined optimal filter bank also maximizes the weighted harmonic mean of the output signal-to-noise ratios, i.e. it is the solution of:

$$\max_{\{\underline{a}_i\}_{i=1}^{N},\, \mathbf{N}_{o(N+1)} \leq \eta^2} \left\{ \frac{1}{\frac{1}{N}\sum_{i=1}^{N} C_i \left(\frac{\mathbf{N}_{oi}}{\mathbf{S}_{oi}}\right)} \right\} \tag{40}$$

Where $C_i \triangleq \sigma_{s_i}^2 K_i^2$.

We have thus presented the equivalence in the filter bank design problem between minimal-noise powers, maximal output signal-to-noise ratios and WMMSE criteria. Furthermore, we can relate the desired frequency responses and the weighting functions to signal and noise spectra by comparing the statistical and deterministic definitions of $\mathbf{R}_i$ and $\underline{d}_i$. This is done under the assumption that $\rho_i = \frac{1}{2}(M_i - 1)$ and $l_i = \frac{1}{2}(M_{N+1} - M_i) - 1$, and therefore Corollary II from the previous section holds, and all the individual filters have linear-phase. It follows that each weighting function represents the spectrum of the corresponding input and each desired frequency response is the cross-spectra of the corresponding input and its signal component divided by the spectrum of the input. Written formally:

$$W_i(f)^2 = F\{E[\bar{x}_i(k)x_i(k+d)]\} \quad i=1,\cdots,N. \tag{41}$$

$$W_i(f)^2 |D_i(f)| = F\{E[x_i(k)s_i(k+d)]\} \quad i=1,\cdots,N. \tag{42}$$

In communication applications the input process is characterized by its autocorrelation sequence and its crosscorrelation with the desired signal. Equations (41) and (42) enable the use of the new design method for these applications by suggesting a way of choosing the weight functions and desired frequency responses in terms of the autocorrelation and crosscorrelation sequences. Furthermore, subject to these relations, equation (40) gives an interpretation of the design criteria in terms of the output SNR's.

IV. On the Complexity of the WMMSE Design Method

The design of an optimal filter bank is composed of the following steps:

(a). Evaluation of R_i and $\underline{d}_i$ from the specified frequency responses (according to (10), (11)).

(b). Calculation of R_i^{-1} and $\underline{a}_i^0$ (to obtain the filter bank for unspecified composite response).

(c). Computation of $\underline{p}$; where if either $\underline{p}=\underline{0}$ or $K_{N+1}=0$ the design is complete.

(d). For specified values of $K_{N+1}>0$, find $\underline{q}$ by solving equation (14).

(e). For specified values of $\eta>0$:

(1). Find the matrix V and the values of $\{d_1 \cdots d_{M_{N+1}}\}$ defined in (20).

(2). Compute $\hat{\underline{p}}$ and solve the non-linear scalar equation (23) for K_{N+1}^2.

(3). Given the value of K_{N+1}^2, $\underline{q}$ is computed via equations (22) and (21a).

(f). Once $\underline{q}$ is known the filters coefficients are obtained by equation (13).

We analyze now the complexity of each of the above steps.

Step (a): In general there are $O(\sum_{i=1}^{N+1} M_i^2)$ integrals to be evaluated in this step. It is highly complicated to evaluate these integrals numerically. However, for weighting functions that are piecewise linear, and basic components that are FIR filters, the integrals that define the matrices R_i can be evaluated analytically.

Let B_i be the number of distinct pieces in the i-th weighting function, then the R_i matrices can be evaluated in $O(\sum_{i=1}^{N+1} M_i^2 B_i)$ operations. The desired frequency responses are present only in $O(\sum_{i=1}^{N+1} M_i)$ integrals, and thus the complexity of step (a) is unaffected whether or not these responses are piecewise linear. For the special case of conventional FIR filters, and piecewise linear desired

responses, the matrices $\mathbf{R}_i$ are Toeplitz matrices, thus only M_i elements have to be evaluated for each matrix, and the integrals involving the desired responses can be evaluated analytically, therefore the overall complexity is $O(\sum_{i=1}^{N+1} M_i \hat{B}_i)$, where $\hat{B}_i$ is the number of distinct pieces in $W_i(f)^2 D_i(f)$.

Step (b) involves solving $(N+1)$ systems of linear equations, or alternatively calculating $(N+1)$ inverses of the matrices $\mathbf{R}_i$. The complexity of this step is thus $O(\sum_{i=1}^{N+1} M_i^3)$. For filters composed of all-pass sections as illustrated in Fig.3.3 the matrices $\mathbf{R}_i$ are Toeplitz matrices. Similarly for filters composed of sections having the same phase-response and powers of a basic magnitude response, as illustrated in Fig. 3.4, the matrices $\mathbf{R}_i$ are Hankel matrices.[1] The first structure coincides with the conventional FIR structure for $\Phi_i(f) = -2\pi f$. Furthermore, this structure seems suitable for the design of filter banks composed of IIR filters. In this case $C_i(f)$ represents an IIR filter designed so that its magnitude response is very close to the desired magnitude response of the i-th filter, and the all-pass sections are used for the phase-correction needed to approximate the desired (linear) phase response. The second structure is especially suitable for the design of filters based on short FIR filters in cascade. In that case, $C_i(f) = e^{-j2\pi f p_i}$ is the delay that guarantees casuality of the i-th filter and $|A_i(f)|$ is the magnitude response of the short prototype FIR filter. This is exactly the structure used in [35,36]. For both structures the matrices $\mathbf{R}_i$ are invertible in $O(M_i^2)$ operations and the overall complexity of step (b) reduces to $O(\sum_{i=1}^{N+1} M_i^2)$.

---

[1] $R_i(m,k) = \int_{-0.5}^{0.5} W_i(f)^2 |C_i(f)|^2 |A_i(f)|^{m+k} df$ gives a Hankel matrix for the structure in Fig. 3.4, and $R_i(m,k) = \int_{-0.5}^{0.5} W_i(f)^2 |C_i(f)|^2 e^{-j\Phi_i(f)(m-k)} df$ gives a Toeplitz matrix for the structure in Fig. 3.3.

Step (c) is of negligible complexity $(O(\sum_{i=1}^{N+1} M_i))$.

Step (d) involves the solution of a set of M_{N+1} linear equations and its complexity is thus $O(M_{N+1}^3)$. This step is of negligible complexity in comparison to step (b) for the general case, but it dominates the complexity of the design when we deal with conventional FIR filters, since the equation matrix in step (d) is not a Toeplitz matrix while all the $\mathbf{R}_i$ matrices are Toeplitz matrices.

Step (e) involves three different operations, out of which the third (reconstructing the vector $\underline{g}$) is of negligible complexity $(O(M_{N+1}^3))$ compared to step (b). The second operation involves solution of a non-linear scalar equation and its complexity can be estimated by $O(M_{N+1}N_{iter})$, where N_{iter} is the number of iterations in the solution (i.e. number of values of K_{N+1}^2 tried until convergence). Since $M_{N+1}>>1$, this complexity can be regarded negligible in comparison to the other design steps. Thus the dominant operation in step (e) is the evaluation of the matrix $\mathbf{V}$ and its complexity is about $O(M_{N+1}^3)$ as discussed in Section VI.

Step (f) which concludes the design process involves matrix vector multiplications and therefore its complexity is $O(\sum_{i=1}^{N+1} M_i^2)$.

Table A-1 summarizes the overall complexity of the design procedure.

TABLE A-1: COMPLEXITY ANALYSIS

Problem Characteristics	Complexity	Dominant Step
General Weighting Functions or Non-FIR Basic Components	$O(\sum_{i=1}^{N+1} M_i^2)$ [2]	(a)
Piecewise Linear Weighting Functions and FIR Basic Components	$O(\sum_{i=1}^{N+1} M_i^3 + \sum_{i=1}^{N+1} M_i^2 B_i)$	(a & b)
Conventional FIR Filters, or Specific Structures (Fig. 2)	$O(M_{N+1}^3 + \sum_{i=1}^{N+1} M_i^2)$	(b & d or e(1))
Conventional FIR Filters, No Composite Response Specification	$O(\sum_{i=1}^{N} M_i^2)$	(b)

Note that no distinction is made between the two types of composite response specifications, since the complexity of steps (d) and (e) is about the same.

V. Design Example

To illustrate the new method, the following design example is presented:

The problem we consider is the design of an octav-band filter bank composed of five filters. The composite response is specified to be flat. The first filter in the bank is a lowpass filter, the last one is a highpass filter and the other three are bandpass filters. The i-th filter has a bandwidth which is twice the bandwidth of the (i-1)-th filter, starting with a lowpass filter having a passband width of 200Hz. The transition bandwidths of the i-th filter are proportional to its bandwidth. Thus, the last (highest) filter has the widest transition band. The individual filters are conventional FIR filters, and the sampling frequency is 8000Hz. In order that all the filters have the same performance, the product

[2]Each operation in this row is a numerical integration.

'filter-length × transition-bandwidth' is set to be about the same for all five filters [18]. For this reason all the weight factors K_i are equal ($K_i = 1$), except for the composite response factor K_6 that takes different values for different designs. For real time applications, an upper bound of 1.68×10^6 multiplies per second is allowed in a particular implementation of the filter bank. This leads to filter lengths ranging from 19 samples to 139 samples (taking advantage of the linear phase). The desired responses $D_i(f)$ $i=1, \cdots, 5$ are the responses of ideal lowpass/bandpass/highpass filters respectively, i.e. set to one in the desired passband and zero elsewhere. Table A-2 summarizes the exact passband / stopband frequencies of the individual filters and their lengths. The magnitude of the desired composite response $D_6(f)$ is unity. All six frequency responses $D_i(f)$ $i=1, \cdots, 6$ have the same linear phase $\psi(f) = -\pi f\, 138$, which corresponds to a delay of 69 samples. Additional delays of $69 - \frac{1}{2}(M_i - 1)$ samples are required so that all five filters have the same delay, and thus the conditions of Corollary II are satisfied and the resulting filters have linear phase.

The weight functions $W_i(f)^2$ for $i=1, \cdots, 6$ are all piecewise constant functions. For each filter in the bank, the weight function $W_i(f)^2$ equals to one in the passband, zero in the transition bands, four in the lower stopband and nine in the upper stopband. For the composite response a unity weight function is used, thus δ_6^2 is the energy of the composite response error. Since $W_i(f)^2$ and $D_i(f)$ $i=1, \cdots, 6$ correspond to real (possibly infinite) sequences, the conditions of Corollary I are satisfied, and the optimal individual filters are all real valued FIR filters. Two extremal values of K_6^2 are used: $K_6^2 = 0$ and $K_6^2 \to \infty$. For the first case, the resulting filter bank is composed of optimal filters that can be designed either by the new method or by the design method in [32,33] since the composite response is of no relevance. This design obtains the minimal weighted MSE for each individual filter $\hat{\delta}_i^2$, and the minimal MSE $\hat{\varepsilon}^2$. However,

since the composite response is ignored in this design, the result is a very poor response as illustrated in Fig. 2.2 by the solid line. The second extremal case is obtained by specifying a flat composite response as a design constraint. This leads to a filter bank with a flat composite response, and the new method minimizes the overall MSE subject to this constraint. The optimal filters obtained using $K_G^2 \to \infty$, are of course degraded with respect to those obtained using $K_G^2=0$, and their MSE is $\tilde{\varepsilon}^2$ i.e., a worst performance than for any finite value of K_G^2. On the other hand these filters minimize the composite response error, which is here zero (i.e., $\delta_G=\hat{\delta}_G=0$), whereas the filters obtained using $K_G^2=0$ have a composite response error of $\delta_G=\tilde{\delta}_G=0.218$. A trade-off between these extremal cases can be obtained either by using finite values of K_G^2 or by choosing a desired point on the $\varepsilon^2(\eta^2)$ curve as illustrated in Fig.3.1.

The frequency responses of the optimal individual filters for the two extremal cases are compared in Fig.3.5. The frequency responses of the filters obtained in the flat composite response design are illustrated in Fig.3.5a, and the frequency responses of the filters obtained in the un-constrained design are in Fig.3.5b. Both are shown on a linear magnitude scale. For further comparison, the frequency response of the fourth individual filter in these two designs is illustrated in Fig.3.6a and Fig.3.6b respectively, on a logarithmic magnitude scale. The values of δ_i^2 obtained in the two extremal designs, and the overall MSE ε^2 are summerized in Table A-2.

It is significant that the two extreme values of ε^2 are quite close to each other (last row in the table), whereas the values of δ_G^2 differ dramatically. Thus, with moderate increase of the MSE a flat composite response is obtained, instead of the poor composite response which results in the design which ignores composite response specifications.

Table A-2: Design Specifications

Filter No. & Type	Filter Length	Lower Stopband [Hz] $D=0, W^2=4$	Passband [Hz] $D=1, W^2=1$	Higher Stopband [Hz] $D=0, W^2=9$	Error In Un-Constrained Design ($\hat{\delta}_i^2$)	Error In Flat-Composite Design ($\tilde{\delta}_i^2$)
1 (LPF)	139	-	0000 - 0200	0300 - 4000	0.504×10^{-2}	1.914×10^{-2}
2 (BPF)	139	0000 - 0200	0300 - 0500	0600 - 4000	0.849×10^{-2}	2.518×10^{-2}
3 (BPF)	79	0000 - 0400	0600 - 1000	1200 - 4000	0.553×10^{-2}	2.881×10^{-2}
4 (BPF)	39	0000 - 0800	1200 - 2000	2400 - 4000	0.816×10^{-2}	3.200×10^{-2}
5 (HPF)	19	0000 - 1600	2400 - 4000	-	0.746×10^{-2}	2.396×10^{-2}
6 (Composite response)	139	-	0000 - 4000	-	46.670×10^{-2}	0
MSE ε	-	-	-	-	1.582×10^{-2}	5.855×10^{-2}

VI. The Complexity of Evaluating the Matrix V.

In this section we investigate the complexity of evaluating the elements of the matrix **V** that appears in Lemma 1. In [59] the following method is applied for evaluating **V** as well as the values of $\{d_n\}_{n=1}^{M_{N+1}}$:

(a). Compute the matrix $\mathbf{C} = \mathbf{R}_{N+1}\mathbf{T}$.

(b). Solve the eigenvalue/eigenvector problem $\mathbf{C}\underline{u}=\lambda\underline{u}$.

It can be shown that for $\mathbf{R}_{N+1}$ and $\mathbf{T}$ which are both hermitian, and $\mathbf{R}_{N+1}$ being p.d., the matrix **C** can be diagonalized.

Now, since **C** is a diagonalizable matrix, there exists a non-singular matrix **U** (whose columns are the eigenvectors) such that: $\mathbf{CU}=\mathbf{U}diag\{d_1 \cdots d_{M_{N+1}}\}$. $\{d_n\}_{n=1}^{M_{N+1}}$ are therefore the eigenvalues of **C**. It can be shown that for $d_n \neq d_m$: $\underline{u}_n^H \mathbf{R}_{N+1} \underline{u}_m = \underline{u}_n^H \mathbf{T} \underline{u}_m = 0$.

(c). If all $\{d_n\}_{n=1}^{M_{N+1}}$ values are distinct, then **V** is obtained from **U** by scaling the columns of **U** as follows: $\underline{v}_n = \underline{u}_n / (\underline{u}_n^H \mathbf{R}_{N+1}^{-1} \underline{u}_n)^{1/2}$.

(d). If the eigenvalue d_n has $m \geq 1$ eigenvectors associated with it, a Grahm-Shmidt orthonormalization process on the sub-space of dimension m of these eigenvectors, will give the m columns of the matrix $\mathbf{V}$ corresponding to this eigenvalue. The orthonormalization process is with respect to the following norm of $\mathbb{C}^{M_{N+1}}$ defined by $(\underline{x},\underline{y})=\underline{x}^H \mathbf{R}_{N+1} \underline{y}$. For $\mathbf{R}_{N+1}$ which is a p.d. hermitian matrix this is a well defined norm, and the case of $m=1$ discussed in section (c) above is only a special case.

The complexity of the matrix multiplication in step (a) above is $O(M_{N+1}^3)$. The complexity of the eigenvector/eigenvalue problem that is solved in step (b) is about $O(M_{N+1}^3)$ using efficient numerical methods [60]. The complexity of the normalization process in steps (c) or (d) is also $O(M_{N+1}^3)$. Thus, the overall complexity of evaluating the elements of the matrix $\mathbf{V}$ is $O(M_{N+1}^3)$. However, this task is certainly more complex than simply inverting an $M_{N+1} \times M_{N+1}$ matrix.

Note that for the special case of a desired flat composite response, $\mathbf{R}_{N+1}=\mathbf{I}$, and step (a) is totally omitted. Furthermore, in step (b) $\mathbf{C}=\mathbf{T}$ is an hermitian p.s.d. matrix, thus it has a unitary diagonalization $\mathbf{V}$ which is the matrix that appears in Lemma 1 (steps (c) and (d) are omitted).

נספח ב' : קיום, יחידות ותכונות כלליות של מחלקה של בעיות אפרוקסימציה וקטורית

I. תאור הבעיה

V הוא מרחב וקטורי לינארי מעל השדה הקומפלקסי $\mathbb{C}$. יהיו $\{v_{ik}\}_{k=1}^{M_i}$ N קבוצות (i = 1,...,N) של וקטורים בלתי תלויים לינארית ב-V, ויהיו $\{d_i\}_{i=1}^{N}$ ו-d_c (N+1) אברים נוספים של V.

יהיו $\{S_i\}_{i=1}^{N}$ N תת-מרחבים לינאריים ממימד סופי $(\dim S_i = M_i)$ של V המוגדרים על ידי:

$$S_i = \{u \in V\ ;\quad u = \sum_{k=1}^{M_i} a_{ik}\, v_{ik}\ ,\quad a_{ik} \in \mathbb{C}\ \}$$

וכן יהיו $\{SD_i\}_{i=1}^{N}$ N תת-מרחבים לינאריים ממימד סופי $(M_i \le \dim SD_i \le M_i+1)$ של V המוגדרים על ידי:

$$SD_i = \{\ u \in V,\ u = \alpha d_i + \hat{u}\ ,\ \hat{u} \in S_i \ \text{and}\ \alpha \in \mathbb{C}\ \}$$

בנוסף יהיו SD_c, S_c שני תת-מרחבים לינאריים ממימד סופי של V המוגדרים על ידי:

$$S_c = \{\ u \in V\ ;\quad u = \sum_{i=1}^{N} \beta_i \hat{u}_i\ ,\ \hat{u}_i \in S_i\} \qquad ,\ \dim S_c = M_c$$

$$SD_c = \{u \in V\ ;\ u = \alpha d_c + \hat{u}\ ,\ \hat{u} \in S_c\ ,\ \alpha \in \mathbb{C}\ \} \qquad M_c \le \dim SD_c \le M_c+1$$

והמקדמים $\{\beta_i\}_{i=1}^{N}$ מהשדה $\mathbb{C}$ מוכתבים מראש.

נגדיר $\{||\cdot||_i\}_{i=1}^{N}$ שהן סמי-נורמות על SD_i ו-$||\cdot||_c$ שהיא סמי-נורמה על SD_c כאשר $||\cdot||_i$ אינן מתאפסות על $S_i - \{o\}$ ו-$||\cdot||_c$ איננה מתאפסת על $S_c - \{o\}$.

נתעניין בבעית הקירוב הסימולטני של: $\underline{d} = (d_1,\dots,d_N) \in V^N$, על ידי $\underline{u} = (u_1,\dots,u_N) \in S_1 \times \dots \times S_N$, כאשר $\sum_{i=1}^{N} \beta_i u_i \in S_c$ מקרב את d_c. לשם כך נתבונן בוקטור $\underline{\delta} \in R^N$, המוגדר על ידי $\delta_i = ||d_i-u_i||_i$, $\underline{\delta} = (\delta_1,\dots,\delta_N)$, ונגדיר "נורמה" נוספת $||\cdot||_T$ על $[o,\infty)^N$ ביחס לסקלריים מ-$[o,\infty)$.

נדרוש מנורמה זו מונוטוניות ביחס לוקטורים ב-$[0,\infty)^N$ קרי:

$$(\forall\ \underline{\delta},\ \underline{\Delta\delta} \in [0,\infty)^N)\ (||\underline{\delta}||_T \leq ||\underline{\delta} + \underline{\Delta\delta}||_T)$$

טיב הקירוב של $\underline{d}$ על ידי $\underline{u}$ ימדד על ידי: $\varepsilon = ||\underline{\delta}||_T \in [0,\infty)$.

והאילוץ ש- $\sum_{i=1}^{N} \beta_i u_i$ מקרב את d_c יאופיין על ידי: $\delta_c = ||d_c - \sum_{i=1}^{N} \beta_i u_i||_c \leq \eta$.

נתעניין בבעיות קיום ויחידות של הקרוב הטוב ביותר תחת האילוץ דנן ובתכונות שלו.

יהא: $\underline{a} = (\underline{a}_1, \underline{a}_2, \dots, \underline{a}_N) \in ¢^{M_a}$ $(M_a = \sum_{i=1}^{N} M_i)$ וקטור המקדמים שמייצג את הקרוב $\underline{u}$ ונתייחס ל- ε ו- δ_c כאל פונקציות של $\underline{a}$, קרי: $\varepsilon = f(\underline{a})$, $\delta_c = g(\underline{a})$.

בנוסף יהא:

$$\bigcup_{i=1}^{N} \{v_{ik}\}_{k=1}^{M_i} \supset \{v_{ck}\}_{k=1}^{M_c}$$

בסיס של S_c (קיים כזה, כי $\bigcup_{i=1}^{N} \{v_{ik}\}_{k=1}^{M_i}$ זו קבוצה פורשית של S_c). אזי קיימת מטריצה $A_{M_c x M_a}$, כך ש-

$$(\forall\ u \in S_c)\ (u = \sum_{k=1}^{M_c} b_k v_{ck})$$

ו- $\underline{b} = A\underline{a}$.

וזאת כי את כל האברים ב- $\bigcup_{i=1}^{N} \{v_{ik}\}_{k=1}^{M_i}$, ניתן להציג כקומבינציה לינארית קבועה של אברי הבסיס (וכל קומבינציה כזו מוכפלת ב-β_i המתאים היא עמודה של A).

לכן: $\delta_c = h(\underline{b}) \underset{\underline{b} = A\underline{a}}{\overset{\uparrow}{=}} g(\underline{a})$.

II. תוצאות

למה 1:

(א) הפונקציות $f(\underline{a})$, $g(\underline{a})$, ו-$h(\underline{b})$ הן קמורות ואי-שליליות.

(ב) $(\forall\ \alpha > 0)\ (\exists M < \infty)\ (\forall\ \underline{a} \in ¢^{M_a},\ ||\underline{a}||_\epsilon > M)\ (f(\underline{a}) > \alpha)$

(ג) $(\forall\ \alpha > 0)\ (\exists M < \infty)\ (\forall\ \underline{b} \in ¢^{M_c},\ ||\underline{b}||_\epsilon > M)\ (h(\underline{b}) > \alpha)$

כש-$||\cdot||_\epsilon$ היא הנורמה האאוקלידית ב-$¢^{M_a}$, $¢^{M_c}$ בהתאמה.

<u>למה 2:</u>

עבור פונקציות $f(\underline{a})$, $g(\underline{a}) = h(A\underline{a})$-ו המקיימות את תנאי למה 1, הרי:

(א) $\underline{a} \in \mathcal{C}^{M_a}$, $\varepsilon_m = \inf f(a)$, מוגדר ומתקבל במינימום.

(ב) $\underline{a} \in \mathcal{C}^{M_a}$, $\eta_m = \inf g(\underline{a})$, מוגדר ומתקבל במינימום.

(ג) לכל $\eta \geq \eta_m$: $\varepsilon(\eta) = \inf\limits_{g(\underline{a}) \leq \eta} f(a)$, מוגדר ומתקבל במינימום.

והקבוצה $D(\eta) \triangleq \{\underline{a} \; ; \; f(\underline{a}) = \varepsilon(\eta) \cap g(\underline{a}) \leq \eta\}$ של המקדמים האופטימליים

היא קבוצה קמורה.

(ד) $\eta_M = \inf\limits_{\varepsilon(\eta)=\varepsilon_m} \{\eta\}$, מוגדר ומתקבל במינימום.

<u>הערה:</u> למה 2 מתקיימת גם עבור פונקציות $f(\underline{a})$, $g(\underline{a})$, $h(\underline{b})$ רציפות (ולאו-דוקא

קמורות), למעט הקמורות של הקבוצה $D(\eta)$, וכאשר $||\cdot||_\epsilon$ מוחלפת בנורמות

אקויולנטיות לה כלשהן ב-$\mathcal{C}^{M_a}$, $\mathcal{C}^{M_c}$ בהתאמה.

ואכן תחת הנחות אלו בלבד תוכח למה 2.

עבור $\eta \geq \eta_M$ הרי $\varepsilon(\eta) = \varepsilon_m$ ואילו עבור $\eta_M \geq \eta \geq \eta_m$, $\infty > \varepsilon(\eta) \geq \varepsilon_m$.

במקרה (הנדיר) שבו $\eta_M = \eta_m$ האילוץ שעל $\sum_{i=1}^{N} \beta_i u_i$ לקרב את d_c, אינו משפיע

על הפתרון האופטימלי. נתעניין במקרה הכללי יותר שבו $\eta_M > \eta_m$ ונחקור אזי

את העקום $\varepsilon(\eta)$ בקטע $[\eta_m, \eta_M]$

<u>משפט 1:</u>

(א) ל-$\eta \in [\eta_m, \eta_M]$, $\varepsilon(\eta)$ היא חסומה, יורדת ממש ב-η, רציפה וקמורה.

(ב) ל-$\eta \in (\eta_m, \eta_M]$ יש ל-$\varepsilon(\eta)$ נגזרות משמאל ומימין שערכן סופי ושיסומנו על ידי

$\varepsilon'_-(\eta)$ ו-$\varepsilon'_+(\eta)$ בהתאמה.

כמוכן קיימת $\varepsilon'_+(\eta_m)$ (אם כי היא יכולה להיות $-\infty$). הפונקציה $\varepsilon'(\eta)$

מונוטונית לא יורדת ב-$[\eta_m, \eta_M]$ ואי-חיובית.

(ג) ל-$\eta \in [\eta_m, \eta_M]$ הרי לכל $\underline{a} \in D(\eta)$, $g(\underline{a}) = \eta$ ולכן:

$D(\eta) = \{\underline{a} \; ; \; f(\underline{a}) = \varepsilon(\eta) \cap g(\underline{a}) = \eta\}$

(ד) כאשר $f(\underline{a})$ קמורה-ממש אזי ל-$\eta \in [\eta_m,\eta_M]$ $\varepsilon(\eta)$ קמורה-ממש, $\varepsilon'(\eta)$
מונוטונית עולה-ממש והקבוצות $D(\eta)$ מכילות וקטור $\underline{a}^*$ יחיד.

עקום טיפוסי של $\varepsilon(\eta)$ מוצג בציור 3.7. על פי סעיף (ד) של המשפט, הלמה הבאה נותנת תנאים מספיקים ליחידות הקירוב האופטימלי.

<u>למה 3:</u>

התנאים הבאים מספיקים על מנת ש-$f(a)$ תהא קמורה-ממש.

(א) $||\cdot||_T$ מונוטונית-ממש ביחס לוקטורים ב-$[0,\infty)^N$, קרי:

$$(\forall\ \underline{\delta} \in [0,\infty)^N,\ \underline{\Delta\delta} \in [0,\infty)^N,\ \underline{\Delta\delta} \neq \underline{0})\ (||\underline{\delta}||_T < ||\underline{\delta}+\underline{\Delta\delta}||_T)$$

(ב) $\dim SD_i = (M_i+1)$, $1 \leq i \leq N$.

(ג) $||\cdot||_i$ הן נורמות קמורות-ממש על SD_i, קרי:

$$(\forall u,v \in SD_i)\ (||u+v||_i = ||u||_i + ||v||_i \implies u = |\alpha|v,\ \ \alpha \in \mathbb{C})$$

ניתן גם להציג תנאים אחרים, כמתואר בהערה הבאה:

<u>הערה:</u>

(א) תחת תנאי (א) דלעיל, ותנאים (ב) ו-(ג) רק עבור $i \in I \subset \{1,...,N\}$, $f(\underline{a})$
אינה בהכרח קמורה-ממש, אך כל הוקטורים $\underline{a}^* \in D(\eta)$ (ל-η נתון) מזדהים
במקדמים $\{a^*_{ik}\}_{k=1}^{M_i}$ עבור $i \in I$.

(ב) בפרט כשתנאי (ג) מתקיים לכל $1 \leq i \leq N$, ובנוסף: (1) $||\cdot||_T$ היא נורמה
קמורה-ממש על $[0,\infty)^N$, או ש-(2). תנאי (ב) מתקיים עבור I כך ש-$|I| = (N-1)$,
אזי לכל η, $D(\eta)$ מכילה וקטור $\underline{a}^*$ יחיד.

נגדיר בעיית קירוב שניה על ידי הגדרת מדד הקירוב : $\varepsilon_+(K) = \varepsilon + K\delta_c$ עבור
$0 \leq K < \infty$. עתה נחפש וקטור $\underline{a} \in \mathbb{C}^{M_a}$ כך שימזער את $\varepsilon_+(K)$ ל-K נתון. נסמן את:
$\varepsilon_+(K) \triangleq L(\underline{a},K)$, אזי על פי ההגדרות הקודמות: $L(\underline{a},K) = f(\underline{a}) + Kg(\underline{a})$.
המשפט הבא קושר את פתרונות בעית קרוב זו, עם פתרונות בעית הקירוב המקורית שהוצגה לעיל.

משפט 2:

(א) $\psi(K) = \inf_{\underline{a}\in \mathbb{C}^{M_a}} L(\underline{a},K)$, מוגדר ומתקבל במינימום.

תהא $\xi(K)$ קבוצת הוקטורים $\underline{a}$ הממזערים את $L(\underline{a},K)$, אזי: $\xi(K) = \bigcup_{\eta\in I_K} D(\eta)$,

כאשר $I_K \triangleq \{\eta \,; \varepsilon'_+(\eta) \geq -K \geq \varepsilon'_-(\eta)\}$ הוא אינטרוול ב-$[\eta_m,\eta_M]$

עבור $K > 0$ ו- $I_o = [\eta_M,\infty)$.

(ב) $\psi(K) = \inf_{\eta\in[\eta_m,\eta_M]} \{\varepsilon(\eta) + K\eta\}$, ועבור $K > 0$ I_K הוא האינטרוול הסגור שבו מתקבל המינימום.

(ג) כאשר $f(\underline{a})$ קמורה-ממש, I_K הוא נקודה בודדת לכל $0 < K$, ו-$\xi(K)$ מכילה וקטור מקדמים יחיד אופטימלי $\underline{a}^*(K)$.

משמעות:

$$\bigcup_{\eta\in[\eta_m,\infty)} D(\eta) \supseteq \bigcup_{K\in[0,\infty)} \xi(K) \supseteq \bigcup_{\eta\in(\eta_m,\infty)} D(\eta)$$

כששיוויון ממש מתקיים באחת ההכלות, והשיוויון מתקיים בהכלה השמאלית אמ"ם $\varepsilon'_+(\eta_m) > -\infty$.

ניתן לפרש את $(-K)$ כשיפוע הפונקציה $\varepsilon(\eta)$, כדלקמן:

עבור קטע שבו הנגזרת $\varepsilon'(\eta)$ קבועה ושווה ל-$(-K)$ הרי I_K יהא אינטרוול של ערכי η, ואילו כאשר $\varepsilon'_+(\eta_o) > \varepsilon'_-(\eta_o)$, יהא אינטרוול של ערכים $-K \in [\varepsilon'_-(\eta_o), \varepsilon'_+(\eta_o)]$ עבורם $\xi(K) = D(\eta_o)$.

ממסקנות אלו נובע שקיים עקום קשיר ומונוטוני לא-עולה $\eta(K)$ כמתואר בציור 3.8.

המשפט הבא מציג בעיה אקויולנטית לבעית הקרוב שהוצגה לעיל:

משפט 3:

בעית הקירוב המקורית שהצגנו אקויולנטית לבעיה הבאה:

$$\varepsilon(\eta) = \operatorname{Inf}_{\left\{\begin{array}{l} \|[\frac{1}{A\underline{a}}]\|^*_c \leq \eta \\ a_1=\dots=a_N=1 \end{array}\right.} \{\|\underline{a}\|^*\}$$

כאשר $A_{M_c\times(M_a+N)}$ היא מטריצה נתונה עם דרגה M_c ו-N עמודות ראשונות שהם עמודות אפסים.

$\|\cdot\|^*_c$ היא סמי-נורמה של $\mathbb{C}^{(M_c+1)}$ שהיא נורמה על $\mathbb{C}^{M_c}$ (כשהאבר הראשון בוקטור הוא אפס), ואילו $\|\cdot\|^*$ היא סמי-נורמה על $\mathbb{C}^{(M_a+N)}$, שהיא נורמה על $\mathbb{C}^{M_a}$ (כש-N האברים הראשוניים בוקטור הם אפס).

בנוסף * $||\cdot||$ ניתנת לפרוק הבא: $||\underline{a}||^* = t(\underline{\delta}(\underline{a}))$, כש-$t(\underline{\delta})$ מונוטונית ביחס לכל קומפוננטה של $\underline{\delta} \in [o,\infty)^N$ ו-$\delta_i = ||p_i\underline{a}||^*$. המטריצות $\{p_i\}_{i=1}^N$ הן מטריצות $(M_a+N) \times (M_a+N)$ שיוצרות חלוקה של $\mathbb{C}^{(M_a+N)}$ לסכום ישר של N תתי-מרחבים $\{\mathbb{C}^{(M_i+1)}\}_{i=1}^N$ הנוצרים על ידי איפוס כל אברי $\underline{a}$ פרט לאבר ה-i בין N האברים הראשונים וקבוצת M_i אברים מתוך M_a האברים האחרים.

III. מרחבי העתקות

נתרכז כעת בתת-מרחבים ליניאריים V של $\{v \; ; \; v:\Omega \to \mathbb{C}\} = U$, כש-$\Omega$ קבוצה כלשהי, ולכל $v \in V$, הצמוד הקומפלקסי $\bar{v}$ המוגדר על ידי $(\forall\, p \in \Omega) \; (v(p) = \overline{v(p)})$ שייך גם כן ל-V.

פעולות החיבור והכפל בסקלר ב-V הן הפעולות הסטנדרטיות ב-$\mathbb{C}$.

בנוסף נניח שמקדמי הקומבינציה הלינארית המגדירה את SD_c $\{\beta_i\}_{i=1}^N$ הם ממשיים.

המשפט הבא מגדיר תנאים מספיקים לכך שכל קבוצת פתרונות $D(\eta)$ (או $\xi(K)$) תכיל וקטור מקדמים ממשי $\underline{a}$.

<u>משפט 4</u>:

אם קיימות $\theta:\Omega \to \mathbb{C}$ ו-$\rho:\Omega \to \Omega$ כך שעבור $\sigma:U \to U$ המוגדרת על ידי:

$$(\forall\, v \in U,\; p \in \Omega) \quad (\sigma(v))(p) = \bar{v}(\rho(p))\,\theta(p)$$

מתקיימים:

1. $\sigma(d_c) = d_c$
2. $\sigma(d_i) = d_i \quad 1 \le i \le N$
3. $\sigma(v_{ik}) = v_{ik} \quad 1 \le i \le N,\; 1 \le k \le M_i$.
4. $||\sigma(v)||_c = ||v||_c \quad \forall\, v \in SD_c$.
5. $||\sigma(v)||_i = ||v||_i \quad \forall\, v \in SD_i,\; 1 \le i \le N$.

אזי בכל אחת מהקבוצות $D(\eta)$ $\eta \in [\eta_m,\eta_M]$ (או $\xi(K)$ $K \in [o,\infty)$) ישנו וקטור $\underline{a}$ ממשי.

<u>הערה</u>: בטרם נמשיך נראה שאכן $\sigma(SD_i) \subset SD_i$ ו-$\sigma(SD_c) \subset SD_c$, וזאת כדלקמן:

$v \in SD_i \Rightarrow v = \alpha d_i + \sum_{k=1}^{M_i} a_{ik}v_{ik} \Rightarrow \sigma(v) = \bar{\alpha}\sigma(d_i) + \sum_{k=1}^{M_i} \bar{a}_{ik}\,\sigma(v_{ik})$,

אך מ-2 ו-3 נקבל כי $\sigma(v) = \bar{\alpha}d_i + \sum_{k=1}^{M_i} \bar{a}_{ik}v_{ik} \in SD_i$.

$v \in SD_c \Rightarrow v = \alpha d_c + \sum_{k=1}^{M_c} b_k v_{ck} \Rightarrow \sigma(v) = \bar{\alpha}\sigma(d_c) + \sum_{k=1}^{M_c} \bar{b}_k\,\sigma(v_{ck})$.

ומ-1, 3 והעובדה ש- $\{v_{ck}\}_{k=1}^{M_c}$ הם אברים מתוך $\{v_{ik}\}_{k=1,i=1}^{M_i,N}$, הרי
$\sigma(v) = \bar{\alpha}d_c + \sum_{k=1}^{M_c} \bar{b}_k v_{ck} \in SD_c$.

המשפט הבא מגדיר תנאים מספיקים לכך שבכל קבוצת פתרונות $D(\eta)$ $(\xi(k))$ יהא וקטור מקדמים <u>a</u> המגדיר העתקה בעלת פזה רצויה. לשם כך נסמן: $\hat{U} = \{v \; ; \; v : \Omega \to R\}$.

<u>משפט 5:</u>

כאשר מתקיים:

1. $d_c(p) = \hat{d}_c(p)\,e^{j\psi(p)}$ $(\forall\, p \in \Omega)$ ו-$\hat{d}_c \in \hat{U}$, ψ.

2. $d_i(p) = \hat{d}_i(p)\,e^{j\psi(p)}$ $(\forall\, p \in \Omega)$ ו-$\hat{d}_i \in \hat{U}$, $1 \leq i \leq N$.

3. לכל $1 \leq i \leq N$ קיימות פרמוטציות $\pi_i(\cdot)$ של $\{1,\ldots,M_i\}$ כך ש:
$v_{ik}(p) = \bar{v}_{i\pi_i(k)}(p)\,e^{j2\psi(p)}$ לכל $1 \leq k \leq M_i$ ו-$p \in \Omega$.

4. לכל $u,v \in SD_c$, $(|u(p)| \geq |v(p)|)$ $(\forall\, p \in \Omega)$ $\Rightarrow$ $||u||_c \geq ||v||_c$.

5. לכל $u,v \in SD_i$ ולכל $1 \leq i \leq N$, $(|u(p)| \geq |v(p)|)$ $(\forall\, p \in \Omega)$ $\Rightarrow$ $||u||_i > ||v||_i$.

אזי בכל קבוצת פתרונות $D(\eta)$ $(\xi(k))$ יהא וקטור מקדמים <u>a</u> שההעתקות u_i המתאימות לו מקיימות: $u_i(p) = \hat{u}_i(p)\,e^{j\psi(p)}$ $(\forall\, p \in \Omega)$, כש-$\hat{u}_i \in \hat{U}$, לכל $1 \leq i \leq N$.

המשפט הבא מגדיר תנאים מספיקים ל-"סימטריה" מסוימת של פתרון בעית הקירוב.

<u>משפט 6:</u>

תהי $S_\Omega = \{\rho:\Omega \to \Omega \; , \; \exists\, \rho^{-1}\}$ חבורת ההעתקות ההפיכות על Ω ביחס לפעולת ההרכב. יהי $\{\omega_i\}_{i=1}^{N}$ coset שמאלי של תת-חבורה סופית מסדר N של S_Ω, קרי $\omega_i = \omega_1\theta_i$, $\{\theta_i\}_{i=1}^{N}$ תת-החבורה הסופית. נגדיר בתוך $L(U,U)$ את ה-coset הימני המתאים $\Omega_i = \Omega_i = \Theta_i\Omega_1$ על ידי ההגדרה $\Omega_i(v(p)) = v(\omega_i(p))$ $(\forall\, p \in \Omega)$ $(\forall\, v \in U)$.

אם מתקיים:

1. $\Theta_i(d_c) = d_c$ $(\forall\ 1 \le i \le N)$

2. $\Omega_i(d_o) = d_i$ $(\forall\ 1 \le i \le N)$ $(d_o \in U)$

3. $\Omega_i(S_o) = S_i$ $(\forall\ 1 \le i \le N)$ $(S_o \subset U)$

4. $\beta_i = 1$ $(\forall\ 1 \le i \le N)$.

5. $||\Theta_i(v)||_c = ||v||_c$ $(\forall\ v \in SD_c)$ $(\forall\ 1 \le i \le N)$

6. $||\Omega_i(v)||_i = ||v||_o$ $(\forall\ v \in SD_o)$, $(\forall\ 1 \le i \le N)$ $(||\cdot||_o)$

7. $||\underline{\delta}^{(1)}||_T = ||\underline{\delta}||_T$ כאשר $\underline{\delta}^{(1)}$ זהו הוקטור שנוצר על ידי סיבוב ציקלי של אברי $\underline{\delta}$ באחד.

אזי בכל קבוצה $D(\eta)$ $(\xi(k))$ יהא וקטור מקדמים $\underline{a}$ שההעתקות u_i המתאימות לו מקיימות $u_i = \Omega_i(u_o)$, $1 \le i \le N$, כש-$u_o \in S_o$.

בנוסף עבור כל וקטור מקדמים $\underline{a}$ בעל תכונה זו, הרי: $\delta_i = \delta_o$, $1 \le i \le N$, ולכן

$$\varepsilon(\eta) = \delta_o \left|\left| \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right|\right|_T, \quad \text{ו-} \quad u_c = \sum_{i=1}^{N} \Omega_i(u_o).$$

לכן הבעיה המקורית זהה (עד כדי קבוע) לבעית הקירוב הבאה: $\min \{||d_o-u_o||_o\}$ $||d_c - \sum_{i=1}^{N} \Omega_i(u_o)||_c = \eta$

ולכן מעורב בה רק וקטור מקדמים ממימד M_a/N.

IV. דוגמה - $L_\infty(T)$

יהא $\Omega = T = \{z \mid |z| = 1\}$, מעגל היחידה ב-$\mathbb{C}$, ו-$V \subset U$ אוסף הפונקציות המדידות והחסומות ב-T, קרי $V = L_\infty(T)$, כאשר נזהה פונקציה ב-$L_\infty(T)$ עם פונקציה מחזורית ב- $L_\infty[-\pi,\pi]$.

נבהיר כעת את משפטים 4 ו-6 במקרה זה.

משפט 4: $\theta = 1$ ו- $\rho(z) = \bar{z}$, אזי: $\sigma(v)\,(e^{j\theta}) = \bar{v}(e^{-j\theta})$, זו העתקה הפיכה של $L_\infty(T)$ על עצמו.

בנוסף נניח שב-SD_i וב-SD_c מוגדר ומתכנס בהחלט טור פורייה הניתן על ידי סדרת המקדמים:

$$\hat{v}_n = F(v) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j\theta n}\, v(e^{j\theta})d\theta$$

ואזי $v = \sigma(v) \iff \hat{v}_n = \widehat{\sigma(v)}_n = \bar{\hat{v}}_n \iff I_m(\hat{v}_n) = o$.

לכן משמעות תנאים 1 - 3 היא שמקדמי טור-פורייה של $\{v_{ik}, d_i, d_c\}$ הם כולם ממשיים.

משמעות תנאים 4 - 5 היא שהסמי-נורמות $||\cdot||_i$, $||\cdot||_c$ ניתנות לזהוי עם סמי-נורמות על מקדמי טורי-פורייה של הפונקציות השונות המקיימות כולן: $||\hat{v}_n|| = ||\bar{\hat{v}}_n||$. לפיכך -

משפט *4:

כאשר טור-פורייה מוגדר ומתכנס בהחלט ב-SD_i וב-SD_c, מקדמי הטורים של הפונקציות v_{ik}, d_i,d_c כולם ממשיים, והסמי-נורמות $||\cdot||_i$, $||\cdot||_c$ ניתנות לזהוי עם סמי-נורמות על מקדמי טור-פורייה, שמקיימות כולן $||\hat{v}_n|| = ||\bar{\hat{v}}_n||$, אזי בכל $D(\eta)$ $(\xi(K))$ קיים וקטור $\underline{a}$ ממשי.

משפט 6:

$\omega_i(z) = z\, e^{j\frac{2\pi i}{N}} e^{j\Delta}$ זהו coset של תת-החבורה הציקלית מסדר N ב-S_T הנוצרת על ידי פונקצית הזזת הזוית ב- $\frac{2\pi}{N}$.

יהיו $\{S_i\}_{i=1}^{N}$ תת-מרחבים זהים זה לזה, של פולינומים טריגונומטריים מאורך $M_i = M$ קרי $v_{ik} = z^{(k-1)}$, $1 \le k \le M$. בברור:

$$\Omega_i(v_{ik})\;(z) = z^{(k-1)}\, e^{j\frac{2\pi(k-1)i}{N}}\, e^{j\Delta}$$

ולכן $\Omega_i(S_o) = S_i$, כש-S_o הוא גם כן תת-מרחב הפולינומים הנ"ל.

האופרטור Θ_i מזיז את אברי U ב- $\frac{2\pi i}{N}$ על פני מעגל היחידה T ולכן המשפט במקרה זה:

משפט *6:

תהי d_c אינווריאנטית להזזה ב- $\frac{2\pi}{N}$ על-פני T, ואילו d_i נוצרות מתוך d_o כלשהו על ידי הזזות ב-$(\Delta + \frac{2\pi i}{N})$. הסמי-נורמה $||\cdot||_c$ אינווריאנטית להזזות ב- $\frac{2\pi}{N}$ על-פני T, ואילו $||\cdot||_i$ נוצרות מ- $||\cdot||_o$ על ידי ההזזות ב-$(\Delta + \frac{2\pi i}{N})$.

אזי קיים סט-פולינומים אופטימלי $(\underline{a} \in D(\eta))$, כך שהפולינום ה-$i$-י נוצר על ידי הזזה ב-$(\Delta + \frac{2\pi i}{N})$ של פולינום אב-טיפוס על פני T.

זהו בדיוק המשפט שמוכח בנספח ג' של העבודה, ומצוטט בסעיף 4.1.

נבהיר כעת את התנאים לקיום למה 3, במקרה זה (יחידות הפתרון).

תנאי (א) של הלמה ($||\cdot||_T$ מונוטונית-ממש) מתקיים למשל לנורמות הולדר עם אינדקס $1 \le s < \infty$ וקבועי משקל חיוביים ממש כלשהם, הוא איננו מתקיים לנורמת המקסימום $(s = \infty)$, ואזי בבירור יתכנו מספר רב של פתרונות לבעית הקירוב.

תנאי (ב) של הלמה $(d_i \notin S_i\ ,1 \le i \le N)$ מתקיים למקרה המקובל שבו S_i מכיל פונקציות רציפות ואילו ל-d_i יש נקודות אי-רציפות על T, או ש-S_i מכיל פולינומים בעלי מספר סופי של נקודות התאפסות ואילו d_i מתאפסת בקבוצה בעלת מידה חיובית (על פי $||\cdot||_i$) על T.

תנאי (ג) של הלמה ($||\cdot||_i$ קמורות-ממש), מתקיים עבור הסמי-נורמות WL_q $1 < q < \infty$ (כש-WL_q מציין שהן מכילות פונקצית משקל $W(z)$ על מעגל היחידה), ובלבד שאלו הן נורמות על SD_i (קרי שפונקצית המשקל מבטיחה שהנורמה של כל אבר ב-SD_i חיובית).

מאידך על WL_∞ ו-WL_1 תנאי זה לא מתקיים ואכן יחידות הפתרון תלויה בסוג הפונקציות המקרבות (קרי, בתת-המרחבים S_i), כמו במקרה של בעית קירוב רגילה (ללא אילוצים על $\sum_{i=1}^{N} \beta_i u_i$).

V. הרחבת התוצאות עבור מספר אילוצים

נתאר להלן את הרחבת התוצאות שתוארו קודם לכן למקרה של מספר אילוצים על קומבינציות לינאריות של הוקטורים u_i.

הבעיה הכללית יותר תוגדר בדומה להגדרות של סעיף I לעיל כשמגדירים $P > 1$ זוגות של תתי-מרחבים לינאריים מתאימים של V $\{S_{cj}\}_{j=1}^{P}$ ו- $\{SD_{cj}\}_{j=1}^{P}$ על ידי שימוש בקומבינציות לינאריות שונות $\{\beta_{ij}\}_{j=1,i=1}^{P,N}$ ובאברים שיש לקרב על ידיהן $\{d_{cj}\}_{j=1}^{P}$.

עתה יוגדרו P פונקציות $\{g_j(\underline{a})\}_{j=1}^{P}$ (ולכן בדומה, גם $h_j(\underline{b})$, כש-$\underline{b} = A_j\underline{a}$

ו-A_j אלו P מטריצות שונות מדרגות $\{M_{cj}\}_{j=1}^{P}$ ומימדים $M_{cj} \times M_a$).

<u>משפט 3</u> מוכלל באופן מיידי כדלקמן:

השיכון של בעית הקירוב יהא ב-$\mathbb{C}^{M_a+N+P}$, וזאת על ידי החלפת $|||\cdot|||_c^*$, ב-P

סמי-נורמות מתאימות $\{|||\cdot|||_{cj}^*\}_{j=1}^{P}$ על $\mathbb{C}^{M_{cj}+1}$ שהן נורמות ב-$\mathbb{C}^{M_{cj}}$ המתאים,

כך ש-

$$\delta_{cj} \triangleq g_j(\underline{a}) = |||[\frac{1}{-A_j\underline{a}}]|||_{cj}^* .$$

<u>ההוכחה</u> היא על ידי חזרה (מייגעת) על ההוכחה של המשפט שמובאת בסעיף VI,

לגבי $P = 1$.

לפיכך בעית התכנות הקמור היסודית המקושרת עם בעית הקירוב שלנו היא:

$$(CP) \quad \begin{cases} \varepsilon(\underline{\eta}) = \text{Inf} \quad f(\underline{a}) \\ g_j(\underline{a}) \le \eta_j, \quad 1 \le j \le p \end{cases}$$

<u>למה 1</u> מוכללת באופן מיידי על ידי הפעלתה לגבי כל הפונקציות $h_j(\underline{b})$ עבור

$1 \le j \le P$, והוכחתה נותרת ללא שינוי ממשי.

<u>למה 3</u> נוגעת רק בפונקציה $f(\underline{a})$ ולכן אינה מושפעת כלל, מההכללה למספר אילוצים.

<u>משפט 4</u> <u>ומשפט 5</u> מוכללים מיידית על ידי שכפול הדרישות שבמשפטים הנ"ל לגבי

SD_c ו- $||\cdot||_c$ לכל תתי המרחבים SD_{cj} ולכל הנורמות $||\cdot||_{cj}$, $1 \le j \le P$.

<u>ההוכחה</u> היא חזרה (מייגעת) על ההוכחות שמובאות בסעיף VI לגבי $P = 1$, כשמשתמשים

בתכונה שלכל וקטור $\underline{\eta}$ פיסיבילי, הקבוצה $D(\underline{\eta})$ היא קבוצה קמורה (מה שנובע

מההכללה של למה 2 כפי שנראה בהמשך).

<u>משפט 6</u> ניתן להכללה בדומה למשפטים 4 ו-5, אולם התנאי $\beta_{ij} = 1$ לכל i ולכל j

שיופיע במשפט ה"מוכלל", יגרום לו להיות במידה רבה נטול ערך, שכן על מנת שתנאי

הסימטריה יתקיים, כל האילוצים חייבים להיות על $\sum_{i=1}^{N} u_i$ <u>בלבד</u>.

<u>משפט 2</u> עוסק למעשה בבעיה הדואלית של (CP) שהוצגה לעיל. לפיכך הוא יהא קשור בפונקציה $L(\underline{a},\underline{K}) = f(\underline{a}) + \sum_{j=1}^{P} K_j g_j(\underline{a})$ וב- $\psi(\underline{K}) = \inf_{\underline{a}\in \not{C}^{M_a}} L(\underline{a},\underline{K})$.

חלק מהתוצאות שמצוטטות בסעיף II למקרה של $P = 1$ עוברות הלאה ל-$P > 1$, אך העדפנו במסגרת עבודה זו לא להתעמק בנושא זה (אם כי הוא איננו קשה במיוחד, ומרבית התוצאות נובעות מהאנליזה של הבעיה (CP) ב-[56,62]).

הרחבת למה 2 ומשפט 1 מעניינת למדי עבור $P > 1$. במקרה זה התחום הפיזיבילי (ערכי $\underline{\eta}$ שעבורם למערכת האי-שיוויונות $g_j(\underline{a}) \geq \eta_j$, $1 \leq j \leq P$, יש פתרון) שיסומן ב- $\Lambda \subset [0,\infty)^P$, הוא בעל צורה עשירה יותר (ביחס למקרה של $P = 1$ שהוצג בסעיף II בנספח זה), ולכן מופיעות תוצאות מעניינות. אם זאת ההוכחות הן במרביתן חזרה (מייגעת) אחר ההוכחות של המקרה הפרטי $P = 1$ המוצגות בסעיף הבא במפורט ולכן נוותר עליהן, למעט איזכור של מספר מקומות בעייתיים.

<u>למה 2</u>

(א) עובדת קיומו של ε_m (מינימום לא מאולץ) מתקבלת בדיוק כמו במקרה של $P = 1$.

(ב) באותו אופן קיימים $\{\eta_{m_j}\}_{j=1}^{P}$ (מינימומים של ערכי האילוצים השונים).

(ג) לכל $\underline{\eta} \in \Lambda$ קיים המינימום המאולץ של $f(\underline{a})$ שערכו $\varepsilon(\underline{\eta})$, והקבוצה של ה-$\underline{a}$-ים שמגשימה אותו $D(\underline{\eta})$ היא קבוצה קמורה. ההוכחה עוקבת אחרי זו של $P = 1$ (כי P סופי). ההבדל העיקרי ביחס למקרה הפרטי של $P = 1$ הוא בתכונות המבניות של Λ שיתוארו בהמשך.

(ד) בדומה למקרה של $P = 1$, ניתן להגדיר את $\{\eta_{M_j}\}_{j=1}^{P}$ שמתקבלים במינימום, וכמובן מקיימים ש- $\eta_{M_j} \geq \eta_{m_j}$. קיים תחום "סופר-פיסיבילי" $\tilde{\Lambda} \supset \prod_{j=1}^{P} [\eta_{M_j},\infty)$ שבו $\varepsilon(\underline{\eta}) = \varepsilon_m$ אך בדומה לתחום Λ , עבור $P > 1$ זה כבר לא יהא אינטרוול חצי-סגור, כי אם קבוצה מורכבת יותר.

<u>למה 4</u>. תכונות התחומים Λ ו-$\tilde{\Lambda}$.

(א) Λ מוכל באורתנט $\prod_{j=1}^{P} [\eta_{m_j},\infty)$, וההיטל שלו על הקואורדינטה ה-j-ית הוא בדיוק $[\eta_{m_j},\infty)$. Λ זו קבוצה קמורה וסגורה.

(ב) נגדיר סדר חלקי על $[0,\infty)^P$ על ידי : $\underline{\eta}_1 < \underline{\eta}_2$ אם"ם $\underline{\eta}_2 - \underline{\eta}_1 \in [0,\infty)^P$, אזי לכל $\underline{\eta}_1 \in \Lambda$ קיים ש- $\{\eta_2;\ \underline{\eta}_2 > \underline{\eta}_1\} \subset \Lambda$.

(ג) קיימים !P וקטורים $\hat{\underline{\eta}}^{(\pi)}$ על שפת Λ (כש-$1 \le \pi \le P!$ זו פרמוטציה של $\{1,\dots,P\}$) המוגדרים על ידי: $\hat{\eta}^{(\pi)}_{\pi(1)} = \eta_{m_{\pi(1)}}$, $\hat{\eta}^{(\pi)}_{\pi(2)} = \inf\limits_{g_{\pi(1)}(\underline{a}) \le \hat{\eta}^{(\pi)}_{\pi(1)}} g_{\pi(2)}(\underline{a})$

וכו'.

(ד) $\tilde{\Lambda}$ מוכל באורתנט $\prod_{j=1}^{P} [\eta_{M_j},\infty)$, וההיטל שלו על הקואורדינטה ה-j-ית הוא בדיוק $[\eta_{M_j},\infty)$, כמוכן הוא מוכל ב-Λ, ו-$\tilde{\Lambda}$ זו קבוצה קמורה וסגורה.

(ה) לכל $\underline{\eta}_1 \in \tilde{\Lambda}$ קיים ש- $\{\underline{\eta}_2 \,;\, \underline{\eta}_2 > \underline{\eta}_1\} \subset \Lambda$.

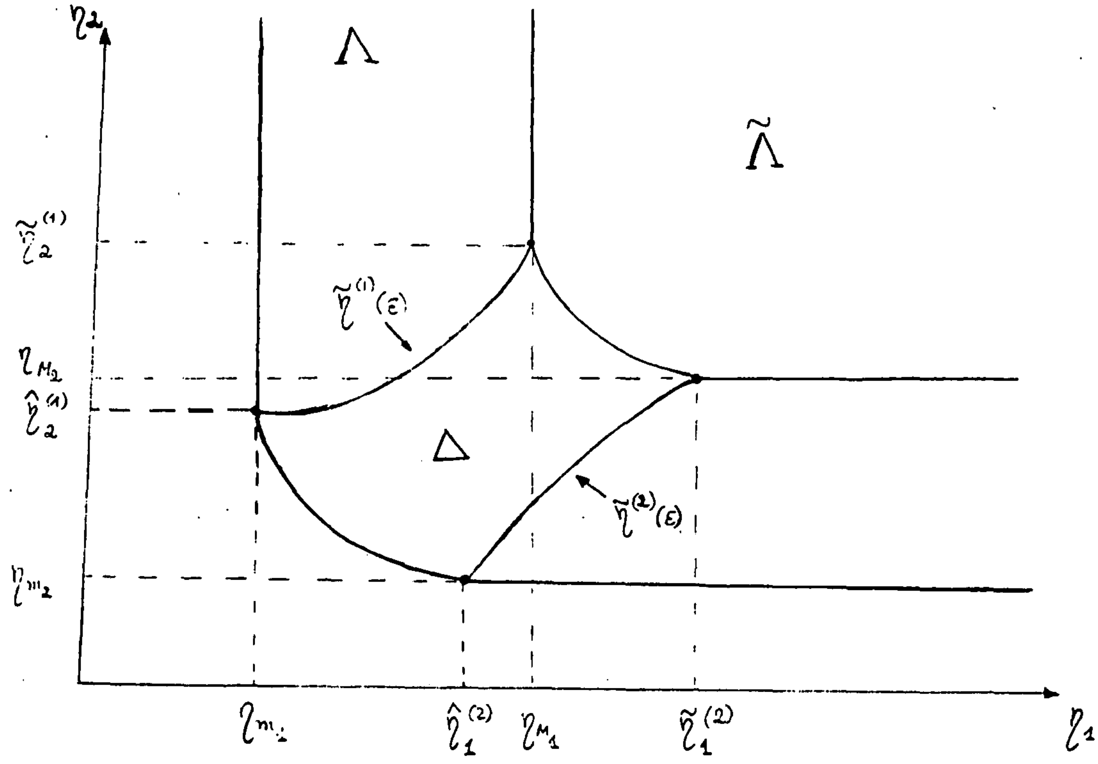
(ו) קיימים !P וקטורים $\tilde{\underline{\eta}}^{(\pi)}$ על שפת $\tilde{\Lambda}$ המוגדרים בדומה לוקטורים $\hat{\underline{\eta}}^{(\pi)}$ על ידי: $\tilde{\eta}^{(\pi)}_{\pi(1)} = \tilde{\eta}_{M_{\pi(1)}}$ $\tilde{\eta}^{(\pi)}_{\pi(2)} = \inf\limits_{\left\{\begin{array}{l} g_{\pi(1)}(\underline{a}) \le \tilde{\eta}^{(\pi)}_{\pi(1)} \\ f(\underline{a}) = \varepsilon_m \end{array}\right\}} g_{\pi(2)}(\underline{a})$

וכו'.

הוכחה:

Λ ($\tilde{\Lambda}$) היא אוסף ה-$\underline{\eta}$ שעבורם למערכת אי-השיוויונות בפונקציות קמורות $g_j(\underline{a}) \le \eta_j$ $1 \le j \le P$ ($g_j(\underline{a}) \le \eta_j$ $1 \le j \le P$, ו- $f(\underline{a}) \le \varepsilon_m$) יש פתרון. מאחר ואלו כולן פונקציות אי-שליליות המקיימות את למות 1 ו-2, הרי התכונות המצוטטות לעיל נובעות מיידית ממשפטים שמופיעים ב-[62].

ציור B-1 מתאר תצורה טיפוסית של Λ ו-$\tilde{\Lambda}$ עבור P = 2.



ציור B-1: תאור של קבוצות Λ ו- $\tilde{\Lambda}$ טיפוסיות.

Fig. B-1: Typical Sets Λ and $\tilde{\Lambda}$.

משפט 1:

נדון במקרה שבו $\tilde{\Lambda} - \Lambda \neq \phi$.

(א) $\varepsilon(\underline{\eta})$ קמורה, רציפה, ומונוטונית לא-עולה ב-$\underline{\eta}$ (ביחס לסדר החלקי שהגדרנו בלמה 4). בקבוצה Λ היא חסומה מלמטה על ידי ε_m ומלמעלה על ידי

$$\varepsilon_M = \sup_{\underline{\eta}\in\Lambda} \varepsilon(\underline{\eta}) < \infty .$$

(ב) ל-$\varepsilon(\underline{\eta})$ יש נגזרות כווניות ב-Λ שהן סופיות לפחות בכל הפנים של Λ . הנגזרות הכווניות הן אי-חיוביות ביחס לכיוונים ב-$[0,\infty)^P$. והפונקציה $D^+\varepsilon(\underline{\eta};\underline{Z})$ (D^-) היא מונוטונית לא יורדת (במובן שניתן ב-4.24 ב-[56]).

(ג) לכל $\varepsilon_m \leq \varepsilon_0 \leq \varepsilon_M$, התחום $\Lambda_{\varepsilon_0} = \{\underline{\eta}\ ;\varepsilon(\underline{\eta}) \leq \varepsilon_0\}$, הוא מהצורה של התחומים Λ ו-$\tilde{\Lambda}$ (כפי שתוארה בלמה 4). נגדיר את P! הוקטורים $\tilde{\underline{\eta}}^{(\pi)}(\varepsilon_0)$, כפי שהוגדרו בלמה 4, אזי $\tilde{\underline{\eta}}^{(\pi)}(\varepsilon)$ הוא עקום רציף ב- ε.

(ד) נסמן ב-Δ את תת-התחום של $\tilde{\Lambda}-\Lambda$ המוכל בין P! העקומים $\tilde{\underline{\eta}}^{(\pi)}(\varepsilon)$, אזי בתוך Δ , $\varepsilon(\underline{\eta})$ יורדת-ממש ביחס ל-$\underline{\eta}$, והקבוצות $D(\underline{\eta})$ הן מהצורה:

$$D(\underline{\eta}) \triangleq \{\ \underline{a}\ \ ;\ f(\underline{a}) = \varepsilon(\underline{\eta})\ \cap\ g_j(\underline{a}) = \eta_j\ \ ,1 \leq j \leq P\} .$$

(ה) כאשר $f(\underline{a})$ קמורה ממש, אזי $\varepsilon(\underline{\eta})$ קמורה ממש ב-Δ , $\underline{\eta}\in\Delta$, $D^+\varepsilon(\underline{\eta}\ ;\ \underline{Z})$ (D^-) היא מונוטונית עולה ממש ב-Δ $\underline{\eta}\in$, וקיים $\underline{a}^*$ יחיד בכל קבוצה $D(\underline{\eta})$ עבור $\underline{\eta}\in\Lambda$.

הוכחה:

זו ההרחבה של ההוכחה של משפט 1 במקרה של $P = 1$, בסיוע תוצאות מסוימות מ-[56] (בעיקר ביחס לפונקציה $D^+\varepsilon(\underline{\eta};\underline{Z})$ (D^-)). התחום Δ הוא האנלוג של הקטע הסגור $[\eta_m,\eta_M]$ עבור $P = 1$, והוא מצויין על גבי ציור B-1 עבור המקרה הטיפוסי של $P = 2$.

VI. הוכחות המשפטים

מבוא

למה 1 ולמה 2 מבוססות על מספר משפטים באנליזה של פונקציות ממשיות, הלקוחים מתוך [63].

בלמה 1 נעשה שימוש אינטנסיבי בתכונות של הסמי-נורמות השונות על מנת להבטיח שהפונקציות $h(\underline{b})$, $g(\underline{a})$, $f(\underline{a})$ כולן קמורות וכן שימוש בעובדה ש-SD_i הם תת-מרחבים לינאריים ע"מ להראות שכאשר חורגים מכדור מספיק גדול כל הפונקציות גדולות כרצוננו. למה 2 נגזרת כמקרה פרטי של מספר משפטים באנליזה של פונקציות ממשיות.

מאחר שהפונקציות $f(\underline{a})$, $g(\underline{a})$ קמורות הרי שבעית הקירוב ניתנת להצגה כבעית תכנות קמור. משפטים 1 ו-2 מוכחים בהתבסס על מספר משפטים בתכנות קמור הלקוחים מתוך [56, 62]. בהוכחת משפט 1 מנצלים רק את התכונות של פונקציות קמורות ומרבית ההוכחה היא self-contained. בהוכחת משפט 2 בחרנו להציג את בעית הקרוב כבעית תכנות קמור פרימלית (primal) והראינו שהבעיה האקויולנטית שהגדרנו במשפט 2 קשורה ללגרנז'יאן של הבעיה הפרימלית (ולמעשה גם קשורה לבעית התכנות הדואלית). לפיכך הקדמנו להוכחת משפט 2 מבוא המאפשר להשתמש בתוצאות של [56,62].

בלמה 3 חקרנו את תכונות הסמי-נורמות השונות המספיקות על-מנת להבטיח יחידות של וקטור המקדמים האופטימלי. אלו בודאי אינם תנאים הכרחיים, אך כשהם לא מתקיימים יש צורך להתייחס ספציפית למבנה של המרחב הוקטורי V (קרי להבדיל בין מרחבי פונקציות שונים למרחבי סדרות וכדומה), וכפועל יוצא שיטות ההוכחה מסתבכות.

משפט 3 הוא משפט אלגברי, שבו משכנים את הבעיה האבסטרקטית במרחב $\mathbb{C}^{M_a+N}$, אם כי כמובן הנורמה המתקבלת שם בדרך כלל אינה נורמה טבעית.

משפטים 4 - 6 מגדירים תנאים מספיקים על ההעתקות השונות כך שבקבוצה $D(\eta)$ $(\xi(K))$ יהיו וקטורים $\underline{a}$ בעלי תכונות רצויות המוגדרות שם. השיטות הן מעיקרן אלגבריות וישנו רק ניצול של הקמירות של $D(\eta)$ הנובעת מלמה 2.

בטרם נציג את ההוכחות המפורטות להלן רשימת המשפטים שאנו משתמשים בהם במהלך ההוכחות. מאחר ונעשה שימוש רק בטופולוגיה של המרחב $\mathbb{C}^{M_a}$ הרשינו לעצמנו להסב בחופשיות את כל המשפטים מהמרחב R^n למרחב $\mathbb{C}^n$.

<u>משפטים שעליהם מבוססות ההוכחות:</u>

T.1 משפט 2.41 [63] (היינה-בורל): כל קבוצה סגורה וחסומה ב-R^n $(\mathbb{C}^n)$ היא קומפקטית.

T.2 משפט 3.6 [63] (בולצנו ויירשטרס): לכל סדרה המוכלת בקבוצה קומפקטית של תת-סדרה המתכנסת לגבול בקבוצה זו.

T.4 משפט 4.16 [63] : לפונקציה רציפה בקבוצה קומפקטית יש מינימום בקבוצה זו.

T.7 משפט 2.36 [63] : לכל אוסף קבוצות סגורות בעל תכונת החיתוך הסופי, המוכלות בקבוצה קומפקטית, חיתוך כל הקבוצות באוסף איננו ריק. (אוסף קבוצות הוא בעל תכונת החיתוך הסופי אם כל חיתוך של מספר סופי של קבוצות מתוכו איננו ריק).

T.3 משפט 4.18 [56] : פונקציה קמורה על R^n $(\mathbb{C}^n)$ היא רציפה בכל מקום.

T.13 משפט 4.35 [56] : אם לפונקציה קמורה ממש מושג המינימום הגלובלי בקבוצה קמורה, אזי הוא מושג בנקודה יחידה בקבוצה זו.

T.6 מסקנה 2.35 [63] : חיתוך של קבוצה קומפקטית וקבוצה סגורה הוא קבוצה קומפקטית.

T.5 מסקנה 4.8 [63] : עבור פונקציה רציפה ב-R^n $(\mathbb{C}^n)$ הקבוצות הבאות:
$\{x;f(x) \le \alpha\}$, $\{x;f(x) \ge \alpha\}$, $\{x;f(x) = \alpha\}$ הן סגורות לכל $\alpha \in R$.

T.8 משפט 4.17 [56] : פונקציה קמורה וחסומה המוגדרת באינטרוול סגור של R היא רציפה באינטרול הפתוח.

T.9 משפט 4.29 [63] : לפונקציה חסומה ומונוטונית לא-עולה באינטרוול סגור יש רק נקודות אי-רציפות מסדר ראשון בו.

T.12 משפט 4.33 [56] : קבוצות הרמה ב-R^n $(\mathbb{C}^n)$ (המוגדרות על ידי $\{x\ ;f(x) \le \alpha\}$) של פונקציה קמורה $f(x)$ הן קמורות.

T.10 משפט 4.25 [56] : לפונקציה $f(x)$ קמורה המוגדרת ב-$[a,b] \subset R$ יש נגזרות מימין ומשמאל בכל $x \in [a,b]$. נסמן את הנגזרת משמאל ב-$D^-f(x)$ והנגזרת מימין $D^+f(x)$, אזי לכל $x_2 > x_1$ קיים $D^+f(x_2) \ge D^-f(x_2) \ge D^+f(x_1) \ge D^-f(x_1)$ (כאשר: $D^+f(b) \triangleq \infty$, $D^-f(a) \triangleq -\infty$).

T.11 משפט 5.8 [56] : תהא $\phi(x,\eta)$ קמורה כראוי על $¢^{n} \times R^{k}$, אזי הפונקציה $\Phi(\eta) = \inf_{x} \phi(x,\eta)$ היא קמורה בכל R^{k}.

T.14 משפט 4.34 [56] : קבוצת הנקודות שבהן פונקציה קמורה משיגה את המינימום שלה בקבוצה קמורה, היא בעצמה קבוצה קמורה.

T.15 משפט 4.22 [56] : התת-דיפרנציאל של f קמורה ב-R^{n} בנקודה x שבה f(x) סופית היא אוסף הוקטורים $\eta \in R^{n}$ שעבורם $D^{+}f(x;z) \geq \eta^{T}z$ לכל $z \in R^{n}$, כש- $D^{+}f(x;z)$ היא הנגזרת מימין של f בנקודה x , בכוון z . נסמן אותו ב-$\partial f(x)$.

T.16 משפט 5.27 [56] : תהא ϕ פונקציה קמורה כראוי וסגורה על R^{n+k} $(¢^{M_a} \times R)$ אזי $(\underline{x}^{*},\lambda^{*})$ נקודת אוכף של הלגרנז'יאן של p_{ϕ} אם"ם $\underline{x}^{*}$ הוא פתרון אופטימלי של p_{ϕ} ו-λ^{*} כופל לגרנז' של p_{ϕ}.

T.17 משפט 5.20 [56] : תהא p_{ϕ} בעלת פתרון אופטימלי, אזי λ^{*} הוא כופל לגרנז' של p_{ϕ} אם"ם $\lambda^{*} \in \partial\Phi(o)$.

<u>הוכחת למה 1:</u>

(א) על פי הגדרתן $\varepsilon \geq 0$, δ_c (כי אלו סמי-נורמות) ולכן כל הפונקציות הנדונות הן אי-שליליות. בנוסף נתבונן ב-$\underline{a}^{(\lambda)} = \lambda\underline{a}^{(1)} + (1-\lambda)\underline{a}^{(2)}$, כאשר $\lambda \in [0,1]$.

יהיו $\underline{u}^{(\lambda)}$, $\underline{u}^{(2)}$, $\underline{u}^{(1)}$ הוקטורים ב- $S_N \times \ldots \times S_1$ המתאימים ל- $\underline{a}^{(\lambda)}$, $\underline{a}^{(2)}$, $\underline{a}^{(1)}$ בהתאמה.

אזי:

$$||d_i-u_i^{(\lambda)}||_i = ||d_i-\lambda u_i^{(1)}-(1-\lambda)u_i^{(2)}||_i \leq \lambda||d_i-u_i^{(1)}||_i + (1-\lambda)||d_i-u_i^{(2)}||_i$$

מקמירות הסמי-נורמה $||\cdot||_i$ על SD_i. לפיכך יהיו $\underline{\delta}^{(\lambda)}$, $\underline{\delta}^{(2)}$, $\underline{\delta}^{(1)}$ וקטורי שגיאות הקרוב ב-$[0,\infty)^N$ אזי: $\lambda\underline{\delta}^{(1)} + (1-\lambda)\underline{\delta}^{(2)} = \underline{\delta}^{(\lambda)} + \underline{\Delta}$ כש-$\underline{\Delta} \in [0,\infty)^N$. מהמונוטוניות של $||\cdot||_T$ והיותה "נורמה" ב-$[0,\infty)^N$ נובע ש:

$$f(\underline{a}^{(\lambda)}) = ||\underline{\delta}^{(\lambda)}||_T \leq \lambda||\underline{\delta}^{(1)}||_T + (1-\lambda)||\underline{\delta}^{(2)}||_T = \lambda f(\underline{a}^{(1)})+(1-\lambda)f(\underline{a}^{(2)})$$

מכאן ש-$f(\underline{a})$ קמורה.

באופן דומה מתכונות הסמי-נורמה $||\cdot||_c$ על SD_c נקבל:

$$g(\underline{a}^{(\lambda)})=||d_c-\sum_{i=1}^{N}\beta_i u_i^{(\lambda)}||_c=||\lambda d_c-\lambda\sum_{i=1}^{N}\beta_i u_i^{(1)}+(1-\lambda)d_c-(1-\lambda)\sum_{i=1}^{N}\beta_i u_i^{(2)}||_c \leq$$
$$\leq \lambda g(\underline{a}^{(1)}) + (1-\lambda)g(\underline{a}^{(2)})$$

ולכן $g(\underline{a})$ קמורה ומאחר והעתקה ליניארית $\underline{b} = A\underline{a}$ שומרת על קמירות גם $h(\underline{b})$ קמורה.

(ב) מתכונות הסמי-נורמה $||\cdot||_i$ על SD_i: $\delta_i = ||d_i-u_i||_i \geq \left|\, ||d_i||_i-||u_i||_i \right|$.

מתכונות "הנורמה" $||\cdot||_T$ על $[0,\infty)^N$:

$$||\; ||u_i||_i ||_T \leq ||(||d_i||_i+|\; ||d_i||_i-||u_i||_i|)||_T \leq ||\; ||d_i||_i||_T +$$
$$+ ||\; |(||d_i||_i - ||u_i||_i)|||_T$$

כי

$$||u_i||_i \leq ||d_i||_i + |(||d_i||_i-||u_i||_i)|$$

ולכן התוצאה נובעת ממונוטוניות $||\cdot||_T$.

ושוב ממונוטוניות $||\cdot||_T$ נקבל: $||\; |(||d_i||_i - ||u_i||_i)|||_T \leq f(\underline{a})$

לכן לסיכום: $f(\underline{a}) \geq \| \|u_i\|_i \|_T - f(\underline{o})$

נסמן: $\hat{f}(\underline{a}) = \| \|u_i\|_i \|_T$ אזי: $f(\underline{a}) \geq \hat{f}(\underline{a}) - f(\underline{o})$

לכן (מאחר ו- $0 < f(\underline{o}) < \infty$)

מספיק על מנת להוכיח את תכונה (ב) להוכיח כי:

$$(A1) \quad (\forall \tilde{\alpha} > 0)\ (\exists M < \infty)\ (\forall \underline{a} \in \mathbb{C}^{M_a},\ \|\underline{a}\|_\epsilon > M)\ (\hat{f}(\underline{a}) > \tilde{\alpha} = \alpha + f(\underline{o}))$$

נניח את שלילת (A1) ונקבל סתירה. שלילת (A1) היא:

$$(\exists \tilde{\alpha} > 0)\ (\forall M)\ (\exists \underline{a}_M \in \mathbb{C}^{M_a})\ (\|\underline{a}_M\|_\epsilon > M \cap \hat{f}(\underline{a}_M) \leq \tilde{\alpha})$$

נבחר סדרה $M_r \uparrow \infty$ ואזי סדרת ה- $\underline{a}_{M_r}$ המתאימה מקיימת $\lim_{r\to\infty} \|a_{M_r}\|_\epsilon = \infty$ (וללא הגבלת כלליות $\|\underline{a}_{M_r}\|_\epsilon > 0$), בעוד ש- $\hat{f}(\underline{a}_{M_r}) \leq \tilde{\alpha}$.

יהא $\underline{b}_{M_r} = \underline{a}_{M_r}/\|\underline{a}_{M_r}\|_\epsilon$, אזי מתכונות הנורמה האאוקלידית $\|\underline{b}_{M_r}\|_\epsilon = 1$ ומתכונות הסמי-נורמות $\|\cdot\|_i$, $\|\cdot\|$ ו-$\|\cdot\|_T$ נובע ש-$\hat{f}(\underline{a})$ הומוגנית חיובית ולכן:

$\hat{f}(\underline{b}_{M_r}) = \hat{f}(\underline{a}_{M_r})/\|\underline{a}_{M_r}\|_\epsilon$ לכן: $0 \leq \lim_{r\to\infty} f(\underline{b}_{M_r}) \leq \lim_{r\to\infty} (\tilde{\alpha}/\|\underline{a}_{M_r}\|_\epsilon) = 0$

<u>מסקנה:</u>

מצאנו סדרה $\underline{b}_{M_r} \in \mathbb{C}^{M_a}$ של וקטורים בעלי $\|\underline{b}_{M_r}\|_\epsilon = 1$, ו-$\lim_{r\to\infty} \hat{f}(\underline{b}_{M_r}) = 0$.

אך הקבוצה $\Omega = \{\underline{b} \in \mathbb{C}^{M_a}\ ; \|\underline{b}\|_\epsilon = 1\}$ היא קבוצה קומפקטית ב-$\mathbb{C}^{M_a}$, (על פי T1), ולכן לכל סדרת אברים ב-Ω יש תת-סדרה מתכנסת ב-Ω (על פי T2).

נתבונן על תת-סדרה כזו של הסדרה $\{\underline{b}_{M_r}\}_{r=1}^{\infty} \in \Omega$ שנסמן על ידי $\{\underline{b}_{M_{r_k}}\}_{k=1}^{\infty}$.

אזי $\lim_{k\to\infty} \underline{b}_{M_{r_k}} = \underline{b} \in \Omega$ ובבירור גם $\lim_{k\to\infty} \hat{f}(\underline{b}_{M_{r_k}}) = 0$ כנובע מהמסקנה דלעיל.

בנוסף $\hat{f}(\underline{a})$ הוא למעשה הפונקציה $f(\underline{a})$ עבור $d_i = 0$.

מאחר ובמקרה זה SD_i הופך ל-S_i, כל הנורמות הנדונות מוגדרות היטב וכמקרה פרטי של חלק (א), נובע ש-$\hat{f}(\underline{a})$ היא פונקציה קמורה ב-$¢^{M_a}$, ולכן רציפה בכל $¢^{M_a}$ (על פי T3).

לכן:
$$0 \underset{k\to\infty}{\leftarrow} \hat{f}(\underline{b}_{M_{r_k}}) \underset{k\to\infty}{\rightarrow} \hat{f}(\underline{b}) \Leftarrow \underline{b}_{M_{r_k}} \underset{k\to\infty}{\rightarrow} \underline{b}$$

ולכן $\hat{f}(\underline{b}) = 0$, $(\exists \underline{b} \in \Omega)$.

מאידך $||\cdot||_T$ היא "נורמה" על $[0,\infty)^N$ ולכן

$$\Leftarrow (\forall\ i)\ (||\sum_{k=1}^{M_i} b_{ik}v_{ik}||_i = 0) \Leftarrow \hat{f}(\underline{b}) = 0$$

מאחר ו-$||\cdot||_i$ הן נורמות ב-S_i, ו-$\{v_{ik}\}_{k=1}^{M_i}$ בסיסים ל-S_i הרי

$$\underline{b} \notin \Omega \Leftarrow ||\underline{b}||_\epsilon = 0 \Leftarrow (\forall i)\quad (\forall k)\quad (b_{ik} = 0)$$

וזו סתירה!!

מכאן ש-(A1) הוכח ולכן תכונה (ב) מתקיימת.

(ג) בדומה להוכחת (ב):

$$h(\underline{b}) = ||d_c - \sum_{k=1}^{M_c} b_k v_{ck}||_c > |(||d_c||_c - ||\sum_{k=1}^{M_c} b_k\ v_{ck}||_c)| \geq ||\sum_{k=1}^{M_c} b_k\ v_{ck}||_c - h(\underline{o})$$

נסמן:

$$\hat{h}(\underline{b}) = ||\sum_{k=1}^{M_c} b_k\ v_{ck}||_c$$

אזי: $h(\underline{b}) \geq \hat{h}(\underline{b}) - h(\underline{o})$ ולכן להוכחת תכונה (ג), מספיק להוכיח את:

$$\text{(A2)}\quad (\forall\ \tilde{\alpha} > 0)\ (\exists M < \infty)\ (\forall\ \underline{b} \in ¢^{M_c}\ ,\ ||\underline{b}||_\epsilon > M)\ (\hat{h}(\underline{b}) > \tilde{\alpha} = \alpha + h(\underline{o}))$$

כי $0 < h(\underline{o}) < \infty$.

נניח את שלילת (A2) ונראה סתירה כמו ב-(ב). נניח את שלילת (A2) קרי:

$$(\exists \tilde{\alpha} > 0)\ (\forall M)\ (\exists \underline{b}_M \in \mathbb{C}^{M_c})\ (||\underline{b}_M||_\epsilon > M\ ,\ \hat{h}(\underline{b}_M) \le \tilde{\alpha})$$

שוב, נבנה סדרה $M_r \uparrow \infty$ ואזי עבור $\underline{b}_{M_r}$ המתאימים $\lim_{r\to\infty} ||\underline{b}_{M_r}||_\epsilon = \infty$,
ו-$\hat{h}(\underline{b}_{M_r}) \le \tilde{\alpha}$. ללא הגבלת כלליות $||\underline{b}_{M_r}||_\epsilon > 0$ ואזי $\hat{\underline{b}}_{M_r} = \underline{b}_{M_r}/||\underline{b}_{M_r}||_\epsilon$.
לכן: $||\hat{\underline{b}}_{M_r}||_\epsilon = 1$, ומכיוון ש-$\hat{h}(\cdot)$ הומוגנית חיובית, הרי:

$$0 \le \lim_{r\to\infty} \hat{h}(\hat{\underline{b}}_{M_r}) \le \lim_{r\to\infty} \tilde{\alpha}/||\underline{b}_{M_r}||_\epsilon = 0 \Leftarrow \hat{h}(\hat{\underline{b}}_{M_r}) = \hat{h}(\underline{b}_{M_r})/||\underline{b}_{M_r}||_\epsilon$$

מכאן ש-$\hat{\underline{b}}_{M_r} \in \Omega$ ו- $\lim_{r\to\infty} \hat{h}(\hat{\underline{b}}_{M_r}) = 0$ $(\Omega = \{\underline{b} \in \mathbb{C}^{M_c}\ , ||\underline{b}||_\epsilon = 1\})$.
מאחר ש-Ω קומפקטית קיימת תת-סדרה $\hat{\underline{b}}_{M_{r_k}} \underset{k\to\infty}{\to} \hat{\underline{b}} \in \Omega$.
מתכונות $||\cdot||_c$ הרי $\hat{h}(\underline{b})$ קמורה, לכן $\hat{h}(\underline{b})$ רציפה, ולכן:

$$0 = \lim_{k\to\infty} \hat{h}(\hat{\underline{b}}_{M_{r_k}}) = \hat{h}(\hat{\underline{b}})$$

מאידך $\hat{h}(\hat{\underline{b}})$ היא $||\cdot||_c$ של אבר ב-S_c ולכן $\hat{h}(\hat{\underline{b}}) = 0 \Rightarrow \sum_{k=1}^{M_c} \hat{b}_k v_{ck} = 0$.
מאחר ו-v_{ck} זהו בסיס ל-S_c הרי $(\hat{b}_k = 0)\ (\forall k) \Rightarrow ||\hat{\underline{b}}||_\epsilon = 0$
ולכן $\hat{\underline{b}} \notin \Omega$ וזו סתירה.

<u>הוכחת למה 2:</u>

1. על פי (א) בלמה 1, $f(\underline{a})$ ו-$g(\underline{a})$ אי-שליליות ולכן חסומות מלמטה. מכאן ש-ε_m, η_m מוגדרים היטב ואי-שליליים. בנוסף על פי (א) בלמה 1, הרי הפונקציות $f(\underline{a})$, $g(\underline{a})$ ו-$h(\underline{b})$ הן כולן רציפות (כי הן קמורות).

2. נבחר $\alpha = \varepsilon_m + 1$, אזי קיים M כך שעבור $||\underline{a}||_\epsilon > M$, $f(\underline{a}) > \varepsilon_m + 1$, מתכונה (ב) של למה 1. מאידך מאחר ו-$\varepsilon_m = \inf_{\underline{a}\in\mathbb{C}^{M_a}} f(\underline{a})$, הרי קיימת סדרת $\underline{a}_r$ שעבורם $f(\underline{a}_r) \downarrow \varepsilon_m$ ובברור ללא הגבלת כלליות $||\underline{a}_r||_\epsilon \le M$, לאור האמור לעיל.

<u>מסקנה:</u> $\varepsilon_m = \inf_{||\underline{a}||_\epsilon \le M} f(\underline{a})$.

כאמור ב-1, $f(\underline{a})$ רציפה והקבוצה $\tilde{\Omega} = \{\underline{a} \in ¢^{M_a}, ||\underline{a}||_\epsilon \leq M\}$, היא קבוצה קומפקטית. מכאן של-$f(\underline{a})$ יש מינימום ב-$\tilde{\Omega}$ (על פי T4) ומאחר ו-ϵ_m הוא האינפימום של $f(\underline{a})$ ב-$\tilde{\Omega}$, הוא מתקבל במינימום הנ"ל.

3. נבחר $\alpha = \eta_m + 1$, אזי קיים M כך שעבור $||\underline{b}||_\epsilon > M$, $h(\underline{b}) > \eta_m + 1$, מתכונה (ג) של למה 1. מאידך $g(\underline{a}) = h(A\underline{a})$ $\forall\, \underline{a} \in ¢^{M_a}$, ו- $\text{rank}A = M_c$ (כי $\{v_{ck}\}_{k=1}^{M_c}$ זהו בסיס ל-S_c), לפיכך:

$$\eta_m = \inf_{\underline{a}\in ¢^{M_a}} g(\underline{a}) = \inf_{\underline{b}\in ¢^{M_c}} h(\underline{b})$$

קיימת סדרת $\underline{b}_r$ שעבורם $h(\underline{b}_r) \downarrow \eta_m$ ובברור ללא הגבלת כלליות $||\underline{b}_r||_\epsilon \leq M$, לאור האמור לעיל.

<u>מסקנה:</u>

$$\eta_m = \inf_{||\underline{b}||_\epsilon \leq M} h(\underline{b})$$

שוב כאמור ב-1, $h(\underline{b})$ רציפה והקבוצה $\tilde{\Omega} = \{\underline{b} \in ¢^{M_c}, ||\underline{b}||_\epsilon \leq M\}$, היא קבוצה קומפקטית. מכאן של-$h(\underline{b})$ יש מינימום ב-$\tilde{\Omega}$ (על פי T4) ומאחר ו-η_m הוא האינפימום ב-$\tilde{\Omega}$ של $h(\underline{b})$, הוא מתקבל במינימום הנ"ל.

4. לאור 3, הרי קיים $\underline{a}_o$ שעבורו $\eta_m = g(\underline{a}_o)$ ולכן לכל $\eta \geq \eta_m$, הקבוצה $\zeta(\eta) \triangleq \{\underline{a} \in ¢^{M_a}; g(\underline{a}) \leq \eta\}$ אינה ריקה. לכן האינפימום של $0 \leq f(\underline{a})$ על $\zeta(\eta)$ קיים וזהו $\epsilon(\eta)$. נבחר $\alpha = \epsilon(\eta) + 1$ אזי קיים M כך שעבור $||\underline{a}||_\epsilon > M$, $f(\underline{a}) > \epsilon(\eta) + 1$, מתכונה (ב) של למה 1. קיימת סדרת $\underline{a}_r \in \zeta(\eta)$ שעבורם $f(\underline{a}_r) \downarrow \epsilon(\eta)$ ובברור ללא הגבלת כלליות $||\underline{a}_r||_\epsilon \leq M$, לאור האמור לעיל.

<u>מסקנה:</u> $\epsilon(\eta) = \inf_{\underline{a}\in\zeta(\eta)\cap B(M)} \{f(a)\}$ כאשר: $B(M) \triangleq \{\underline{a}\in ¢^{M_a}, ||\underline{a}||_\epsilon \leq M\}$.

עקב הרציפות של $g(\underline{a})$ (מ-1) הרי $\zeta(\eta)$ סגורה (על פי T5), ומאחר ו-$B(M)$ קומפקטית הרי גם $A(\eta)=\zeta(\eta)\cap B(M)$ קומפקטית (על פי T6). מאחר ו-$f(\underline{a})$ רציפה ב-$¢^{M_a}$ (על פי 1) הרי האינפימום $\epsilon(\eta)$ מתקבל על ידי מינימום.

מאחר ו-$g(\underline{a})$ קמורה הקבוצה $\zeta(\eta)$ קמורה (על פי T12), ולכן גם הקבוצה $D(\eta)$ קמורה לאור קמירות $f(\underline{a})$ ו-T14.

5. מאחר ו-ε_m מתקבל במינימום על פי 2, הרי קיים $\underline{a}_o \in \not{C}^{M_a}$ כך ש- $f(\underline{a}_o) = \varepsilon_m$.
בנוסף $g(\underline{a}_o) < \infty$, ולכל $\eta \geq g(\underline{a}_o)$ הרי $\underline{a}_o \in A(\eta)$ ולכן: $\varepsilon(\eta) = \varepsilon_m$.
<u>מסקנה:</u> הקבוצה $c_m \triangleq \{\eta \,;\, \varepsilon(\eta) = \varepsilon_m\}$ אינה ריקה.
בנוסף מכיון ש- $\varepsilon(\eta)$ סופי רק ל- $\eta \geq \eta_m$ הרי $c_m \subset [\eta_m, \infty)$. יתר על-כן, מאחר ו-$A(\eta_1) \subset A(\eta_2)$ לכל $\eta_1 \leq \eta_2$, הרי $\varepsilon(\eta)$ מונוטונית לא-עולה ב-η ב-$[\eta_m, \infty)$. בנוסף $\forall\, \eta\ \varepsilon(\eta) \geq \varepsilon_m$ על פי הגדרת ε_m. מכאן שאם $\eta_1 \in c_m$ הרי $[\eta_1, \infty) \subset c_m$. כמסקנה, מכך נובע ש-$c_m$ הוא אינטרוול יחיד בתוך $[\eta_m, \infty)$.
לפיכך האינפימום של η ב-c_m שיסומן על ידי η_M תמיד מוגדר והוא מקיים $\eta_M \geq \eta_m$. האינפימום מתקבל במינימום אם"ם c_m הוא אינטרוול חצי-סגור (מלמטה).

6. קיים M כך ש-$\underline{a} \notin B(M) \Leftarrow f(\underline{a}) > \varepsilon_m + 1$ מתכונה (ב) של למה 1.
לכן הקבוצה $\nu = \{a \in \not{C}^{M_a}, f(\underline{a}) = \varepsilon_m\}$ מוכלת בתוך $B(M)$ והיא לא ריקה על פי 2.
עבור כל $\eta \geq \eta_M \geq \eta_m$ גם $\zeta(\eta)$ אינה ריקה (ראה 4 דלעיל).
עתה נגדיר: $\nu(\eta) = \nu \cap \zeta(\eta) = \{\underline{a} \in \not{C}^{M_a}, f(\underline{a}) = \varepsilon_m, g(\underline{a}) \leq \eta\}$

על פי הגדרת η_M, הרי $\nu(\eta)$ אינו-ריק עבור כל $\eta > \eta_M$. $f(\underline{a})$ ו-$g(\underline{a})$ קמורות ולכן רציפות ולכן הקבוצות ν ו-$\zeta(\eta)$ סגורות (על פי T5).
לכן (על פי T6) הקבוצה $\nu(\eta)$ סגורה לכל $\eta > \eta_M$.

האוסף $\{\nu(\eta)\}_{\eta > \eta_M}$ הוא אוסף של קבוצות סגורות בעל תכונת החיתוך הסופי, שכן $\bigcap_{i=1}^{n} \nu(\eta_i) = \nu(\min_{1 \leq i \leq n} \{\eta_i\}) \neq \phi$ (מכיון ש-$\zeta(\eta_2) \supset \zeta(\eta_1)$ לכל $\eta_1 > \eta_2$).
קבוצות אלו מוכלות ב-$B(M)$ שהיא קומפקטית (על פי T1), ולכן $\bigcap_{\eta > \eta_M} \nu(\eta)$ אינו ריק, (על פי T7). יהא $\underline{a}_o \in \bigcap_{\eta > \eta_M} \nu(\eta)$ אזי:

$$f(\underline{a}_o) = \varepsilon_m \ \&(\forall\, \eta > \eta_M)\ (g(\underline{a}_o) \leq \eta)$$

מכאן ש- $f(a_o) = \varepsilon_m \ \&\ g(\underline{a}_o) \leq \eta_M$, דהיינו $\varepsilon(\eta_M) = \varepsilon_m$ ולכן $\eta_M \in c_m$ והאינפימום מתקבל במינימום.

<u>הוכחת משפט 1</u>:

(א) הראינו כבר ב-5 בהוכחת למה 2, ש-$\varepsilon(\eta)$ לא-עולה ב-$[\eta_m,\eta_M]$. מתוך למה 2 חלק (ג), נובע ש- $\varepsilon(\eta) < \infty$ לכל $\eta \geq \eta_m$, לכן $\varepsilon(\eta)$ חסומה מלמעלה ב-$[\eta_m,\eta_M]$ על ידי $\varepsilon(\eta_m) = \varepsilon_M$. בנוסף $\varepsilon(\eta_M) = \varepsilon_m \geq 0$ ולכן $\varepsilon(\eta)$ חסומה גם מלמטה ב-$[\eta_m,\eta_M]$.

על מנת להדאות ש-$\varepsilon(\eta)$ קמורה ב-$[\eta_m,\eta_M]$ נגדיר את הפונקציה: $\phi(\underline{a},\eta)$ על $\mathcal{C}^{M_a} \times R$ כדלקמן:

$$\phi(\underline{a},\eta) = \begin{cases} f(\underline{a}) & \underline{a} \in \zeta(\eta) \\ +\infty & \underline{a} \notin \zeta(\eta) \end{cases}$$

ואזי, $\varepsilon(\eta) = \inf_{\underline{a}\in\mathcal{C}^{M_a}} \phi(\underline{a},\eta)$ (על פי הגדרת $\varepsilon(\eta)$), והיא סופית עבור $\eta \geq \eta_m$.

על פי ההגדרה $\phi(\underline{a},\eta) > -\infty$, כי $-\infty < f(\underline{a})$ בכל $\mathcal{C}^{M_a}$, וכן עבור η גדול מספיק $\zeta(\eta)$ אינה ריקה ולכן $\phi(\underline{a},\eta)$ אינה $+\infty$ זהותית. לפיכך $\phi(\underline{a},\eta)$ היא פונקציה כראוי (proper). יהיו $\underline{a}_1 \in \zeta(\eta_1)$ ו-$\underline{a}_2 \in \zeta(\eta_2)$, אזי מקמירות $g(\underline{a})$: $\lambda\underline{a}_1 + (1-\lambda)\underline{a}_2 \in \zeta(\lambda\eta_1 + (1-\lambda)\eta_2)$ לכל $0 \leq \lambda \leq 1$, ומקמירות $f(\underline{a})$ נובע ש-

$$\lambda\phi(\underline{a}_1,\eta_1) + (1-\lambda)\phi(\underline{a}_2,\eta_2) \geq \phi(\lambda\underline{a}_1 + (1-\lambda)\underline{a}_2,\lambda\eta_1 + (1-\lambda)\eta_2)$$

לכן $\phi(\underline{a},\eta)$ קמורה כראוי בכל $\mathcal{C}^{M_a} \times R$, ולפי T11 $\varepsilon(\eta)$ קמורה ב-R.

הפונקציה $\varepsilon(\eta)$ קמורה וחסומה באינטרוול סגור ולכן היא רציפה באינטרוול הפתוח (על פי T8). מאחר והיא גם מונוטונית לא עולה הרי קיימים גבולות בקצוות האינטרוול על פי T9. מאותה סיבה הרי $\varepsilon(\eta_M) \leq \varepsilon(\eta) \leq \lim_{\eta\uparrow\eta_M} \varepsilon(\eta) \triangleq \varepsilon(\eta_M^-)$ וכן:

$$\varepsilon(\eta_m^+) \triangleq \lim_{\eta\downarrow\eta_m} \varepsilon(\eta) \leq \varepsilon(\eta_m) = \varepsilon_M$$

נניח כי $\varepsilon(\eta_M^-) > \varepsilon(\eta_M)$ ונקבל סתירה כדלקמן:

נעביר דרך הנקודה $(\eta_M,\varepsilon(\eta_M))$ קו-ישר בזוית מעל 90°, מספיק קטנה כך שהנקודה (η_m,ε_M) תמצא מתחתיו של הקו. מאחר ו-$\varepsilon(\eta_M) > \varepsilon(\eta_M^-)$ הרי קיימת $h > 0$ קטנה מספיק כך שהנקודה $(\eta_M-h,\varepsilon(\eta_M-h))$ היא מעל הקו הישר הנ"ל. מאחר והקטע (η_m,η_M-h) הפונקציה $\varepsilon(\eta)$ היא רציפה ו-$\varepsilon_M \geq \varepsilon(\eta_M^+)$ הרי יש נקודה $(\eta_2,\varepsilon(\eta_2))$ שנמצאת על הקו הנ"ל כש-$\eta_m < \eta_2 < \eta_M-h$.

עבור שלושת הערכים $\eta_2 < (\eta_M-h) < \eta_M$ נוצרה סתירה לקמירות של $\varepsilon(\eta)$.

את הרציפות ב-η_m נוכיח על ידי ארגומנט דומה לזה שהופעל בהוכחת למה 2 סעיף 6. יהא $B(M)$ כדור מספיק גדול, כך ש-

$$f(\underline{a}) > \varepsilon(\eta_m^+) + 1 \Leftarrow ||\underline{a}||_\varepsilon > M$$

ויהא:

$$\beta = \{\underline{a} \in \not{C}^{M_a},\ f(\underline{a}) \leq \varepsilon(\eta_m^+)\}$$

אזי $\beta \subset B(M)$ וסגורה כי $f(\underline{a})$ רציפה ב-$B(M)$, ולא ריקה כי $\varepsilon(\eta_m^+) \geq \varepsilon_m$ על פי מונוטוניות $\varepsilon(\eta)$. בנוסף גם $\beta(\eta) = \beta \cap \zeta(\eta)$ היא לא-ריקה לכל $\eta > \eta_m$ על פי המונוטוניות של $\varepsilon(\eta)$ והגדרת: $\varepsilon(\eta_m^+) \triangleq \lim_{\eta \downarrow \eta_m} \varepsilon(\eta)$.

יתר על כן, כמו בסעיף 6, בהוכחת למה 2, $\beta(\eta) \subset B(M)$ והן קבוצות סגורות (כי גם $g(\underline{a})$ רציפה ולכן $\zeta(\eta)$ סגורות). $B(M)$ סגורה וחסומה ב-$\not{C}^{M_a}$ ולכן קומפקטית (T1).

$\{\beta(\eta)\}_{\eta>\eta_m}$ אוסף תת-קבוצות סגורות של $B(M)$. בנוסף $\beta(\eta_1) \subset \beta(\eta_2)$ לכל $\eta_1 < \eta_2$, על פי הגדרתן ולכן:

$$\phi \neq \beta(\min_{1\leq i\leq n}\eta_i) = \bigcap_{i=1}^{n} \beta(\eta_i)$$

מאחר ולאוסף יש את תכונת החיתוך הסופי הלא-ריק הרי (על פי T4) גם $\bigcap_{\eta>\eta_m} \beta(\eta)$ אינו-ריק. תהא $\underline{a}_o \in \bigcap_{\eta>\eta_m} \beta(\eta)$ אזי:

$$\varepsilon(\eta_m) \leq \varepsilon(\eta_m^+) \Leftarrow g(\underline{a}_o) \leq \eta_m \ \&\ f(\underline{a}_o) \leq \varepsilon(\eta_m^+) \Leftarrow (\forall\ \eta>\eta_m)(g(\underline{a}_o)\leq\eta\ \&\ f(\underline{a}_o)\leq\varepsilon(\eta_m^+))$$

ולכן $\varepsilon(\eta_m^+) = \varepsilon(\eta_m)$ והושגה רציפות ב-η_m.

פונקציה לא-עולה, קמורה, רציפה וחסומה באינטרוול הסגור $[\eta_m,\eta_M]$ יכולה לא-לרדת רק בתת-אינטרוול סגור $[\eta_o,\eta_M]$, וזאת מכיון שאם ב-$[\eta_1,\eta_2]$ הפונקציה $\varepsilon(\eta)$ קבועה ו-$\eta_2 < \eta_M$ אזי ישנו h קטן מספיק כך ש-
$\varepsilon(\eta_2-h) = \varepsilon(\eta_2) > \varepsilon(\eta_2+h)$ ו-$\varepsilon(\eta_2)$ מעל הקו המחבר את $(\eta_2-h\ ,\varepsilon(\eta_2-h))$ עם $(\eta_2+h\ ,\varepsilon(\eta_2+h))$ וזו סתירה לקמירות של $\varepsilon(\eta)$. בנוסף על פי הגדרת η_M בלמה 2 הרי $\varepsilon(\eta)$ יורדת ממש בסביבה קטנה של η_M, ולכן יורדת ממש בכל $[\eta_m,\eta_M]$.

(ב) על פי משפט T10, הרי כל הנגזרות משמאל ומימין של $\varepsilon(\eta)$ קיימות ב-$[\eta_m,\eta_M]$ (אם כי אולי ערכן אינסופי) והפונקציה $\varepsilon'(\eta)$ מונוטונית לא-יורדת. מאחר ו-$\varepsilon(\eta)$ היא מונוטונית לא-עולה בכל $[\eta_m,\infty)$ (על פי הגדרתה), הרי $\varepsilon'(\eta)$ היא אי-חיובית ב-$[\eta_m,\eta_M]$. נותר רק להוכיח ש-$\varepsilon'(\eta)$ סופית בכל $(\eta_m,\eta_M]$. עקב המונוטוניות של $\varepsilon'(\eta)$ וחסימותה מלמעלה, מספיק להוכיח כי $-\infty < \varepsilon'_-(\eta_m+h)$ עבור כל $h > 0$ קטן כרצוננו.

מהקמירות של $\varepsilon(\eta)$ נובע שלכל $0 \leq \Delta \leq h$:

$$\frac{\varepsilon(\eta_m) - \varepsilon(\eta_m+h)}{(-h)} \leq \frac{\varepsilon(\eta_m+h-\Delta) - \varepsilon(\eta_m+h)}{(-\Delta)}$$

מאחר ש-$\varepsilon(\eta)$ חסומה ב-$[\eta_m,\eta_M]$ הרי לכל $h > 0$ (קבוע):

$$-\infty < \frac{\varepsilon(\eta_m) - \varepsilon(\eta_m+h)}{(-h)} \triangleq a_h$$

ולכן לכל סדרה $\Delta_n \to 0^+$ הגבול:

$$-\infty < a_h \leq \lim_{n\to\infty} \frac{\varepsilon(\eta_m+h-\Delta_n) - \varepsilon(\eta_m+h)}{(-\Delta_n)} \triangleq \varepsilon'_-(\eta_m+h)$$

ולכן $\varepsilon'_-(\eta_m+h)$ סופי לכל $h > 0$ קטן כרצוננו.

(ג) נניח בשלילה, שקיים $\eta_1 \in [\eta_m,\eta_M]$ ו-$\underline{a}_o \in D(\eta_1)$, כך ש- $g(\underline{a}_o) < \eta_1$.
נגדיר $\eta_2 \triangleq g(\underline{a}_o)$, אזי $\underline{a}_o \in \zeta(\eta_2)$, ולכן: $\varepsilon(\eta_2) \leq f(\underline{a}_o) = \varepsilon(\eta_1)$,
כש-$\eta_2 < \eta_1$ עבור $\eta_1 > \eta_m$ זו סתירה לעובדה ש-$\varepsilon(\eta)$ פונקציה יורדת
ממש באינטרוול, ועבור $\eta_1 = \eta_m$ זו סתירה להגדרת η_m בלמה 2.
מכאן שהוכחנו בדרך השלילה את הטענה הדרושה.

(ד) 1. נוכיח את הקמירות ממש של $\varepsilon(\eta)$ בדומה להוכחה ב-(א) דלעיל. יהיו
$\eta_1 \neq \eta_2 \in [\eta_m,\eta_M]$, $\lambda \in (0,1)$, $\underline{a}_1 \in D(\eta_1)$-ו, $\underline{a}_2 \in D(\eta_2)$
כלשהם. נסמן: $\underline{a}_\lambda \triangleq \lambda\underline{a}_1 + (1-\lambda)\underline{a}_2 \in \phi^{M_a}$. מקמירות $g(\underline{a})$
נובע ש-$\underline{a}_\lambda \in \zeta(\lambda\eta_1 + (1-\lambda)\eta_2)$, ולכן על פי הגדרת $\varepsilon(\eta)$ הרי
$\varepsilon(\lambda\eta_1 + (1-\lambda)\eta_2) \leq f(\underline{a}_\lambda)$. מהקמירות ממש של $f(\underline{a})$:

$$f(\underline{a}_\lambda) < \lambda f(\underline{a}_1) + (1-\lambda)f(\underline{a}_2) = \lambda\varepsilon(\eta_1) + (1-\lambda)\varepsilon(\eta_2)$$

ומכאן ש-$\varepsilon(\eta)$ קמורה ממש.

2. על מנת להוכיח ש-$\varepsilon'(\eta)$ פונקציה מונוטונית עולה-ממש, עלינו רק להראות
ש-$\varepsilon'_-(\eta_2) > \varepsilon'_+(\eta_1)$ לכל $\eta_M \geq \eta_2 > \eta_1 \geq \eta_m$.
נסמן $h \triangleq (\eta_2-\eta_1)/2 > 0$, אזי מהקמירות ממש של $\varepsilon(\eta)$ נובע ש:

$$\frac{\varepsilon(\eta_2-h)-\varepsilon(\eta_2)}{(-h)} > \frac{\varepsilon(\eta_1+h)-\varepsilon(\eta_1)}{(h)}$$

באותו ארגומנט כמו בהוכחת סעיף (ב) הרי:

$$\varepsilon'_-(\eta_2) \geq \frac{\varepsilon(\eta_2-h)-\varepsilon(\eta_2)}{(-h)} \ , \ \frac{\varepsilon(\eta_1-h)-\varepsilon(\eta_1)}{h} \geq \varepsilon'_+(\eta_1)$$

והמונוטוניות ממש של $\varepsilon'(\eta)$ נובעת מאיחוד שלושת האי-שיוויונות דלעיל.

3. ידוע לנו כבר ש-$D(\eta)$ אינן ריקות עבור $\eta \geq \eta_m$. כמו כן
מקמירות $g(\underline{a})$ נובע ש-$\zeta(\eta)$ הן קבוצות קמורות (על פי T12),
ומהקמירות ממש של $f(\underline{a})$ נובעת היחידות של $\underline{a}^*$ (על פי T13).

<u>הוכחת למה 3:</u>

יהא: $\underline{a}^{(1)} \neq \underline{a}^{(2)} \in \mathbb{C}^{M_a}$, $\lambda \in (0,1)$. נגדיר: $\underline{a}^{(\lambda)} = \lambda\underline{a}^{(1)} + (1-\lambda)\underline{a}^{(2)}$

אזי מהקמירות של $f(\underline{a})$ נובע: $f(\underline{a}^{(\lambda)}) \leq \lambda f(\underline{a}^{(1)}) + (1-\lambda)f(\underline{a}^{(2)})$.

אם נגדיר את $\underline{\delta}^{(2)}$, $\underline{\delta}^{(1)}$, $\underline{\delta}^{(\lambda)}$ כוקטורי שגיאות הקירוב המתאימים ל- $\underline{a}^{(2)}$, $\underline{a}^{(1)}$,

$\underline{a}^{(\lambda)}$, הרי בדומה להוכחת למה 1, גם כאן: $\lambda\underline{\delta}^{(1)} + (1-\lambda)\,\underline{\delta}^{(2)} = \underline{\delta}^{(\lambda)} + \underline{\Delta}$

כש- $\underline{\Delta} \in [0,\infty)^N$.

עבור $\underline{\Delta} \neq \underline{0}$ נקבל מתנאי (א) של הלמה:

$$f(\underline{a}^{(\lambda)}) = ||\underline{\delta}^{(\lambda)}||_T < ||\underline{\delta}^{(\lambda)}+\underline{\Delta}|| \leq \lambda||\underline{\delta}^{(1)}||_T + (1-\lambda)||\underline{\delta}^{(2)}||_T =$$
$$= \lambda f(\underline{a}^{(1)}) + (1-\lambda)f(\underline{a}^{(2)})$$

לכן הקמירות ממש של $f(\underline{a})$ מובטחת ובלבד שנוכיח שלכל $\underline{a}^{(1)} \neq \underline{a}^{(2)}$

ו-$\lambda \in (0,1)$, הרי $\underline{\Delta} \neq \underline{0}$.

נוכיח זאת בדרך השלילה. נניח של-$\underline{a}^{(1)} \neq \underline{a}^{(2)}$ ו-$\lambda \in (0,1)$ מתקבל $\underline{\Delta} = \underline{0}$

קרי:

$$(\forall\ 1 \leq i \leq N) \qquad (\lambda\delta_i^{(1)} + (1-\lambda)\delta_i^{(2)} = \delta_i^{(\lambda)})$$

נסמן על ידי; $\underline{u}^{(2)}$, $\underline{u}^{(1)}$, $\underline{u}^{(\lambda)}$ את הוקטורים ב- $S_1 \times \ldots \times S_N$ המתאימים

ל- $\underline{a}^{(2)}$, $\underline{a}^{(1)}$, $\underline{a}^{(\lambda)}$ בהתאמה.

אזי נתון שלכל $1 \leq i \leq N$:

$$\lambda||d_i-u_i^{(1)}||_i + (1-\lambda)||d_i - u_i^{(2)}||_i = ||d_i-\lambda u_i^{(1)}-(1-\lambda)u_i^{(2)}||_i$$

מהלינאריות של $||\cdot||_i$ ומתנאי (ג) של הלמה (קמירות ממש), נובע ש-

$$\lambda(d_i-u_i^{(1)}) = |\alpha_i|(1-\lambda)\ (d_i-u_i^{(2)}) \qquad 1 \leq i \leq N$$

נפריד לשני מקרים:

(1) $\lambda = |\alpha_i|\ (1-\lambda) \neq 0$ לכל $1 \leq i \leq N$, אזי על ידי צמצום ב-λ נקבל:

$1 \leq i \leq N$, $u_i^{(1)} = u_i^{(2)}$ ולכן $\underline{a}_i^{(1)} = \underline{a}_i^{(2)}$ לכל i ולכן $\underline{a}_1 = \underline{a}_2$ בסתירה לנתון.

(2) קיים i כך ש- $(1-\lambda) |\alpha_i| \neq \lambda$, אזי עבור i-זה, ישנה קומבינציה לא אפס זהותית של d_i ואברי הבסיס של S_i שמתאפסת. לכן $d_i \in S_i$ ואזי:

$M_i = \dim SD_i = \dim S_i$ וזו סתירה לתנאי (ב) של הלמה.

מכיון שבשני המקרים התקבלה סתירה, ההוכחה הושלמה.

<u>הוכחת ההערה:</u>

(א) מאחר ו-$D(\eta)$ קמורה (על פי למה 2) הרי לכל $\underline{a}^{*(1)} \neq \underline{a}^{*(2)} \in D(\eta)$, ולכל $\lambda \in (0,1)$, $\underline{a}^{*(\lambda)} \in D(\eta)$, ולכן: $f(\underline{a}^{*(\lambda)}) = f(\underline{a}^{*(1)}) + (1-\lambda) f(\underline{a}^{*(2)})$.

בעקבות הוכחת הלמה דלעיל: $\underline{\delta}^{(\lambda)} = \lambda\underline{\delta}^{(1)} + (1-\lambda)\underline{\delta}^{(2)}$ על פי תנאי (א) של הלמה. יהא $i \in I$ כלשהו, אזי שוב בעקבות הוכחת הלמה דלעיל מקמירות-ממש של $||\cdot||_i$ והעובדה ש-$\dim SD_i = (M_i+1)$ נובע ש-
$0 \neq (1-\lambda)|\alpha_i| = \lambda$ ולכן: $\underline{a}_i^{*(1)} = \underline{a}_i^{*(2)}$ כאמור בהערה.

(ב) 2. אם תנאי (ג) מתקיים לכל i, אך תנאי (ב) מתקיים רק עבור $(N-1)$ i-ים, אזי נסמן ב-i_o, את זה שעבורו תנאי (ב) לא מתקיים. מחלק (א) של ההערה ברור ש- $\underline{a}_i^{*(1)} = \underline{a}_i^{*(2)}$ לכל $i \neq i_o$, ולכן $\delta_i^{(1)} = \delta_i^{(2)}$ לכל $i \neq i_o$. מאידך $f(\underline{a}^{*(1)}) = f(\underline{a}^{*(2)})$ מחייב שגם $\delta_{i_o}^{(1)} = \delta_{i_o}^{(2)}$ לאור המונוטוניות ממש של $||\cdot||_T$ על פי תנאי (א), והעובדה ש- $\delta_i^{(1)} = \delta_i^{(2)}$ ל- $i \neq i_o$. מאחר וגם $||\cdot||_{i_o}$ היא קמורה ממש, הרי:

$$\lambda(d_{i_o}-u_{i_o}^{(1)}) = |\alpha_{i_o}|(1-\lambda)\ (d_{i_o}-u_{i_o}^{(2)})$$

עתה $\delta_{i_o}^{(1)} = \delta_{i_o}^{(2)} \Leftarrow (1-\lambda)|\alpha_{i_o}| = \lambda$ ולכן $\underline{a}_{i_o}^{*(1)} = \underline{a}_{i_o}^{*(2)}$ כאמור בהערה.

1. מתנאי (א) של הלמה דלעיל ראינו שעבור $\underline{a}^{*(1)} \neq \underline{a}^{*(2)}$ מתקיים $\underline{\delta}^{(\lambda)} = \lambda\underline{\delta}^{(1)} + (1-\lambda)\ \underline{\delta}^{(2)}$ על מנת לקבל שיוויון של

$$f(\underline{a}^{*(2)}) = f(\underline{a}^{*(1)}) = f(\underline{a}^{*(\lambda)})$$

עבור $||\cdot||_T$ שהיא נורמה קמורה-ממש ב-$[0,\infty)^N$, השיוויון הנ"ל מתקיים רק כאשר $\underline{\delta}^{(2)} (1-\lambda) |\alpha| = \lambda\underline{\delta}^{(1)}$. מכיון ש-$||\cdot||_T$ מונוטונית ממש הרי $0 \neq (1-\lambda) |\alpha| = \lambda$, כדי לקבל $f(\underline{a}^{*(2)}) = f(\underline{a}^{*(1)})$. עתה מתנאי (ג) של הלמה דלעיל (קמירות ממש של כל ה-$||\cdot||_i$) נובע ש-

$\lambda(d_i-u_i^{(1)}) = |\alpha_i|\,(1-\lambda)\,(d_i-u_i^{(2)})$ ולכן $\underline{\delta}^{(1)} = \underline{\delta}^{(2)}$, יכפה $0 \neq (1-\lambda)\,|\alpha_i| = \lambda$
לכל $N \geq i \geq 1$ ולכן $\underline{a}^{*(1)} = \underline{a}^{*(2)}$, כאמור בהערה.

הוכחת משפט 2:

משפט 2 הוא מקרה פרטי של משפטי הדואליות של בעיות תכנות קמור (ראה ב-[56,62]).
ניתן להוכיחו ישירות ללא שימוש בטרמינולוגיה של תכנות קמור ודואליות, אך
העדפנו הוכחה המנצלת מושגים אלה. נגדיר את הבעיה הפרימלית הסטנדרטית עבור η
(שנסמן ב-SP_η) כדלקמן: $SP_\eta : \inf \{f(\underline{a})\ ;\ \underline{a} \in ¢^{M_a}, \eta-g(\underline{a}) \geq 0\}$.

זו בעיה סטנדרטית של תכנות קמור כמתואר ב-[56,sec.5.2], כי $f(\underline{a})$ קמורה
ו-$g(\underline{a}) - \eta$ קעורה. נשייך לבעיה זו את הפונקציה:

$$\phi_\eta(\underline{a},\omega) = \begin{cases} f(\underline{a}) & \underline{a} \in \zeta(\eta-\omega) \\ \infty & \underline{a} \notin \zeta(\eta-\omega) \end{cases}$$

כש-$\omega \in R$. זו פונקציה קמורה כראוי בכל $¢^{M_a} \times R$ (הוכחה ראה בהוכחת
משפט 1, סעיף א), קבוצות הרמה שלה $S(\phi_\eta,\alpha)$ הן על פי הגדרתן מהצורה:

$$S(\phi_\eta,\alpha) = \{(\underline{a},\omega);\ \ f(\underline{a}) \leq \alpha\ \&\ \underline{a} \in \zeta(\eta-\omega)\}$$

נראה שהקבוצות $S(\phi_\eta,\alpha)$ הן סגורות לכל α ואזי ϕ_η זו פונקציה סגורה בכל $¢^{M_a} \times R$
(על פי ההגדרה של [56,sec. 4.2]). תהא $\{(\underline{a}_k, \omega_k)\}_{k=1}^{\infty}$ סדרה המתכנסת
ל-$(\underline{a},\omega)$ (דהיינו: $\omega_k \to \omega$ ו- $\underline{a}_k \to \underline{a}$), כך ש-$(\underline{a}_k,\omega_k) \in S(\phi_\eta,\alpha)$.
אזי $f(\underline{a}_k) \leq \alpha$ לכל k ולכן גם $f(\underline{a}) \leq \alpha$ (על פי הרציפות של f לאור T3).
בנוסף $\underline{a}_k \in \zeta(\eta-\omega_k)$ לכל k קרי $g(\underline{a}_k) + \omega_k \leq \eta$. מאחר ו-$\omega_k \to \omega$
ו-$\underline{a}_k \to \underline{a}$ והפונקציה $g(\underline{a}) + \omega$ רציפה בכל $¢^{M_a} \times R$ (לאור T3) הרי
$g(\underline{a}) + \omega \leq \eta$ ולכן $\underline{a} \in \zeta(\eta-\omega)$, מכאן ש-$(\underline{a},\omega) \in S(\phi_\eta,\alpha)$ ולכן
אלו קבוצות סגורות.

עתה הבעיה הפרימלית שקשורה ל-$\phi_\eta(\underline{a},\omega)$ היא: $P\phi_\eta \; : \; \inf_{\underline{a}\in\mathbb{C}^{M_a}} \{\phi_\eta(\underline{a},0)\}$

ולבעיה זו ידוע לנו שקבוצת הפתרונות היא הקבוצה הקמורה והלא-ריקה $D(\eta)$ עבור $\eta_m \geq \eta$. פונקצית הפרטורבציה של הבעיה הפרימלית $\Phi_\eta(\omega)$ היא:

$$\Phi_\eta(\omega) = \inf_{\underline{a}\in\mathbb{C}^{M_a}} \{\phi_\eta(\underline{a},\omega)\} = \varepsilon(\eta-\omega)$$

כאשר P_{ϕ_η} ו-$\Phi_\eta(\omega)$ מוגדרות על פי [56,sec. 5.2], ולאור T11 $\Phi_\eta(\omega)$ היא פונקציה קמורה ב-ω.

עבור $\eta \geq \eta_m$, $\Phi_\eta(0) < \infty$ על פי למה 2 סעיף (ג). התת-דיפרנציאל של Φ_η ב-$\omega = 0$ שיסומן על ידי $\partial\Phi_\eta(0)$ מכיל את כל ה-λ כך שלכל ω:

$$\Phi_\eta(\omega) = \varepsilon(\eta-\omega) \geq \Phi_\eta(o) + \lambda\omega = \varepsilon(\eta) + \lambda\omega$$

לאור T15 הרי :

$$\partial\Phi_\eta(o) = \{\lambda; \;\; \varepsilon'_+(\eta) \geq -\lambda \geq \varepsilon'_-(\eta)\}$$

וזו קבוצה לא-ריקה לכל $\eta \in (\eta_m,\eta_M]$ על פי משפט 1 סעיף (ב), לכן על פי ההגדרה של [56, sec. 5.2] הבעיה P_{ϕ_η} יציבה, לכל $\eta \in (\eta_m,\eta_M]$ וגם ל-$\eta = \eta_m$ אם $\varepsilon'_+(\eta_m) > -\infty$.

על פי ההגדרה של [56, sec. 5.3] הרי הלגרנז'יאן של P_{ϕ_η} הוא בדיוק $L(\underline{a},K) - K\eta$ עבור $K \geq 0$ ו-$-\infty$ עבור $K < 0$ (כשכופל לגרנז' λ שווה ל-K).

עתה ניגש להוכחת משפט 2.

(א) לאור T16 ו-T17 לכל $\underline{a}^* \in D(\eta)$ עבור $\eta > \eta_m$ (ולכל $\underline{a}^* \in D(\eta_m)$ כאשר $\varepsilon'_+(\eta_m) > -\infty$), ולכל $K^* \geq 0$ כך ש-$\varepsilon'_-(\eta) \geq -K^* \geq \varepsilon'_+(\eta)$, $(\underline{a}^*,+K^*)$ היא נקודת אוכף של הלגרנז'יאן של P_{ϕ_η}, ולפיכך גם $\underline{a}^* \in \xi(K^*)$.

לפיכך: $\xi(K) \supset \bigcup_{\eta\in I_K} D(\eta)$ כאשר: $I_K = \{\eta \; ; \; \varepsilon'_+(\eta) \geq -K > \varepsilon'_-(\eta)\}$.

עבור $K = 0$ הרי $L(\underline{a},0) = f(\underline{a})$ ולכן בברור $\xi(o) = \bigcup_{\eta\in[\eta_M,\infty)} D(\eta)$ על פי למה 2 סעיף (ד).

עבור $K > 0$ הרי לכל $\eta \in I_K$ קיים $0 > \varepsilon'_-(\eta)$ ולכן $I_K \subset [\eta_m,\eta_M]$. על פי משפט 1 סעיף (ב) $\varepsilon'(\eta)$ היא פונקציה מונוטונית לא-יורדת, ובנוסף $\varepsilon'_+(\eta_M) = 0$ ו- $\varepsilon'_-(\eta_m) = -\infty$, ולכן I_K הוא אינטרוול לא ריק בתוך $[\eta_m,\eta_M]$ לכל $0 < K < \infty$, ולכן גם $\xi(K)$ היא קבוצה לא-ריקה לכל K ומתקבל ערך סופי ל-$\psi(K)$.

יהא $\eta \in I_K$ כלשהו ו-$\underline{a} \in D(\eta)$, אזי: $\psi(K) = L(\underline{a},K) = \varepsilon(\eta) + K\eta$ ולכן:

$$\psi(K) = \inf_{\eta\in[\eta_m,\eta_M]} \{\varepsilon(\eta) + K\eta\}$$

נניח שקיים: $\underline{a} \in \xi(K) - \bigcup_{\eta\in I_K} D(\eta)$ עבור $K \in (0,\infty)$ כלשהו, אזי $g(\underline{a}) \in [\eta_m,\infty)$ על פי למה 2 סעיף (ב). נסמן ב-$\eta_0 \triangleq g(\underline{a})$, אזי אם $\underline{a} \notin D(\eta_0)$ הרי $f(\underline{a}) > \varepsilon(\eta_0)$ ולכן קיים $\underline{a}^* \in D(\eta_0)$ כך ש- $L(\underline{a}^*,K) < L(\underline{a},K)$, בסתירה להנחה ש-$\underline{a} \in \xi(K)$. לפיכך $\underline{a} \in D(\eta_0)$ ו-$\eta_0 \notin I_K$. אך $\underline{a} \in \xi(K)$ ולכן: $\psi(K) = L(\underline{a},K) = \varepsilon(\eta_0) + K\eta_0$ אך מכאן שלכל $\omega \in R$ $\varepsilon(\eta_0) + K\eta_0 \le \varepsilon(\eta_0-\omega) + K(\eta_0-\omega)$ ולכן $K \in \partial\Phi_{\eta_0}(0)$. לאור T15 נקבל ש-$\eta_0 \in I_K$, בסתירה להנחה. לסיכום הוכחנו ש- $\xi(K) = \bigcup_{\eta\in I_K} D(\eta)$ לכל $\infty > K \geq 0$.

(ב) כבר ראינו ל-$K > 0$ כי $\psi(K) = \inf_{\eta\in[\eta_m,\eta_M]} \{\varepsilon(\eta) + K\eta\}$ ושהאינפימום מתקבל בכל $\eta \in I_K$ שהוא אינטרוול לא ריק.

עבור $K = 0$, לאור למה 2 סעיף (ד), הרי $\varepsilon_m = \psi(0)$ ניתן גם כן לרישום בצורה דלעיל. בנוסף אם האינפימום מתקבל ב-η_0 כלשהו הרי $\varepsilon(\eta_0) \le \varepsilon(\eta_0-\omega)-K\omega$, לכל $(\eta_0-\eta_M) \le \omega < \infty$. אך עבור $\omega \le (\eta_0-\eta_M)$, הרי $\varepsilon(\eta_M)-K(\eta_0-\eta_M) \le \varepsilon(\eta_0-\omega)-K\omega$ (על פי הגדרת η_M בלמה 2). לפיכך בהכרח אם האינפימום מתקבל ב-η_0 הרי $K \in \partial\Phi_{\eta_0}(0)$ ולכן $\eta_0 \in I_K$ (לאור T15 והגדרת I_K). מכאן שעבור $K > 0$ I_K הוא בדיוק האינטרוול שבו מתקבל המינימום של $\varepsilon(\eta) + K\eta$ ב-$[\eta_m,\eta_M]$.

מאחר ו- $\varepsilon(\eta) + K\eta$ היא פונקציה רציפה ב-$[\eta_m,\eta_M]$ לכל $\infty > K \geq 0$ (על פי משפט 1, סעיף (א)), הרי לאור T5, I_K הוא אינטרוול סגור.

(ג) כאשר $f(\underline{a})$ קמורה-ממש הרי $K\eta + \varepsilon(\eta)$ היא פונקציה קמורה-ממש לכל K ב-$[\eta_m,\eta_M]$ שהיא קבוצה קמורה (על פי משפט 1 סעיף (ד)), ולאור T13 המינימום שלה (שמושג על פי סעיף (ב) דלעיל) מושג בנקודה יחידה שהיא בדיוק I_K עבור $K > 0$, ובנקודה זו $D(\eta)$ מכילה וקטור מקדמים יחיד אופטימלי (על פי משפט 1 סעיף (ד)) שהוא $\underline{a}^*(K)$. עבור המקרה $K = 0$, הרי לאור T13 המינימום של $f(\underline{a})$ ב-$\mathbb{C}^{M_a}$, (המושג על פי למה 2 סעיף (א)) מושג בוקטור יחיד $\underline{a}^*$ שלו מתאימה $\eta_M = g(\underline{a}^*)$. מכאן ש- $D(\eta) = \phi$ עבור $\eta_M < \eta$ (כי $\varepsilon(\eta) = \varepsilon_m$) ולכן $\xi(o)$ מכילה וקטור יחיד זה.

<u>הוכחת המשמעות:</u>

ההכלה השמאלית נובעת מסעיף (א) של המשפט. ההכלה הימנית וטענת השיוויון נובעות מסעיף (א) ומהעובדה שלכל $\eta_m < \eta$ (וגם עבור $\eta = \eta_m$ כאשר $-\infty < \varepsilon'_+(\eta_m)$) הקבוצה $\partial\Phi_\eta(0)$ אינה ריקה, כפי שהראינו במבוא להוכחה, ולכן קיים כופל לגרנז' K^* ל- P_{ϕ_η} על פי T17. הפרשנות של $(-K)$ כשפוע $\varepsilon(\eta)$ נובעת מהגדרת I_K בסעיפים (א) ו-(ב) וכנ"ל המקרים הפרטיים השונים שמתוארים. העקום $\eta(K)$ הוא קשיר ומונוטוני לא-עולה לאור סעיף (א) דלעיל וסעיף (ב) של משפט 1.

<u>הוכחת משפט 3:</u>

I. $||\cdot||_i$ על SD_i, ניתנת לזהוי עם סמי-נורמה $||\cdot||_i^*$ על $\mathbb{C}^{(M_i+1)}$.
$||\cdot||_c$ על SD_c ניתנת לזהוי עם סמי-נורמה $||\cdot||_c^*$ על $\mathbb{C}^{(M_c+1)}$.

בשני המקרים הללו פשוט נזהה את הקבוצות הפורשות $\{d_i,-v_{i1},\dots,-v_{iM_i}\}$ ו-$\{d_c,-v_{c1},\dots,-v_{cM_c}\}$ עם הבסיס הסטנדרטי של וקטורי יחידה ב-$\mathbb{C}^{(M_i+1)}$ וב-$\mathbb{C}^{(M_c+1)}$. עתה נשכן את $\{\mathbb{C}^{(M_i+1)}\}_{i=1}^N$ כתת מרחבים בתוך $\mathbb{C}^{(M_a+N)}$ על ידי שרשור N הוקטורים.

לכל וקטור ב-$\underline{a} \in \mathbb{C}^{(M_a+N)}$ הרי:

$$\delta_c = \tilde{g}(\underline{a}) = || [-\frac{1}{A\underline{a}}-] ||_c^*$$

כאשר המטריצה A מכילה N עמודות אפסים (בהתאמה לאברים d_i ב-SD_i-ים).

בנוסף: $|| \, ||\underline{a}_i||_i^* ||_T = ||\underline{\delta}||_T \overset{\Delta}{=} \tilde{f}(\underline{a}) = \varepsilon$. אך $||\cdot||_T$ ו-$||\cdot||_i^*$ ליניאריים ביחס לסקלריים, ולכן גם $\tilde{f}(\underline{a})$ ליניארית ביחס לסקלרים מ-$\mathbb{C}$. בנוסף $||\cdot||_i^*$ ו-$||\cdot||_T$ מקיימים את אי-שיוויון המשולש ו-$||\cdot||_T$ מונוטונית ביחס לוקטורים ב-$[0,\infty)^N$ ולכן $\tilde{f}(\underline{a})$ מקיימת את אי-שיוויון המשולש, לפיכך $||\underline{a}||^* \overset{\Delta}{=} \tilde{f}(\underline{a})$ היא סמי-נורמה על $\mathbb{C}^{(M_a+N)}$. מכאן התוצאה שאת הבעיה היסודית שהגדרנו ניתן לשכן בתוך:

$$\varepsilon(\eta) = \inf_{\substack{\left|\left|\frac{1}{A\underline{a}}\right|\right|_c^* \le \eta \\ a_1=\dots=a_N=1}} \{||\underline{a}||^*\}$$

כש: $||\cdot||^*$ היא סמי-נורמה על $\mathbb{C}^{(M_a+N)}$. $A_{M_c\times(M_a+N)}$ היא מטריצה נתונה.
$||\cdot||_c^*$ היא סמי-נורמה על $\mathbb{C}^{(M_c+1)}$.

II . על מנת לזהות את הבעיה האקויולנטית לבעיתנו, נמצא את כל התנאים על A, $||\cdot||_c^*$ ו-$||\cdot||^*$.

(א) A מכילה N עמודות ראשונות שהן עמודות אפסים.

(ב) $||\cdot||_c^*$ אינה מתאפסת על תת-המרחב ממימד M_c שנוצר כשהאיבר הראשון בוקטור מאופס, למעט בוקטור האפס.

(ג) $||\cdot||^*$ אינה מתאפסת על תת-המרחב ממימד M_a הנוצר כאשר N האברים הראשונים מאופסים, למעט בוקטור האפס.

(ד) $||\underline{a}||^*$ היא פונקציה של הוקטור $\underline{\delta} \in [0,\infty)^N$ הנוצר על ידי חישוב $||\cdot||^*$ על N תתי-מרחב מתאימים ממימד (M_i+1) הנוצרים על ידי איפוס כל אברי $\underline{a}$ שלא מזוהים עם SD_i $(1 \le i \le N$-ל$)$, וזו פונקציה מונוטונית ביחס לכל קומפוננטה של $\underline{\delta}$.

תנאים (א) - (ג) ברורים. תנאי (ד) מאפשר לזהות את $||\cdot||^*$ על תתי-המרחב המתאימים עם $||\cdot||_i^*$ (שכן זו פונקציה ליניארית ביחס לסקלר ב-$\mathbb{C}$ ומקיימת את אי-שיוויון המשולש שם), כשהפונקציה $t(\underline{\delta}) = ||\underline{a}||^*$ מחליפה את $||\cdot||_T$. בברור $t(\cdot)$ ליניארית ביחס לסקלר מ-$[0,\infty)$, על פי הליניאריות של $||\cdot||^*$, ואינה מתאפסת פרט ל-$\underline{\delta} = \underline{0}$ על פי תנאי (ג) על $||\underline{a}||^*$. היא מקיימת את אי-שיוויון המשולש ב-$[0,\infty)^N$

לאור העובדה ש-$||\cdot||^*$ מקיימת את אי-שיוויון המשולש (הוכחה - על ידי בחירת וקטורים $\underline{a}$ ו-$\underline{b}$ כך שבתת המרחב ה-i-י:

$$\left.\begin{array}{l} \frac{\delta_{2_i}}{\delta_{1_i}} \underline{a}_i = \underline{b}_i \\ ||\underline{a}_i||^* = \delta_{1_i} \end{array}\right\} \Leftarrow t(\underline{\delta}_1+\underline{\delta}_2) = ||\underline{b}+\underline{a}||^* \geq ||\underline{b}||^* + ||\underline{a}||^* = t(\underline{\delta}_2) + t(\underline{\delta}_1).$$

ומונוטונית ביחס לוקטורים ב-$[0,\infty)^N$ על פי תנאי (ד).

<u>הוכחת משפט 4:</u>

כפי שמוסבר בהמשך המשפט הרי עבור $v \in SD_i$ המיוצג על ידי $v = \alpha d_i + \sum_{k=1}^{M_i} a_{ik}v_{ik}$,

הרי $\sigma(v) = \bar{\alpha}d_i + \sum_{k=1}^{M_i} \bar{a}_{ik}v_{ik}$.

לפיכך על פי נתון 5,

$$(\forall\, \underline{a} \in \mathbb{C}^{M_a}) \quad ||d_i - \sum_{k=1}^{M_i} a_{ik}v_{ik}||_i = ||d_i - \sum_{k=1}^{M_i} \bar{a}_{ik}v_{ik}||_i$$

ולכן $f(\underline{a}) = f(\bar{\underline{a}})$. בדומה לכך לכל $v \in SD_c$ המיוצג על ידי

$$v = \alpha d_c + \sum_{i=1}^{N} \beta_i \sum_{k=1}^{M_i} a_{ik}v_{ik}$$

הרי

$$\sigma(v) = \bar{\alpha}d_c + \sum_{i=1}^{N} \bar{\beta}_i \sum_{k=1}^{M_i} \bar{a}_{ik}v_{ik}$$

לפיכך על פי נתון 4:

$$(\forall\, \underline{a} \in \mathbb{C}^{M_a}) \quad ||d_c - \sum_{i=1}^{N} \beta_i \sum_{k=1}^{M_i} a_{ik}v_{ik}||_c = ||d_c - \sum_{i=1}^{N} \bar{\beta}_i \sum_{k=1}^{M_i} \bar{a}_{ik}v_{ik}||_c$$

ומאחר וכל ה-$\{\beta_i\}_{i=1}^{N}$ ממשיים הרי ש-$g(\underline{a}) = g(\bar{\underline{a}})$.

מכאן שחמשת התנאים של משפט 4 מבטיחים:

$$\begin{cases} f(\underline{a}) = f(\bar{\underline{a}}) \\ g(\underline{a}) = g(\bar{\underline{a}}) \end{cases}$$

מכאן ואילך נוכיח את המשפט על סמך שתי עובדות אלו בלבד. לכן לכל מרחב וקטורי אחר, מערכת תנאים שתבטיח אותן, תגרור קיומו של וקטור $\underline{a}$ ממשי בכל קבוצה $D(\eta)$ $(\xi(K))$. יהא $\underline{a} \in D(\eta)$ (קיים וקטור כלשהו כזה לכל $\eta_m \geq \eta$). אזי מהאמור לעיל ומהגדרת $D(\eta)$ גם $\bar{\underline{a}} \in D(\eta)$. מאחר ו-$D(\eta)$ קבוצה קמורה (על פי למה 2) הרי גם $\frac{1}{2}\underline{a} + \frac{1}{2}\bar{\underline{a}} \in D(\eta)$.

אך $\mathrm{Re}\,\underline{a} = \frac{1}{2}\underline{a} + \frac{1}{2}\bar{\underline{a}}$ הוא וקטור ממשי ולכן לבכל $D(\eta)$ קיים וקטור ממשי. התוצאה עבור הקבוצות $\xi(K)$ נובעת מיידית מכאן, עקב משפט 2 סעיף (ג).

<u>הוכחת משפט 5:</u>

נגדיר לכל $\underline{a} \in ₵^{M_a}$ את $\underline{a}^o, \underline{a}^e \in ₵^{M_a}$ על ידי:

$$\begin{cases} a^e_{ik} = (a_{ik} + \bar{a}_{i\pi_i(k)})/2 \\ a^o_{ik} = (a_{ik} - \bar{a}_{i\pi_i(k)})/2 \end{cases}$$

ולכן $\underline{a}^e + \underline{a}^o = \underline{a}$ וכן:

$$\begin{cases} a^e_{ik} = \bar{a}^e_{i\pi_i(k)} \\ a^o_{ik} = -\bar{a}^o_{i\pi_i(k)} \end{cases}$$

בצורה דומה להעתקות $u_i \in S_i$ המתאימות לוקטור המקדמים $\underline{a}_i$ נגדיר את u^e_i, u^o_i כהעתקות ב-S_i המתאימות ל-$\underline{a}^o_i$, $\underline{a}^e_i$ בהתאמה.

מתכונה 3 הנתונה נובע ש-

$$\bar{u}^e_i(p)\, e^{j\psi(p)} = u^e_i(p) e^{-j\psi(p)}$$

וכן

$$\bar{u}^o_i(p) e^{j\psi(p)} = -u^o_i(p) e^{-j\psi(p)}$$

$(\forall p \in \Omega)$, $1 \leq i \leq N$.

לכן $u^e_i(p) e^{-j\psi(p)} \in R$ ו- $u^o_i(p) e^{-j\psi(p)} \in I$ $(I \triangleq \{z | z = jy, y \in R\})$.

עתה:

$$|d_i(p) - u_i(p)| = |d_i(p) e^{-j\psi(p)} - u_i(p) e^{-j\psi(p)}|$$

מאחר ש-$u_i(p) = u_i^e(p) + u_i^o(p)$ ועל פי תכונה 2 הנתונה, נקבל:

$$(A3) \quad |d_i(p) - u_i(p)| = |\hat{d}_i(p) - u_i^e(p)e^{-j\psi(p)} - u_i^o(p)e^{-j\psi(p)}| \underset{\uparrow}{=}$$

פירוק לחלק ממשי ומדומה

$$= \{|\hat{d}_i(p) - u_i^e(p)e^{-j\psi(p)}|^2 + |u_i^o(p)e^{-j\psi(p)}|^2\}^{1/2} \underset{\uparrow}{=}$$

שוב תכונה 2 הנתונה

$$= \{|d_i(p) - u_o^e(p)|^2 + |u_i^o(p)|^2\}^{1/2} \geq |d_i(p) - u_i^e(p)|$$

היות ואי-שיוויון זה מתקיים לכל $1 \leq i \leq N$ ולכל $p \in \Omega$, הרי מתכונה 5 הנתונה, ומהמונוטוניות של $||\cdot||_T$ נובע ש-$f(\underline{a}) \geq f(\underline{a}^e)$.

באופן דומה יהא $u_c \in S_c$ ההעתקה המתאימה לוקטור המקדמים $\underline{a}$ ויהיו u_c^e ו-u_c^o ההעתקות המתאימות ל-$\underline{a}^e$, $\underline{a}^o$ בהתאמה, אזי מהתוצאות שקיבלנו לעיל:

$$(A4) \quad u_c^e(p)e^{-j\psi(p)} = \sum_{i=1}^{N} \beta_i u_i^e(p)e^{-j\psi(p)} = \sum_{i=1}^{N} \beta_i \bar{u}_i^e(p)e^{+j\psi(p)} = \bar{u}_c^e(p)e^{j\psi(p)} \in R$$

$$(A5) \quad u_c^o(p)e^{-j\psi(p)} = \sum_{i=1}^{N} \beta_i u_i^o(p)e^{-j\psi(p)} = -\sum_{i=1}^{N} \beta_i \bar{u}_i^o(p)e^{j\psi(p)} = -\bar{u}_c^o(p)e^{j\psi(p)} \in I$$

כאשר ניצלנו את הנתון ש-$\beta_i = \bar{\beta}_i$, $1 \leq i \leq N$. עתה באופן אנלוגי לאי-השוויון (A3), נקבל על ידי שמוש בתכונה 1, ב-(A4), (A5) ובכך ש- $u_c = u_c^e + u_c^o$:

$$(A6) \quad |d_c(p) - u_c(p)| = |\hat{d}_c(p) - u_c^e(p)e^{-j\psi(p)} - u_c^o e^{-j\psi(p)}| =$$

$$= \{|\hat{d}_c(p) - u_c^e(p)e^{-j\psi(p)}|^2 + |u_c^o(p)e^{-j\psi(p)}|^2\}^{1/2} =$$

$$= \{|d_c(p) - u_c^e(p)|^2 + |u_c^o(p)|^2\}^{1/2} \geq |d_c(p) - u_c^e(p)|$$

מ-(A6) ומתכונה 4 הנתונה הרי $||d_c - u_c||_c \geq ||d_c - u_c^e||_c$ ולכן: $g(\underline{a}) \geq g(\underline{a}^e)$.

יהא $\underline{a} \in D(\eta)$ (קיים כזה לכל $\eta \geq \eta_m$ כי $D(\eta)$ אינו ריק). אזי קיים עבורו $\underline{a}^e$ כך ש- $f(\underline{a}^e) \leq \varepsilon(\eta)$ ו- $g(\underline{a}^e) \leq \eta$, ועל פי הגדרת $\varepsilon(\eta)$ בלמה 2, $\underline{a}^e \in D(\eta)$.

נגדיר:
$$\hat{u}_i^e(p) \triangleq u_i^e(p)e^{-j\psi(p)} \in R$$

אזי $\hat{u}_i^e \in \hat{U}$ ו- $u_i^e(p) = \hat{u}_i^e(p)e^{j\psi(p)}$ $(\forall\, p \in \Omega)$ ולכל $1 \leq i \leq N$. מכאן שב-$D(\eta)$ ישנו וקטור מקדמים $\underline{a}$ שהעתקותיו u_i מקיימות $u_i(p) = \hat{u}_i(p)e^{j\psi(p)}$ כש- $\hat{u}_i \in \hat{U}$ וזו בדיוק טענת המשפט. המקרה של $\xi(K)$ נובע מכאן מידית על פי משפט 2 סעיף (ג).

הוכחת משפט 6:

יהא $\underline{a} \in D(\eta)$ ויהיו $\{u_i\}_{i=1}^N$ ההעתקות המתאימות לוקטור זה.

נתבונן בהעתקות $\{\hat{u}_i\}_{i=1}^N$ מהצורה: $\hat{u}_i = \frac{1}{N}\sum_{k=1}^N \Omega_i\Omega_k^{-1}(u_k)$. היות ו-$\omega_k \in S_\Omega$ הרי ω_k^{-1} קיים ולכן גם Ω_k^{-1} קיים. לפיכך (מהסגירות ביחס להרכב, לפעולות לינאריות וביחס ל-Ω_i, Ω_k^{-1} של U) הרי $\hat{u}_i \in U$ מוגדר היטב.

על פי תכונה 3 של הנתון הרי $S_o \supseteq \Omega_k^{-1}(u_k)$ ועל פי אותה תכונה $\hat{u}_i \in S_i$ (נזכור שכל ה-S_i כולל כמובן S_o הם תת-מרחבים לינאריים של U). מכאן שלהעתקות $\{\hat{u}_i\}_{i=1}^N$ מתאים וקטור מקדמים כלשהו $\hat{\underline{a}} \in ¢^{M_a}$. בנוסף $\Omega_i \in L(U,U)$ (אלו אופרטורים לינאריים מ-U לעצמו) ולכן ניתן להציג את $\hat{u}_i$ כדלקמן: $\hat{u}_i = \Omega_i(\hat{u}_o)$ כש-$\hat{u}_o = \frac{1}{N}\sum_{k=1}^N \Omega_k^{-1}(u_k)$ ולכן (מתכונה 3 הנתונה) $\hat{u}_o \in S_o$. לכן הוקטור $\hat{\underline{a}}$ בעל התכונה המצוטטת במשפט ולא נותר אלא להוכיח ש-$\hat{\underline{a}} \in D(\eta)$.

תחילה נחשב את $\hat{\delta}_i$ המתאימים לו:

$$\hat{\delta}_i = ||d_i-u_i||_i \underset{(1)}{=} ||\Omega_i(d_o-\hat{u}_o)||_i \underset{(2)}{=} ||d_o-\hat{u}_o||_o \underset{(3)}{\le} \frac{1}{N}\sum_{k=1}^N ||d_o - \Omega_k^{-1}(u_k)|| \underset{(4)}{=}$$

$$= \frac{1}{N}\sum_{k=1}^N ||\Omega_k^{-1}(d_k-u_k)||_o \underset{(5)}{=} \frac{1}{N}\sum_{k=1}^N ||d_k-u_k||_k = \frac{1}{N}\sum_{k=1}^N \delta_k$$

כאשר (1) דלעיל נובע מתכונה 2 הנתונה ולינאריות Ω_i, (2) מתכונה 6 הנתונה, (3) מאי-שיוויון המשולש והגדרת $\hat{u}_o$, (4) שוב מתכונה 2 ולינאריות Ω_k^{-1} ו-(5) מתכונה 6 הנתונה.

לפיכך $\hat{\underline{\delta}} \le \frac{1}{N}\sum_{k=1}^N \underline{\delta}^{(k)}$ כש-$\underline{\delta}^{(k)}$ זהו הוקטור הנוצר מ-$\underline{\delta}$ על ידי סיבוב ציקלי ב-k אברים. על פי המונוטוניות של $||\cdot||_T$, אי-שיוויון המשולש עבורה ותכונה 7 הנתונה:

$$f(\hat{\underline{a}}) = ||\hat{\underline{\delta}}||_T \le ||\frac{1}{N}\sum_{k=1}^N \underline{\delta}^{(k)}||_T \le \frac{1}{N}\sum_{k=1}^N ||\underline{\delta}^{(k)}||_T = ||\underline{\delta}||_T = f(\underline{a})$$

עתה נחשב את $\hat{\delta}_c$ המתאים ל-$\hat{\underline{a}}$:

$$g(\hat{\underline{a}}) = \hat{\delta}_c \underset{(1)}{=} ||d_c - \sum_{i=1}^N \hat{u}_i||_c \underset{(2)}{=} ||d_c - \frac{1}{N}\sum_{k=1}^N \sum_{i=1}^N \Theta_i\Theta_k^{-1}(u_k)||_c \underset{(3)}{=}$$

$$= ||d_c - \frac{1}{N}\sum_{j=1}^N \Theta_j\ (\sum_{k=1}^N u_k)||_c \underset{(4)}{=} ||\ \frac{1}{N}\sum_{j=1}^N \Theta_j\ (d_c-u_c)\ ||_c \underset{(5)}{\le}$$

$$\le \frac{1}{N}\sum_{j=1}^N ||\Theta_j(d_c-u_c)||_c \underset{(6)}{=} \frac{1}{N}\sum_{j=1}^N ||d_c-u_c||_c = \delta_c = g(\underline{a})$$

כאשר (1) דלעיל נובע מתכונה 4 הנתונה, (2) מלינאריות Ω_i והגדרתן, (3) מטענת העזר שתוכח בהמשך ולינאריות Θ_j, (4) מהגדרת u_c ותכונה 1 הנתונה, (5) מאי-שיוויון המשולש, ו-(6) מתכונה 5 הנתונה.

קיבלנו לכן ש- $g(\hat{\underline{a}}) \le g(\underline{a})$ ו- $f(\hat{\underline{a}}) \le f(\underline{a})$, וראינו כבר שאם $\underline{a} \in D(\eta)$ תכונות אלו מחייבות $\hat{\underline{a}} \in D(\eta)$ (הוכחת משפט 5). סיימנו בזאת את המשפט כי הטיעון ל-$\xi(K)$ מיידי (ראה שם). נותר לנו להוכיח את טענת העזר הבאה:

<u>טענת עזר</u>:

עבור כל תת-חבורה סופית מסדר N של L(U,U) $\{\Theta_i\}_{i=1}^{N}$ הרי: $\sum_{i=1}^{N}\Theta_i\,\Theta_k^{-1} = \sum_{j=1}^{N}\Theta_j$.

ההוכחה טריויאלית:

מאחר ו- $\{\Theta_i\}_{i=1}^{N}$ זו תת-חבורה הרי היא סגורה ביחס להפכי ולהרכב לכן $\Theta_i\,\Theta_k^{-1}$ הוא איבר מהחבורה. בנוסף אם $\Theta_{i_1}\Theta_k^{-1} = \Theta_{i_2}\Theta_k^{-1}$ הרי $\Theta_{i_1} = \Theta_{i_2}$, קל להראות שבמקרה זה סדר החבורה קטן מ- N ולכן זו סתירה לנתון. מכאן ש- $\{\Theta_i\Theta_k^{-1}\}_{i=1}^{N}$ הם כל אברי החבורה ולכן בודאי: $\sum_{i=1}^{N}\Theta_i\Theta_k^{-1} = \sum_{j=1}^{N}\Theta_j$.

<u>הערה</u>:

מטענת העזר גם ברור מדוע תכונה 5 מוגדרת היטב, קרי $\Theta_j(SD_c) = SD_c$. מכיון ש- $\Theta_j(d_c) = d_c$ על פי תכונה 1 ו- Θ_j לינאריות, הרי כל שצריך לאמת הוא ש- $S_c = \Theta_j(S_c)$ אך מכיון ש- $u \in S_c$ <= $u_i \in S_i$, $u = \sum_{i=1}^{N} u_i$ <= $u = \sum_{i=1}^{N} u_i$, $u_i \in \Theta_i(S_1)$ על פי תכונה 3 <= $\Theta_j(u) = \sum_{i=1}^{N}\Theta_j u_i$ $u_i \in \Theta_i(S_1)$ <= מאחר ו- Θ_i הפיכים <= $\Theta_j(u) = \sum_{i=1}^{N}\Theta_j\Theta_i(\tilde{u}_i)$, $\tilde{u}_i \in S_1$ <= מאחר ו- Θ_i הם אברי חבורה הרי - $\Theta_j(u) = \sum_{i=1}^{N}\tilde{\tilde{u}}_i$, $\tilde{\tilde{u}}_i \in \Theta_{\pi(i)}(S_1)$ כש-$\pi(\cdot)$ זו פרמוטציה כלשהי <= על פי הגדרת S_c ותכונה 3, הרי $\Theta_j(u) \in S_c$, לכן לבסוף $\Theta_j(S_c) \subset S_c$ והשיוויון נובע מכך ש- Θ_j^{-1} קיים ונמצא בחבורה ולכן גם $\Theta_j^{-1}(S_c) \subset S_c$.

I. Conditions for the Optimality of Prototype translated Filter Banks

We begin with introducing the design problem in a general framework, which is applicable to almost any design criteria, and for a generalized structure of the filter bank. Although this complicates the presentation, it enables to obtain results for both the WMMSE and the Min-Max criteria.

Problem Statement:

(A). The filter bank is composed of N individual digital filters. The i-th filter is a linear combination of M_i basic components having the frequency responses $E_{ik}(f)$ $k=1, \ldots, M_i$ (for a conventional FIR filter, the basic components are delays, and thus $E_{ik}(f)$ takes the form of $E_{ik}(f) = e^{-j2\pi(k+l_i)f}$). Since all the results derived in this section are for a complex filter bank, the coefficients of the linear combinations (denoted by a_{ik}, $k=1, \ldots, M_i$), are assumed to be complex numbers. The conditions for realness of the optimal coefficients and phase linearity of the filters for a wide class of design criteria are presented elsewhere. With the above definition of its components, the frequency response of the i-th filter is:

$$H_i(f) \triangleq \sum_{k=0}^{M_i-1} a_{ik} E_{ik}(f) \qquad (1)$$

(B). The *desired* frequency response of the i-th filter is denoted by $D_i(f), i=1, \ldots, N$. The error between this desired frequency response and the frequency response of the corresponding filter is measured by a semi-norm (denoted by $\| \ \|_i$) which is defined on the space of complex functions with real argument which are periodic with a periodicity of 1. Thus, the i-th filter response error is defined as:

$$\delta_i \triangleq \|D_i(f) - H_i(f)\|_i \qquad (2)$$

For example, in the WMMSE design: $\delta_i \triangleq [\int_{-0.5}^{0.5} | W_i(f)(D_i(f) - H_i(f))|^2 df]^{1/2}$, whereas in the Min-Max design: $\delta_i \triangleq \sup_{-0.5 \le f \le 0.5} \{| W_i(f)(D_i(f) - H_i(f))|\}$. Since $\| \ \|_i$ is only a semi-norm there may exist non-zero functions having a zero-norm. However, the design problem is well posed only if $\|H_i(f)\|_i > 0$ for every filter defined by (1) with a non-zero set of coefficients.

(C). The composite response of the filter bank (denoted by $H_c(f)$ is the sum of the responses of all N filters, i.e.: $H_c(f) \triangleq \sum_{i=1}^{N} H_i(f)$. The specification on the composite response is given by an allowed tolerance η with respect to the desired composite frequency response $D_c(f)$. The error between the desired composite response and the composite response of the filter bank is measured by another semi-norm (denoted by $\| \ \|_c$) and therefore:

$$\delta_c \triangleq \|D_c(f) - H_c(f)\|_c \qquad (3)$$

Thus, the tolerance specification is $\delta_c \le \eta$, and for example, a flat composite response is specified by $D_c(f) = 1$. In both the WMMSE and Min-Max designs $\| \ \|_c$ usually has the same form as $\| \ \|_i$, but uses a different weight function (denoted by $W_c(f)$).

(D). The overall performance of the filter bank is measured by yet another semi-norm (denoted by $\| \ \|_T$), which takes into account the response errors of all the filters. Let $\underline{\delta} \in R_+^N$ be the vector, whose i-th component is δ_i - the error of the i-th filter, then the overall error is defined as:

$$\varepsilon \triangleq \|\underline{\delta}\|_T \qquad (4)$$

Where $\| \ \|_T$ is defined on R_+^N, the space of N dimensional real vectors with non-

negative components. For example, in the WMMSE design: $\|\underline{\delta}\|_T = [\sum_{i=1}^{N} K_i^2\delta_i^2]^{1/2}$, whereas in the Min-Max design either: $\|\underline{\delta}\|_T = \underset{1\le i\le N}{Max}\{|K_i|\delta_i\}$ or $\|\underline{\delta}\|_T = \sum_{i=1}^{N} |K_i|\delta_i$, with K_i being appropriate weighting factors.

(E). The design of an optimal filter bank with a specified composite response is therefore the solution of the following optimization problem:

$$\underset{\{a_{ik}\}_{i=1,k=1}^{N\ M_i}|\delta_c\le\eta}{Min} \{\varepsilon\} \tag{5}$$

The existence of a solution (not necessarily a unique one) to this design problem is guaranteed for large enough η, provided that the following three assumptions hold:

(AS1). The overall error is defined such that an increase in the errors of some of the individual filters never decreases the overall error, i.e., for every $\underline{\delta}$, $\underline{\Delta\delta}\in\mathbb{R}_+^N$, $\|\underline{\delta} + \underline{\Delta\delta}\|_T \ge \|\underline{\delta}\|_T$.

(AS2). For every filter defined by (1) with a non-zero set of coefficients, $\|H_i(f)\|_i > 0$ for $i=1, \ldots, N$.

(AS3). Since some of the basic components of different filters might be identical, let the composite response of the filter bank be rewritten as a linear combination of P *linearly independent basic components:*

$$H_c(f) = \sum_{k=0}^{P-1} b_k E_{ck}(f). \tag{6}$$

Now, the assumption is that for every composite response defined by (6) with at least one of the b_k-s which is non-zero, $\|H_c(f)\|_c > 0$.

If the design problem is well-posed, then these assumptions hold. For example, both the WMMSE and Min-Max designs of conventional FIR filter banks obey these assumptions as long as each one of the weight functions is non-zero on a

set of positive measure. The proof of this existence theorem is beyond the scope of this paper.

(F). Definition: A Prototype Translated Filter Bank (PTFB), is a filter bank with the property: $H_i(f - \Delta - \frac{i}{N}) = H_o(f), i=1, \ldots, N$, for some value of Δ and some function $H_o(f)$.

A Complex Uniform Filter Bank (CUFB) is characterized by:

$$D_i(f - \Delta - \frac{i}{N}) = D_o(f);\ i=1, \ldots, N \tag{7a}$$

$$D_c(f - \frac{i}{N}) = D_c(f);\ i=1, \ldots, N \tag{7b}$$

and the error criterion satisfies

$$\|A(f - \Delta - \frac{i}{N})\|_o = \|A(f)\|_i; \text{for any function } A(f) \text{ and every } 1 \le i \le N \tag{7c}$$

$$\|A(f - \frac{i}{N})\|_c = \|A(f)\|_c; \text{ for any function } A(f) \tag{7d}$$

$$\|\underline{\delta}^{(r)}\|_T = \|\underline{\delta}\|_T; \text{ for any } \underline{\delta} \in R_+^N \tag{7e}$$

where $\underline{\delta}^{(r)}, r=1, \ldots, N$, denotes the vector obtained by a cyclic shift of the components of the vector $\underline{\delta}$, i.e.: $\delta_i^{(r)} = \delta_{(i+r) mod N}$. The interpretation of (7a)-(7e) is that the individual filters specifications are translated versions of a prototype specification, the composite response specifications are invariant under frequency translation, and each one of the filters has the same contribution to the error measure.

The following theorem relates any CUFB design problem with an optimal solution which is a PTFB.

Theorem 1: Let the basic components of the filters have the following property: For any frequency response of the i-th filter $H_i(f)$ defined by (1), there exists a set of coefficients of the $(i-1)modN$-th filter having the frequency response $H_i(f - \frac{1}{N})$.

Then, any CUFB design problem has at least one optimal solution which is a PTFB.

Proof: Due to assumptions (AS1)-(AS3) there exists at least one optimal solution, denoted by $\{H_i(f)\}_{i=1}^{N}$. Define: $\hat{H}_i(f) \triangleq \frac{1}{N}\sum_{k=1}^{N} H_k(f - \frac{(k-i)}{N})$, for $i = 1, \ldots, N$. It is easy to verify that the theorem's condition implies the existence of a set of coefficients $\{a_{ik}\}$ for which $\hat{H}_i(f)$ is the frequency response of the i-th filter in the filter bank. Furthermore, it is easily verified that this filter bank is a PTFB. Using the triangle inequality, and homogenity of $\| \; \|_i$ together with (7a) and (7c) results in $\hat{\delta}_i \triangleq \|D_i(f) - \hat{H}_i(f)\|_i \leq \frac{1}{N}\sum_{k=1}^{N}\|D_o(f) - H_k(f - \Delta - \frac{k}{N})\|_o$. Using (7a) and (7c) with respect to $\| \; \|_k$ transfers the above inequality into: $\hat{\delta}_i \leq \frac{1}{N}\sum_{k=1}^{N} \delta_k$. Following assumption (AS1), the triangle inequality and homogenity of $\| \; \|_T$ results in the following inequality: $\hat{\varepsilon} \triangleq \|\hat{\underline{\delta}}\|_T \leq \|\frac{1}{N}\sum_{k=1}^{N} \underline{\delta}^{(k)}\|_T \leq \frac{1}{N}\sum_{k=1}^{N}\|\underline{\delta}^{(k)}\|_T$. However, from (7e) this is equivalent to $\hat{\varepsilon} \leq \varepsilon$. Thus, the PTFB achieves the minimal error, and it only remains to show that $\hat{\delta}_c \triangleq \|D_c(f) - \hat{H}_c(f)\|_c \leq \delta_c$. Rearranging the expression for the composite response of the PTBF as $\hat{H}_c(f) = \frac{1}{N}\sum_{j=1}^{N} H_c(f - \frac{j}{N})$, using the triangle inequality and homogenity of $\| \; \|_c$ results in: $\hat{\delta}_c \leq \frac{1}{N}\sum_{j=1}^{N}\|D_c(f) - H_c(f - \frac{j}{N})\|_c$. From (7b) and (7d) follows that $\hat{\delta}_c \leq \delta_c$, and the proof is completed.

The following corollary restates the condition of Theorem 1 for the conventional structure of filter banks:

Corollary 1: For conventional FIR filter banks (i.e., with delays as basic components) any CUFB design problem has at least one PTFB optimal solution, provided that all N individual filters have the same length (i.e., $M_i = M$) and the same delay.

When the conditions in this corollary hold, the optimal PTFB can be obtained as the solution of a much simplier design problem as suggested by the following theorem:

Theorem 2: Under the conditions of Corollary 1 the optimal PTFB is obtained as follows:

(a). Solve the following FIR filter design problem:

$$\underset{\{a_{ok}\}_{k=0}^{M-1}}{Min} \| D_o(f) - \sum_{k=0}^{M-1} a_{ok} e^{-j2\pi f(k+l)} \|_o \tag{8a}$$

Subject to the following constraint imposed by the composite response specification:

$$\delta_c \triangleq \| D_c(f) - N \sum_{m=\lceil \frac{l}{N} \rceil}^{\lfloor \frac{M-1+l}{N} \rfloor} a_{o(mN-l)} e^{-j2\pi(f+\Delta)mN} \|_c \leq \eta \tag{8b}$$

(b). The coefficients of the PTFB filters are obtained from the optimal prototype filter (the solution of (8)) by:

$$a_{ik} \triangleq a_{ok} e^{-j2\pi(\frac{i}{N}+\Delta)(k+l)} \qquad i=1,\ldots,N;\ k=0,\ldots,M-1. \tag{9}$$

Proof: From the definition of the PTFB, and the structure of conventional filter banks implied by Corollary 1, equation (9) follows immediately. From (2), (7a) and (7c) it follows that in any PTFB solution for a CUFB design problem all the N components of $\underline{\delta}$ are equal. Thus, due to assumption (AS1) the optimal prototype minimizes each component of $\underline{\delta}$, i.e. it is the

solution of (8a). Equation (8b) follows from the definition of the PTFB, equation (3), and the well-known result: $\sum_{i=0}^{N-1} e^{-j2\pi \frac{ik}{N}} = \begin{cases} N & k \equiv 0(mod N) \\ 0 & \text{elsewhere} \end{cases}$.

Since Δ in the PTFB and CUFB definitions is arbitrary, its value can always be chosen so that the prototype filter is a lowpass filter.

For any filter bank structure, for which Theorem 1 holds, the analogue of Theorem 2 can be easily derived. Since in the sequel we concentrate on the conventional FIR structure, we omit the derivation.

II. WMMSE Design of Uniform Filter Banks.

The WMMSE design of filter banks with specified composite response has been presented in [61]. Here we define the WMMSE CUFB design problem, and as a consequence of Theorem 2, obtain a simplified algorithm for the design of the optimal (in the WMMSE sense) PTFB solution.

The semi-norms $\| \|_i$, $\| \|_c$ and $\| \|_T$ corresponding to the WMMSE design criteria have already been defined as examples in the discussion of the previous section. The definition of the WMMSE CUFB in terms of the WMMSE design specifications is therefore:

$$D_i(f - \Delta - \frac{i}{N}) = D_o(f) \qquad 1 \le i \le N \tag{10a}$$

$$D_c(f - \frac{i}{N}) = D_c(f) \qquad 1 \le i \le N \tag{10b}$$

$$W_i(f - \Delta - \frac{i}{N}) = W_o(f) \qquad 1 \le i \le N \tag{10c}$$

$$W_c(f - \frac{i}{N}) = W_c(f) \qquad 1 \le i \le N \tag{10d}$$

$$K_i = 1 \quad 1 \le i \le N \tag{10e}$$

And the optimal prototype filter is the solution of:

$$\min_{\{a_{ok}\}_{k=0}^{M-1}} \varepsilon^2 \triangleq N \int_{-0.5}^{0.5} |W_o(f)|^2 |D_o(f) - \sum_{k=0}^{M-1} a_{ok} e^{-j2\pi f(k+l)}|^2 df \tag{11a}$$

Subject to:

$$\delta_c^2 \triangleq \int_{-0.5}^{0.5} |W_c(f)|^2 |D_c(f) - N \sum_{m=\lceil \frac{l}{N} \rceil}^{\lfloor \frac{M-1+l}{N} \rfloor} a_{o(mN-l)} e^{-j2\pi(f+\Delta)mN}|^2 df \le \eta^2 \tag{11b}$$

This convex programming problem is equivalent to [56, sec. 4.5]:

$$\min_{\{a_{ok}\}_{k=0}^{M-1}} \{\varepsilon^2 + K^2 \delta_c^2\} \tag{12}$$

For some value of $K(\eta)$. Differentiating (12) with respect to the unknown variables and rearranging the resulting equations leads to the solution:

$$\underline{a}_o = R_o^{-1}[\underline{d}_o + H^T \underline{q}] \tag{13a}$$

Where $\underline{a}_o \in \mathbb{C}^M$ is the vector whose elements are the coefficients of the prototype filter. The elements of the P.D. Hermetian $M \times M$ matrix R_o are:

$$R_o(m,k) = \int_{-0.5}^{0.5} |W_o(f)|^2 e^{j2\pi f(m-k)} df \quad ; m,k = 0, \ldots, M-1 \tag{13b}$$

And the elements of $\underline{d}_o \in \mathbb{C}^M$ are:

$$d_o(m) = \int_{-0.5}^{0.5} |W_o(f)|^2 D_o(f) e^{j2\pi f(m+l)} df \quad ; m,k = 0, \ldots, M-1 \tag{13c}$$

The elements of the $Q \times M$ matrix H (with $Q = \lfloor \frac{M-1+l}{N} \rfloor - \lceil \frac{l}{N} \rceil + 1$) are:

$$H(m,k) = \begin{cases} 1 & k = mN + (\lceil \frac{l}{N} \rceil N - l) \\ 0 & \text{elsewhere} \end{cases} \quad ; m = 0, \ldots, Q-1, k = 0, \ldots, M-1 \tag{13d}$$

The vector $\underline{q} \in \mathbb{C}^Q$ is a correction vector due to the composite response specification and is defined as:

$$\underline{q} = \left[\frac{1}{K^2} R_c^{-1} + NHR_o^{-1}H^T\right]^{-1} (R_c^{-1}\underline{d}_c - NHR_o^{-1}\underline{d}_o) \tag{13e}$$

Where the elements of the P.D. Hermitian $Q \times Q$ matrix R_c are:

$$R_c(m,k) = \int_{-0.5}^{0.5} |W_c(f - \Delta)|^2 e^{j2\pi f(m-k)} df \quad ; m,k = 0, \ldots, M-1 \tag{13f}$$

And the elements of $\underline{d}_c \in \mathbb{C}^Q$ are:

$$d_c(m) = \int_{-0.5}^{0.5} |W_c(f - \Delta)|^2 D_c(f - \Delta) e^{j2\pi f(m + \lceil \frac{l}{N} \rceil)N} df \; ; m = 0, \ldots, Q-1 \tag{13g}$$

Equations (13a)-(13g) describe the optimal prototype given $K^2(\eta^2)$ and are the simplified version of (10)-(15) in [61]. The overall error (ε^2) and the composite response error (δ_c^2) of the optimal PTFB are:

$$\varepsilon^2 = N\left[\int_{-0.5}^{0.5} |W_o(f)D_o(f)|^2 df - \underline{d}_o^* R_o^{-1}\underline{d}_o + \underline{q}^* H^T R_o^{-1} H\underline{q}\right] \tag{14a}$$

$$\delta_c^2 = \int_{-0.5}^{0.5} |W_c(f)D_c(f)|^2 df - \underline{d}_c^* R_c^{-1}\underline{d}_c + \underline{q}^* R_c^{-1}\underline{q} \tag{14b}$$

where $\underline{v}^*$ denotes conjugate transposition of $\underline{v}$. The value of $K(\eta)$ is obtained, similar to [61, equations (16)-(24)], by simultaneously diagonalization of R_c^{-1} and $T \triangleq NHR_o^{-1}H^T$. Similar to [61], the optimal prototype has real coefficients provided that $|W_o(f)|^2$, $|W_c(f - \Delta)|^2$, $D_o(f)$ and $D_c(f - \Delta)$ are Fourier

transforms of real sequences. Furthermore, if both $D_o(f)$ and $D_c(f-\Delta)$ have the linear phase $e^{-j2\pi f(l+(M-1)/2)}$, and $2l+(M-1) \equiv 0 mod N$, the optimal prototype has a linear phase [61]. However, phase linearity of the original specifications (i.e., $D_i(f)$) imposes $l = -(\frac{M-1}{2})$, i.e., both $D_o(f)$ and $D_c(f-\Delta)$ have zero phase.

Usually $|W_o(f)|^2$, $|W_c(f-\Delta)|^2$, $D_o(f)$ and $D_c(f-\Delta)$ are piecewise constant functions, and all the integrals appearing in (13) and (14) can be evaluated analytically. Thus, the complexity of the design is $O(M^2+\alpha(M/N)^3)$. $O(M^2)$ operations are needed for the inversion of the M dimensional Toeplitz matrix R_o, and for obtaining the solution $\underline{a}_o$ via (12a). $O(\alpha(M/N)^3)$ operations are needed for the simultaneous diagonalization of the $Q \times Q$ symmetric matrices R_c^{-1} and $NHR_o^{-1}H^T$, where $Q \cong M/N$, and $\alpha > 1$ represents the relative complexity of unitary diagonalization of a matrix compared to its inversion. In comparison, the original algorithm of [61] involves inversion of N Toeplitz matrices, then a simultaneous diagonalization of two $M \times M$ matrices, and thus its complexity is $O(NM^2+\alpha M^3)$. For example, the design of a typical digital filter bank with 32 filters, each of them is an FIR filter of length 256 takes a few CPU seconds on a 16 bit computer using the new algorithm whereas a direct design requires the solution of an optimization problem with 8192 variables !

III. Real WMMSE Uniform Filter Banks

In many applications, real outputs are desired while preserving the efficient implementation of the PTFB. This is done by adding together proper pair of outputs of the PTFB. The following result is an immediate consequence of equation (9):

Lemma 1: Adding pairs of outputs of a PTFB with a real prototype filter having a zero-phase response, results in a *real* filter bank if and only if $\Delta \equiv 0 mod(\frac{1}{2N})$.

Two different structures of real filter banks can thus be obtained from the same real prototype filter. The first one (denoted as structure (A)) corresponds to $\Delta = 0$ with the i-th and $(N-i)$-th outputs being added, yielding $\lceil \frac{N+1}{2} \rceil$ real filters; whereas in the second structure (denoted by (B)) $\Delta = \frac{1}{2N}$ and the i-th and $(N-1-i)$-th outputs are added, yielding $\lceil \frac{N}{2} \rceil$ real filters. The main difference between these two structures is in the bandwidth of the first (lowpass) filter in the resulting real filter bank.

For both structures the design specifications are no-longer those of a CUFB.

A RUFB, is a filter bank characterized by (7b), (7d), (7e) and:

$$D_i(f) = C_i[D_o(f + \frac{i}{2N}) + D_o(f - \frac{i}{2N})] \tag{15a}$$

$$A(f) = \overline{A(-f)} \Rightarrow \|A(f + \frac{i}{2N})\|_i = \|A(f - \frac{i}{2N})\|_i \tag{15b}$$

for any function $A(f)$

Where $i = 0,2,4,\ldots,$ for structure (A), and $i = 1,3,5,\ldots,$ for structure (B), and $C_i \triangleq 1$ except for $C_o = C_N \triangleq 0.5$.

Real filter banks obtained from a PTFB have the following property:

$$H_i(f) = C_i[H_o(f + \frac{i}{2N}) + H_o(f - \frac{i}{2N})] \tag{16}$$

Where the values of i in (16) are determined by the structure used in combining pairs of outputs of the PTFB. In general the optimal solution of the RUFB design problem does not possess the property implied by (16). Thus, unlike the CUFB case, in general any PTFB provides only a *sub-optimal* solution to the RUFB

design problem. However, one can impose equation (16) as part of the design specifications to guarantee an efficient implementation via a PTFB. In this context the question of finding the optimal prototype filter arises naturally. We concentrate therefore on the WMMSE RUFB design subject to equation (16). Specifically we use the following weight functions:

$$|W_i(f)|^2 = \begin{cases} 1 & |f - \frac{i}{2N}| \in [0, F_p] \text{ or } |f + \frac{i}{2N}| \in [0, F_p]; \; F_p \le \frac{1}{2N} \\ 0 & |f - \frac{i}{2N}| \in (F_p, F_s] \text{ or } |f + \frac{i}{2N}| \in (F_p, F_s]; \; F_s \le \frac{1}{N} \\ w & \text{elsewhere} \end{cases} \tag{17}$$

[1] [2]

Fig.4.1a illustrates a typical weight function and Fig.4.1b illustrates the equivalent low-pass weight function $|W_o(f)|^2$ used in the design of the optimal prototype of the corresponding CUFB. The composite response specification imposed on the prototype filter of the WMMSE RUFB design is given by (11b) as in the CUFB case. The overall error of the WMMSE RUFB design, which is obtained by substituting (7),(16),(15a), (10e), in (1),(2),(4) is for structure (A):

$$\varepsilon_A^2 = \begin{cases} \varepsilon_{cx}^2 + N\psi_N & ; \; N \text{ odd} \\ \varepsilon_{cx}^2 + N\psi_{\frac{N}{2}} & ; \; N \text{ even} \end{cases} \tag{18a}$$

and for structure (B):

$$\varepsilon_B^2 = \begin{cases} \varepsilon_{cx}^2 + N\psi_N & ; \; N \text{ odd} \\ \varepsilon_{cx}^2 + N(2\psi_N - \psi_{\frac{N}{2}}) & ; \; N \text{ even} \end{cases} \tag{18b}$$

Where ε_{cx}^2 is the overall error of the WMMSE CUFB design given in (11a), and ψ_N is the following correction term:

[1] For $F_s > \frac{1}{2N}$ this definition becomes inaccurate for $i=1$ (the first filter in structure (B)). To accommodate for this particular case in the design process equations (17)-(20) have to be slightly modified. However, this modification complicates the presentation and therefore is omitted here.

[2] $W_i(f)$ is periodic with a period of 1, and in that context we interpret the word "elsewhere".

$$\psi_N \triangleq \frac{1}{N}\sum_{i=1}^{N-1}\Bigg\{\int_{-0.5}^{0.5}(|W_o(f)|^2 - w)|D_o(f + \frac{i}{N}) - H_o(f + \frac{i}{N})|^2 df$$

$$+ \int_{-0.5}^{0.5}(2|W_o(f)|^2 - w)\mathrm{Re}[(\overline{D_o(f)} - \overline{H_o(f)})(D_o(f + \frac{i}{N}) - H_o(f + \frac{i}{N}))]df\Bigg\} \quad (18c)$$

where 'superbar' denotes complex conjugation. Due to the existence of a non-zero correction term ψ_N, ε_A^2 and ε_B^2 differ from the overall error ε_{cx}^2 of the WMMSE CUFB. Thus, the optimal prototype for the WMMSE RUFB design is *different* from the optimal prototype for the WMMSE CUFB design. Furthermore, for an even number of filters, the optimal prototype for structure (A) is not the optimal prototype for structure (B). For optimization purposes the correction term ψ_N is rewritten as the following quadratic form:

$$\psi_N = \frac{1}{2}\underline{a}_o^*(R_N^* + R_N)\underline{a}_o - \underline{d}_N^*\underline{a}_o - \underline{a}_o^*\underline{d}_N + \psi_o \quad (19)$$

where ψ_o is a constant (independent of a_o, and the elements of the $M\times M$ matrix R_N are:

$$R_N(m,k) = -\frac{1}{N}(3R_o(m,k) - 2w) + (R_o(m,k) - w)\delta(m \equiv k \bmod N)$$

$$+ (2R_o(m,k) - w)\delta(k \equiv l \bmod N) \quad (20a)$$

and the elements of $\underline{d}_N \in \mathbb{C}^M$ (assuming that $D_o(f) = 0$ for $|f| > \frac{1}{2N}$) are:

$$d_N(m) = -\frac{1}{N}d_o(m) + d_o(m)\delta(m \equiv l \bmod N) \quad ; m=0,\ldots,M-1 \quad (20b)$$

Therefore, the design of the optimal prototype for the WMMSE RUFB is via equations (13a)-(13g) with R_o and $\underline{d}_o$ being replaced by the following expressions, respectively:

(a). $R_o + \frac{1}{2}(R_N + R_N^*)$; $\underline{d}_o + \underline{d}_N$ for N *odd* .

(b). $R_o + \frac{1}{2}(R_{\frac{N}{2}} + R_{\frac{N}{2}}^*)$; $\underline{d}_o + \underline{d}_{\frac{N}{2}}$ for N *even* and *structure* (A).

(c). $R_o + R_N + R_N^* - \frac{1}{2}(R_{\frac{N}{2}} + R_{\frac{N}{2}}^*)$ $\underline{d}_o + 2\underline{d}_N - \underline{d}_{\frac{N}{2}}$ for N *even* and *structure* (B).

The issues of phase-linearity, and design complexity are similar to the $CUFB_o$ case.

IV. Min-Max Design of Uniform Filter Banks.

Optimal *Min-Max* design of filter banks with specified composite response involves linear programming techniques. The linear program for the i-th filter is of M_i unknown variables and P_i linear constraints, where the continuous frequency response error of the i-th filter is uniformly sampled with a sampling interval of $\frac{1}{P_i}$ [25,26]. For the whole filter bank design, the overall linear programming effort involves $(\sum_{i=1}^{N} M_i + 1)$ unknown variables and $\sum_{i=1}^{N+1} P_i$ linear constraints including the composite response specifications. We shall concentrate on conventional filter banks with delays as basic components, and assume that all the individual filters have the same length (i.e., $M_i = M$). Furthermore, it is assumed that the sampling interval is the same for all $(N+1)$ frequency responses (i.e. $P_i = P$). Each iteration of the linear programming algorithm involves $(N+1)$ FFT's of dimension M each, followed by interpolations by a factor of P/M [26]. This step which has a complexity of $O((N+1)M\log_2 M + (N+1)\alpha P)$ operations per iteration determines the complexity of the algorithm, where α is the number of operations per output sample of the interpolation filter. For typical values of $N = 32$, $M = 256$ and $P = 1024$ this algorithm is quite complex.

A Min-Max CUFB is defined by design specifications that satisfy (10) using the weighted L_∞ norm for $\| \|_i$ and $\| \|_c$. As a particular instance of Corollary 1

and Theorem 2, the optimal Min-Max CUFB is a PTFB whose prototype filter is the solution of:

$$\underset{\{a_{ok}\}_{k=0}^{M-1}}{Min} \underset{f\in[-0.5,0.5]}{Sup} \{ | W_o(f) || D_o(f) - \sum_{k=0}^{M-1} a_{ok} e^{-j2\pi f(k+l)} | \} \tag{21a}$$

Subject to:

$$\underset{f\in[-0.5,0.5]}{Sup} \{ | W_c(f) || D_c(f) - N \sum_{m=\lceil \frac{l}{N} \rceil}^{\lfloor \frac{M-1+l}{N} \rfloor} a_{o(mN-l)} e^{-j2\pi(f+\Delta)mN} | \} \le \eta \tag{21b}$$

This modified design problem involves only M variables and $2P$ constraints, thus the complexity of the design is reduced by at least a factor of N. The optimal prototype filter has real coefficients and linear-phase provided that the conditions presented in section II are fulfilled (i.e., $l = -\frac{(M-1)}{2}$, $D_o(f)$ and $D_c(f-\Delta)$ are zero-phase functions, and $|W_o(f)|$, $|W_c(f-\Delta)|$, $D_o(f)$ and $D_c(f-\Delta)$ are Fourier transforms of real sequences).

Whereas the designs of the optimal prototypes for the WMMSE RUFB and WMMSE CUFB have the same complexity, the design of the optimal prototype for the Min-Max RUFB involves the evaluation of $(N+1)$ different frequency responses and is thus more complex than the design of the Min-Max CUFB. Nevertheless, equations (15)-(17) can be used to reduce the complexity of the linear programming algorithm in the design of a RUFB as well.

It can be shown that for $|W_c(f)|$ and $|W_o(f)|$ which are positive on a set of non-zero measure, there exists a solution of (21a)-(21b), which is obtained with equality in (21b) provided that $\eta \ge \eta_m$. Where the minimal composite response error η_m is the solution of:

$$\eta_m = \underset{\{b_m\}_{m=0}^{Q-1}}{Min} \{ \underset{f\in[-0.5,0.5]}{Sup} | W_c(f-\Delta) || D_c(f-\Delta) - N \sum_{m=0}^{Q-1} b_m e^{-j2\pi f N(m+\lceil \frac{l}{N} \rceil)} | \} \tag{22}$$

With $Q \triangleq \lfloor \frac{M-1+l}{N} \rfloor - \lceil \frac{l}{N} \rceil + 1$.

The design of the optimal Min-Max prototype for $\eta = \eta_m$ can thus be done in two steps:

(a). Solve (22) to obtain η_m and $\{b_m\}_{m=0}^{Q-1}$.

(b). Substitute $a_{o(mN+\lceil \frac{l}{N} \rceil N - l)} = b_m$ and solve (21a) to complete the impulse response of the prototype filter.

Each one of these steps involves a linear programming solution of a constrained FIR digital filter design problem, similar to the one presented in [40]. For the particular case of a flat composite response (i.e., $D_c(f) = 1$), the unique solution of (22) is $b_m = \delta(m + \lceil \frac{l}{N} \rceil)$, with $\eta_m = 0$. Thus, the optimal prototype for a *specified flat composite response* with zero tolerance is the N-th band FIR filter designed in [40]. The linear programming approach is quite satisfactory for short length prototype designs (typically, for $M \leq 100$). However, for values of M which are several hundreds this approach becomes impractical.

V. Approximate Optimal Window Design

The following theorem relates the optimal Min-Max CUFB design with the well-known Window Method for the design of FIR filter banks.

Theorem 3:

(a). For any CUFB specifications the Window Method results in a PTFB.

(b). If $D_c(f) = \sum_{i=0}^{N-1} D_i(f)$ then in the Window Method the composite response is $H_c(f) = D_c(f) * W(f)$ with $W(f)$ being the frequency response of the window sequence.

(c). For $D_0(f)$ which is an ideal LPF of bandwidth $\frac{1}{2N}$ and a flat composite response specification ($D_c(f) = 1$), the optimal Min-Max CUFB for zero tolerance ($\eta = \eta_m = 0$) can be obtained with the Window method.

Proof: The proof of *part (b)* of this theorem is given in [37], and this is essentially the motivation for using the Window Method in filter bank designs. The proof of *part (a)* is: In the Window method $H_i(f) = D_i(f)*W(f)$. Using (10a) and elementary properties of the convolution operator leads to $H_i(f) = [D_o*W](f+\frac{i}{N}+\Delta) \triangleq H_o(f+\frac{i}{N}+\Delta)$.

The proof of *part (c)* is: Due to part (a) the Window Method results in a PTFB. Due to part (b) this PTFB has the composite response $H_c(f) = 1*W(f) = w_o$. Thus, for any window with $w_o = 1$ the Window Method results in a PTFB with $\eta = \eta_m = 0$. For $D_o(f)$ which is an ideal LPF of bandwidth $\frac{1}{2N}$, the coefficients of the corresponding impulse response $d_o(k)$ are zero *only* for $k \equiv 0(modN)$. Therefore, any PTFB with a flat composite response can be designed by the Window method, provided that a proper window sequence is used. In particular the optimal PTFB can be obtained by the Window method.

Due to this theorem the Optimal Window (in the Min-Max sense) leads to the solution of the Min-Max CUFB design problem. Since for $D_o(f)$ which is an ideal LPF of Bandwidth $\frac{1}{2N}$ any window sequence results in a flat composite response, the Optimal Window is any solution of:

$$\underset{\{w_k\}_{k=-|l|}^{|l|}}{Min} \; \underset{f\in[-0.5,0.5]}{Sup} \{ | W_o(f) | | D_o(f) - \sum_{k=-|l|}^{|l|} d_o(k) w_k e^{-j2\pi fk} | \} \tag{23}$$

Where we assumed that $l = -(\frac{M-1}{2})$ to assure phase linearity of all the filters in the optimal PTFB, and $\{d_o(k)\}_{k=-|l|}^{|l|}$ are the coefficients of the desired (infinite) impulse response, i.e.:

$$d_o(k) \triangleq \int_{-0.5}^{0.5} D_o(f) e^{j2\pi fk} df = \begin{cases} \frac{1}{\pi k}\sin(\frac{\pi k}{N}) & k \neq 0 \\ \frac{1}{N} & k = 0 \end{cases} \tag{24}$$

For $|W_o(f)| = |W_o(-f)|$ there is always an Optimal Window with zero-phase, real coefficients, and satisfying $w_o = 1$. Restricting the solution of the design problem to have these properties, the Optimal Window is the Solution of:

$$\underset{\{w_k\}_{k=1}^{L}}{Min} \left\{ \delta_w \triangleq \underset{f \in [0,0.5]}{Sup} |W_o(f)||D_o(f) - \frac{1}{N} - \sum_{k=1}^{L} 2\frac{w_k}{\pi k}\sin(\frac{\pi k}{N})\cos(2\pi f k)| \right\} \quad (25)$$

This problem can be solved by linear programming as in [40]. It seemed at first that this is a Chebyshev approximation problem and thus the Remez exchange method can be applied. This is not true, since the Chebyshev approximation problem with unknown coefficients $a_{ok} \triangleq w_k d_o(k)$ results in general with $a_{o(mN)} \neq 0$, contradicting the constraints $d_o(mN) = 0$. For this reason the method presented in [39] is sub-optimal. Similar to [21], we represent the error induced in the Window Method as follows:

$$\delta_w = Max\{ \underset{f \in [0,\frac{1}{2N}]}{Sup} |W_o(f)||J_w(\frac{1}{2N}-f) + J_w(\frac{1}{2N}+f)|,$$

$$\underset{f \in [\frac{1}{2N},0.5]}{Sup} |W_o(f)||J_w(f-\frac{1}{2N}) + J_w(1-\frac{1}{2N}-f)|\} \quad (26a)$$

Where:

$$J_w(f) = \int_f^{0.5}\{1 + \sum_{k=1}^{L} 2w_k\cos(2\pi\vartheta k)\}d\vartheta = (0.5-f) - \sum_{k=1}^{L}\frac{w_k}{\pi k}\sin(2\pi f k) \quad (26b)$$

In [22] the following approximation of δ_w (denoted by $\tilde{\delta}_w$) is used:

$$\tilde{\delta}_w \triangleq Max\{ \sup_{f \in [0, \frac{1}{2N}]} |W_0(f)| Max[|J_w(\frac{1}{2N} - f)|, |J_w(\frac{1}{2N} + f)|],$$

$$\sup_{f \in [\frac{1}{2N}, 0.5]} |W_0(f)| Max[|J_w(f - \frac{1}{2N})|, |J_w(1 - \frac{1}{2N} - f)|]\} \quad (27)$$

It is easily verified that $\delta_w \leq 2\tilde{\delta}_w$ for any window sequence. Furthermore, in [21,22] this approximation was used for most of the known window sequences and for many design examples, always leading $\delta_w \geq \tilde{\delta}_w$. Let the approximate optimal window (AOW) denote the sequence with minimal value of $\tilde{\delta}_W$. We therefore expect the error of the PTFB designed using the AOW sequence to be within a factor of two from the minimal error. Combining (26b) and (27) and rearranging the resulting expression, the AOW is the solution of:

$$\min_{\{w_k\}_{k=1}^{[L]}} \{ \sup_{\vartheta \in [0, 0.5]} [|\tilde{W}(\vartheta)| |0.5 - \vartheta - \sum_{k=1}^{[L]} \frac{w_k}{\pi k} \sin(2\pi\vartheta k)|]\} \quad (28a)$$

Where:

$$|\tilde{W}(\vartheta)| \triangleq Max\{|W_0(\frac{1}{2N} - \vartheta)|, |W_0(\frac{1}{2N} + \vartheta)|, |W_0(\vartheta - \frac{1}{2N})|, |W_0(1 - \frac{1}{2N} - \vartheta)|\} \quad (28b)$$

And in (28b) we interpret $|W_0(\lambda)|$ as zero for $\lambda < 0$ or $\lambda > 0.5$. Unlike the original problem stated in (25) this is a Chebyshev approximation problem, thus the Remez exchange method can be applied, and the AOW is easily obtained even for filter lengths of several hundreds. Solving (28) with the Remez exchange method appears to require a special program. However, the available program in [27] is suitable for this purpose by applying the following algorithm:

(a). Design an optimal (in the Min-Max sense) *differentiator* of length M for the weight function $|\tilde{W}(0.5-\vartheta)|$ using [27] with an absolute error criterion.

(b). Let $\{a_{|l|},\ldots, a_1, 0, -a_1,\ldots,-a_{|l|}\}$ be the coefficients of this optimal differentiator, then the desired window sequence is given by $w_k \triangleq 2\pi k(-1)^{k+1}a_k,\ k = 1,\ldots,|l|$.

The AOW was defined for the CUFB with an odd length prototype filter. For CUFB with an even length prototype filter, Theorem 3 holds, although (21b) is inaccurate since now l is not an integer. The AOW is defined as the argmin of δ_w defined in (27) and for even length sequences it is the solution of:

$$\underset{\{w_k\}_{k=1}^{M/2}}{Min}\left\{\underset{\vartheta\in[0,0.5]}{Sup}\left[\,|\hat{W}(\vartheta)|\left|0.5-\sum_{k=1}^{\frac{M}{2}}\frac{2w_k}{\pi(2k-1)}\sin\pi\vartheta(2k-1)\right|\,\right]\right\} \tag{29}$$

With $|\hat{W}(\vartheta)|$ defined by (28b). The expression above results from the analogue of (26b) for even length sequences. The available program in [27] can be used to solve (29) using the following algorithm:

(a). Design an optimal (in the Min-Max sense) *Hilbert transformer* of length M for the weight function $|\hat{W}(\vartheta)|$ using the program in [27].

(b). Let $\{a_{\frac{M}{2}},\ldots, a_1, -a_1,\ldots,-a_{\frac{M}{2}}\}$ be the coefficients of this filter, then the desired window sequence is given by $w_k \triangleq \pi(k-\frac{1}{2})a_k,\ k=1,\ldots,\frac{M}{2}$.

For a general (Non-Uniform) filter bank with specified flat composite response the AOW can be defined in a similar manner, where the weight function $|\hat{W}(\vartheta)|$ given in (28b) is properly changed. The AOW results in an approximation to the optimal filter bank among those designed by the window method. Therefore it usually leads to a superior performance as compared with conventional window sequences. However, for Non-Uniform filter banks the AOW design is not necessarily an approximation to the optimal (in the Min-Max sense) filter bank. For this reason we concentrated on the CUFB case although the AOW may be useful in the more general case as well.

We compare the design using the AOW sequence with the optimal Min-Max prototype via the design example 1 in [40]. This example corresponds to the design of a CUFB composed of eight FIR filters, each one of them has 39 taps. The passband of the prototype is $f \in [0, 0.10625]$, whereas its stopband is $f \in [0.14375, 0.5]$. The desired passband deviation equals the desired stopband deviation (i.e., $W_o(f)$ is as in Fig.1b with $w=1$). The optimal Min-Max prototype has passband ripple of $0.35dB$, and stopband attenuation of $33dB$. The AOW design results in passband ripple of $0.43dB$ and stopband attenuation of $32dB$. Thus, the degradation due to the sub-optimality of the AOW design is very small. Similar results have been obtained for other design examples. For longer filters (typically above one hundred taps), the optimal Min-Max design is too complex. Thus, we compare the design using the AOW sequence with various sub-optimal methods, namely: The conventional window method in [37] using the Kaiser Window (denoted by KW), the sub-optimal Min-Max design of [39] (denoted by TMX), and the un-specified composite response design in [27] (denoted by DMX). The comparison is via a CUFB design example. Since TMX is applicable only to odd length filters we consider a bank of $N = 16$ filters, having each an impulse response of length $M = 123$ samples. A flat composite response is specified, i.e. $D_c(f) = 1$, and the desired prototype frequency response is that of an ideal LPF of bandwidth $\frac{1}{32}$ (i.e., $D_o(f) = 1$ for $|f| \le \frac{1}{32}$ and zero elsewhere). All four design methods are applied to design a PTFB. For these specifications KW, AOW and TMX guarantee a flat composite response ($\delta_c = 0$). The prototype weight function is:

$$|W_o(f)| \triangleq \begin{cases} 1 & |f| \le F_p \\ 0 & F_p < |f| < F_s \\ K & F_s \le |f| \le 0.5 \end{cases} \qquad F_p < \frac{1}{32} < F_s \tag{30}$$

The transition bandwidth is $F_s - F_p \triangleq \frac{0.55}{32}$. For KW, DMX and TMX $(F_s + F_p)/2 = \frac{1}{32}$ whereas for the AOW $(F_s + F_p)/2$ is optimally set according to [21]. Table B-1 summarizes the performance obtained for three typical values of K = 1,10,50. As expected, the DMX design has the smallest passband ripple, and largest stopband attenuation. However, it results in a poor composite response, with a ripple of up to 8.73db for K = 50. For K = 1 all four methods results in a similar performance, whereas for $K \gg 1$ the AOW is certainly preferred on TMX and KW. Since TMX is applicable only for $(N-1) \geq K \geq 1$ [39], it is not used for K = 50. Figures 4.2 to 4.5 illustrate the shape of the response of the prototype filter which results in the DMX, TMX, KW and AOW designs, respectively, for K = 10. In Fig.4.2 the composite response of the DMX filter bank is also illustrated.

It is obvious in Fig.4.3 that the low stopband attenuation of the TMX results due to the spurious peaks in the frequency bands around $\frac{1}{32} + \frac{m}{16}$, $m = 1, \ldots, 14$. If these bands are excluded from the stopband by proper change of $|W_0(f)|$ similar performance to the AOW is achieved. Figures 4.4 and 4.5 illustrate the shape of the filters designed using KW and AOW, respectively. Note that the AOW results in a nearly equiripple response.

Table C-1: CUFB Design Example.

	KW [37]		AOW		TMX [39]		DMX [27]		
	A_p	A_s	A_p	A_s	A_p	A_s	A_p	A_s	A_c
$K=1$	0.22	38.10	0.25	38.45	0.14	41.57	0.14	41.68	0.03
$K=10$	0.22	38.10	1.10	46.68	1.03	30.25	0.54	50.08	4.61
$K=50$	0.22	38.10	3.09	51.22	-	-	1.08	58.18	8.73

A_p - Passband Ripple in dB.
A_s - Stopband Attenuation in dB.
A_c - Composite Response Ripple in dB.

I. Statistical Model of the A/S System

We present below a statistical model of a fixed time reference [6,13] A/S system with WOLA synthesis and quantization of the output of the analysis stage. Although most of the A/S systems contain the DFT transform in the analysis stage so that the quantizer input is the Discrete Short Time Fourier Transform (DSTFT), of the input signal [5], some contain other linear regular transforms (such as the DCT, or Hadamard transform). Therefore, the modeling of an A/S system with an arbitrary linear regular transform is presented, in order to apply the new design method to various coding systems. A schematic description of the A/S system is given in Figs. 5.1, 5.2. In order to simplify the presentation of the new design method, we frequently use infinite sums, so as not to bother with explicitly stating their limits, although they are actually finite unless explicitly stated. The time domain sequences as well as the analysis and synthesis windows have real values, whereas the transform values may be complex. A detailed description of the model is given below.

(A). The input signal sequence $x(n)$ is multiplied by the sliding analysis window $h(\cdot)$ having a length of L_h samples. A time aliasing operation reduces these L_h values into a vector of length M (the transform size), denoted by $\underline{x}_s$. These vectors (for different time instances) are decimated with a decimation factor R, $1 \le R \le M$. The m-th element of the vector $\underline{x}_{sR}$ is given by:

$$x_{sR}(m) = \sum_{r=-\infty}^{\infty} h(sR-m-Mr)x(m+Mr), \quad 0 \le m \le M-1 \tag{1}$$

A linear regular transform of size M operates on these vectors and results in output vectors $\underline{X}_{sR}$ of M elements each. This transform is represented by the $M \times M$ dimensional matrix T, whose (k,m) element is denoted by $t(k,m)$, $0 \le k,m \le M-1$. The k-th element of the vector $\underline{X}_{sR}$ is given by:

$$X_{sR}(k) = \sum_{m=0}^{M-1} t(k,m)\, x_{sR}(m) \qquad 0 \le k \le M-1 \tag{2}$$

If the DFT transform is applied (i.e., $t(k,m) = \exp(-j\frac{2\pi km}{M})$), the sequence of vectors $\underline{X}_{sR}$ is the DSTFT of the input signal. For the general transform represented by T we denote the sequence of vectors $\underline{X}_{sR}$ as the discrete short-time transform (DSTT) of the input signal. This completes the analysis part of the system.

(B). Let the modified DSTT (MDSTT) sequence of vectors be denoted by $\hat{\underline{X}}_{sR}$. Before describing in detail the various quantizers considered (i.e., the various types of mappings of $\hat{\underline{X}}_{sR}(\underline{X}_{sR})$), we describe the WOLA synthesis which is used in all the systems considered here.

An inverse transform operates on the MDSTT and results in time domain vectors $\hat{\underline{x}}_{sR}$ of M elements each. The inverse transform is represented by the matrix T^{-1} of dimensions $M \times M$, whose (m,k) element is denoted by $t^{-1}(m,k)$ for $0 \le m,k \le (M-1)$. The m-th element of $\hat{\underline{x}}_{sR}$ is thus given by:

$$\hat{x}_{sR}(m) = \sum_{k=0}^{M-1} t^{-1}(m,k)\, \hat{X}_{sR}(k) \quad , \quad 0 \le m \le M-1 \tag{3}$$

The sequence of time domain vectors $\hat{\underline{x}}_{sR}$ is interpolated with an interpolation factor R, and a weighted overlap-add operation reconstructs the output sequence $y(n)$. This operation is done using a synthesis filter $f(\cdot)$ of L_f samples, and is described by:

$$y(n) = \sum_{s=-\infty}^{\infty} f(n-sR)\hat{x}_{sR}((n)_M) \tag{4}$$

where $(n)_M$ denotes $n(\text{mod } M)$.

(C). Two types of quantization approaches are considered:

1. Fine Quantization (FQ)

When fine quantization is used, there are typically about 5 or more bits per input sample and the output signal to noise ratio (SNR) is quite high. It is well-known that for this case the quantization error can be reasonably modeled as an additive noise vector that is uncorrelated with the input signal [66]. We shall further assume that these noise vectors are samples of a wide-sense stationary process with zero-mean (typically, assumed to be uncorrelated), and known covariance sequence. Let $\underline{V}_{sR}$ denote the noise vector added to $\underline{X}_{sR}$, then:

$$\hat{X}_{sR}(k) = V_{sR}(k) + X_{sR}(k) \quad , \quad 0 \le k \le M-1 \tag{5}$$

For convenience of the presentation we denote by $\underline{v}_{sR}$ the inverse transform of the vector $\underline{V}_{sR}$ and use the covariance sequence of $\underline{v}_{sR}$ rather than the covariance sequence of $\underline{V}_{sR}$. Thus,

$$\hat{x}_{sR}(m) = x_{sR}(m) + v_{sR}(m) \quad , \quad 0 \le m \le M-1 \tag{6}$$

and:

$$E[v_{sR}(m)] = E[v_{sR}(m)\, x(n)] = 0 \qquad \forall\, n,s; \;\; 0 \le m \le M-1 \tag{7a}$$

$$E[v_{sR}(m) v_{(s+d)R}(n)] = \Psi_{m,n}(dR) \qquad \forall\, s,d; \;\; 0 \le m,n \le M-1 \tag{7b}$$

For example, if the vectors $\underline{V}_{sR}$ are uncorrelated both in time and in the transform domain, then $\Psi_{m,n}(dR) = 0$ for $d \neq 0$. Furthermore, if the DFT transform is used, the matrix whose elements are $\Psi_{m,n}(0)$ is a Circulant matrix whose first row is the IDFT of dimension M of the sequence $(E[V_{sR}(0)^2], \ldots, E[V_{sR}(M-1)^2])$. The latter sequence is closely related to the bit allocation used in the different bands.

2. Matrix Quantization (MQ)

The matrix quantization of blocks of $B \geq 1$ DSTT vectors is explained below:

The vectors $\{\underline{X}_{sBR}, \ldots, \underline{X}_{(sB+(B-1))R}\}$, are regarded as an $M \times B$ matrix $\mathbf{X}_{s(BR)}$. The code book contains L different matrices $\{C^{(1)}, \ldots, C^{(L)}\}$ and the space of all $M \times B$ complex matrices is divided into L distinct sets $\{A^{(1)}, \ldots, A^{(L)}\}$, such that $\bigcup_{l=1}^{L} A^{(l)} = \mathbb{C}^{M \times B}$. In the time instance s, the coder selects the index $1 \leq i_s \leq L$, according to:

$$i_s = l \qquad \text{iff} \qquad \mathbf{X}_{s(BR)} \in A^{(l)} \quad 1 \leq l \leq L \tag{8a}$$

When a specific index value l is received by the decoder, it uses the matrix $C^{(l)}$ as $\hat{\mathbf{X}}_{s(BR)}$ and therefore the sequence $\{\underline{c}_0^{(l)}, \ldots, \underline{c}_{B-1}^{(l)}\} = T^{-1}C^{(l)}$ is used as the sequence of vectors $\{\hat{\underline{x}}_{sBR}, \ldots, \hat{\underline{x}}_{(sB+(B-1))R}\}$ in the synthesis. This sequence is assumed to be real even for complex transforms (one can easily guarantee this property by an appropriate selection of the code book matrices). The output samples of the A/S system are the result of a WOLA synthesis (using (4)), applied to the concatenated sequence:

$$\hat{\underline{x}}_{(sB+d)R} = \underline{c}_d^{(i_s)} \quad , \quad 0 \leq d \leq B-1, \;\; 1 \leq i_s \leq L, \;\; -\infty < s < \infty \tag{8b}$$

It is quite difficult to design a complete codebook for large values of L (e.g., 1024 and above), and therefore product codes are usually used. In a product code the index i_s is essentially a vector of length $J \geq 1$, i.e.: $i_s = (i_s(1), \ldots, i_s(J))$, where $1 \leq i_s(k) \leq L_k$ with $L = \prod_{k=1}^{J} L_k$. The reduction in complexity is obtained by designing J independent codebooks, with the k-th codebook being described by the pair of sets $\{A^{(1,k)}, \ldots, A^{(L_k,k)}\}$, $\{C^{(1,k)}, \ldots, C^{(L_k,k)}\}$, and the coding-decoding scheme is:

$$i_s(k) = l \quad \text{iff} \quad \mathbf{X}_{s(BR)} \in A^{(l,k)} \quad , \quad 1 \le l \le L_k,\ 1 \le k \le J \tag{9a}$$

$$\hat{\underline{x}}_{(sB+d)R} = \sum_{k=1}^{J} \underline{c}_d^{(i_s(k),k)} \qquad 0 \le d \le (B-1),\ 1 \le i_s \le L,\ \infty < s < \infty \tag{9b}$$

The linear combination of the representative vectors is not an inherent property of product codes, but seems reasonable when the synthesis is linear and high quality reconstruction of the time domain waveform is desired. We give below four particular cases of this general framework, which are usually used:

(a). *Vertical Vector Quantization* :corresponds to $B = J = 1$.

(b). *Scalar Quantization with Fixed Bit Allocation* : corresponds to $B=1$, $J=M$ (or $M/2$ for the DFT transform), L_k is the number of quantization levels of the k-th sample of the DSTT vector, and $C^{(l,k)}$ is the k-th unit vector multiplied by the value of the l-th quantization level of the k-th sample.

(c). *Horizontal Vector Quantization* : This is the generalization of the scalar quantization to the case $B > 1$, where $C^{(l,k)}$ is a matrix in which only the k-th row is non-zero.

(d). *Cascaded Vector Quantization* : This is a variant of the vertical vector quantization, with $B=1$, $J>1$. A coarse vector quantizer of size L_1 is first designed and then the residual vectors $\underline{X}_{sR} - C^{(i_s)}$ are used as input sequence for a second vector quantizer of size L_2, etc.

In analyzing this type of quantization we make the following assumptions:

(a). The quantizer is unbiased, i.e:

$$C^{(l,k)} = E[\mathbf{X}_{s(BR)} \mid \mathbf{X}_{s(BR)} \in A^{(l,k)}] \qquad 1 \le k \le J,\ 1 \le l \le L_k \tag{10a}$$

This property guarantees an unbiased output signal, and is automatically satisfied by many optimal matrix quantizers (e.g., those designed under a

minimum-mean-square-error criterion).

(b). The expected value of the input sample $x(sBR+d)$ given that $\mathbf{X}_{s(BR)} \in A^{(l,k)}$ depends only on the delay value d, and is known at least for several values of d (as specified in the sequel). Let:

$$G^{(l,k)}(d) = E[x(sBR+d) \mid \mathbf{X}_{s(BR)} \in A^{(l,k)}] \tag{11a}$$

(c). The occurrence probabilities of codewords in each one of the J codebooks is known (measured), as well as the probabilities of occurrence of a specific pair of codewords, with a specific delay $d(BR)$ between them, i.e.:

$$P^{(l,k)} = \text{Prob}\,\{\mathbf{X}_{s(BR)} \in A^{(l,k)}\} \tag{11b}$$

$$F^{(l,k),(i,j)}(d) = \text{Prob}\,\{\mathbf{X}_{s(BR)} \in A^{(l,k)} \cap \mathbf{X}_{(s+d)(BR)} \in A^{(i,j)}\} \tag{11c}$$

There are $\hat{L} \triangleq \sum_{k=1}^{J} L_k$ different codewords in the union of the J codebooks and therefore for each value of the delay d, $\hat{L}$ different values of $G^{(\cdot)}(d)$ and $\hat{L}^2$ values of $F^{(\cdot)}(d)$ have to be known. If one assumes that different codebooks' outputs are independent, then this large amount of measurements is highly reduced (nevertheless, this assumption is not needed in the sequel).

(d). We assume that the input signal $x(n)$ is composed of samples of a wide-sense stationary process with zero-mean and known autocorrelation denoted by $\rho(d)$. If the input is a speech signal, $\rho(\cdot)$ represents the long-term autocorrelation sequence of the speech, thus taking advantage of the non-flatness of the speech spectrum in the design process. Furthermore, although in various steps, this sequence appears inside infinite summations, the final result is that only the terms of $\rho(d)$ for $|d| \le L_h + L_f$ are used in the design of the optimal A/S system, and these terms appear only in the case of FQ. A similar remark applies to the covariance sequence of the additive

noise that is used in modeling the FQ effect.

In assumptions (b) and (c) above, we assumed knowledge of values which depend on the codebook used, the analysis window, and type of transform. Therefore, in the sequel we assume that all these factors have already been determined, and concentrate on selecting the optimal synthesis filter for a given analysis and coding system (quantizer). Since usually the design of a MQ is based on a typical training sequence, this sequence can also be used to determine $P^{(\cdot)}$, $F^{(\cdot)}$ and $G^{(\cdot)}(d)$ which will afterwards determine the structure of the synthesis filter. Again, to simplify the presentation, we will use the sequences $F^{(\cdot)}(d)$ and $G^{(\cdot)}(d)$ inside infinite summations, but in the final result it is implied that only the values for $|d| \leq L_f + L_h$ are actually required.

II. Error Measures.

Physically realizable A/S systems introduce signal delay. It is known that the delay of the A/S system depicted in Figs. 5.1, 5.2 is an integer multiple of the transform size [6]. We therefore, assume that the A/S system has a delay Mr_o, with r_o being an integer. Since in an ideal system the reconstructed signal $y(n)$ coincides with the delayed input signal, we define the output error signal as:

$$\varepsilon(n) \triangleq y(n) - x(n - Mr_o) \tag{12}$$

The assumption that the quantizer is unbiased ((7a),(10a)) guarantees that the output error signal has zero mean. The WOLA synthesis is a time varying operation due to the embedded interpolation, and therefore when quantization is applied $\varepsilon(n)$ is *not* a wide-sense stationary process. Thus, in order to measure the error induced by the A/S system with different synthesis filters, we generalize the usual MSE error measure for a class of non-stationary processes which

of course includes the error process $\varepsilon(n)$ above.

Let the autocorrelation sequence of $\varepsilon(n)$ be denoted by $\varphi(\cdot)$, i.e.:

$$\varphi(d,m) \triangleq E[\varepsilon(m+d)\,\varepsilon(m)] \quad \forall\, d,m \tag{13}$$

In Section V expressions for $\varphi(\cdot)$ are given for FQ and MQ. Furthermore, this function has the following two properties (as proven in Section V):

$$|\varphi(d,m)| \leq \text{Const.} \qquad \forall\, d,m \tag{14a}$$

$$\varphi(d,m+lN) = \varphi(d,m) \quad \forall\, d,m,l \tag{14b}$$

where $N \triangleq BRM/gcd(BR,M)$ (where $gcd(BR,M)$ is the greatest common divisor of the two integer numbers BR and M, and $B = 1$ for the FQ case).

Let $G(f)$ be a *non-negative* real and symmetric weight function in $L_1[-0.5,0.5]$, whose Fourier coefficients, $g(n) \triangleq \int_{-0.5}^{0.5} G(f)\, e^{j2\pi fn} df$, converges absolutely, i.e.,

$$\lim_{l\to\infty} \sum_{n=-l}^{l} |g(n)| < \infty \tag{15}$$

The first error measure we consider is given by:

$$U \triangleq \sum_{d=-\infty}^{\infty} g(d) \left[\frac{1}{N} \sum_{m=0}^{N-1} \varphi(d,m)\right] \tag{16}$$

which due to (14a) and (15) is well-defined. Furthermore, it has a spectral domain interpretation as the natural generalization of the MSE criterion for processes with bounded periodic autocorrelation sequence (i.e., satisfying (14)). The following theorem (whose proof is given in Section V) summarizes this interpretation.

Theorem 1: Let:

$$u_{l,r} \triangleq \frac{1}{(2r+1)} \int_{-0.5}^{0.5} E \mid \sum_{n=l-r}^{l+r} \varepsilon(n) e^{-j2\pi f n} \mid^2 G(f) df \qquad \forall\, l,r \tag{17}$$

Then:

(a). For any value of l, $\lim_{r\to\infty} (u_{l,r}) = U$. Therefore, $U \geq 0$.

(b). For $\varepsilon(n)$ that is wide-sense stationary process, having a spectrum $S(f) \in L_2[-0.5,0.5]$, it follows that:

$$U = \int_{-0.5}^{0.5} S(f)\, G(f) df \tag{18}$$

The error measure U defined in (16) is based on the error signal in the time domain. An alternative approach is to define the error in the transform domain (i.e., in the domain in which the quantization is done). Let $\underline{Y}_{sR}$, $-\infty \leq s \leq \infty$ denote the DSTT sequence of vectors generated from the reconstructed signal $y(n)$. In an ideal A/S system, $y(n)$ is a delayed version of the DSTT of the input signal (with delay of Mr_0 samples). Thus, the sequence of error vectors in the transform domain is:

$$\underline{E}_{sR} \triangleq \underline{Y}_{sR} - \underline{X}_{(sR-Mr_0)} \tag{19}$$

Since the DSTT is a linear operation and the error in time domain $\varepsilon(n)$ has zero mean, it follows immediately that $\underline{E}_{sR}$ has also zero mean. With the DFT transform in mind, it is natural to consider the expected value of the Euclidian norm of the random error vectors $\underline{E}_{sR}$, i.e.:

$$v_{sR} \triangleq \mathbf{E}[\|\underline{E}_{sR}\|^2] = \sum_{k=0}^{M-1} \mathbf{E}[\mid Y_{sR}(k) - X_{sR-Mr_0}(k) \mid^2] \tag{20}$$

It follows from the definition of the DSTT ((1) and (2)), and of $\varepsilon(n)$ (in (12)), that

v_{sR} can be represented in terms of the autocorrelation sequence $\varphi(\cdot)$ as:

$$v_{sR} = \sum_{r=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} h(sR-t)\, h(sR-r)\, \varphi(r-t,t)\, a((r)_M, (t)_M) \qquad (21)$$

where $a(i,j)$ is the (i,j)-th element of the $M \times M$ matrix defined as:

$$A = T^* T \qquad (22)$$

where * denotes 'conjugate transpose'.

Since the A/S system is time-varying, the error signal $\underline{E}_{sR}$ is not a wide-sense stationary process, and v_{sR} depends on the time instance sR. However, due to (14b), the sequence v_{sR} is periodic with a period of N, i.e.:

$$v_{(sR+lN)} = v_{sR} \qquad \forall\, s,l \ \text{ integers} \qquad (23)$$

We now define the second error measure as the time average of v_{sR} over one period of this sequence, i.e.:

$$V \triangleq \frac{1}{N} \sum_{m=0}^{N-1} v_{sR+m} \qquad (24)$$

Substituting (24) in (21) results in:

$$V = \sum_{d=-\infty}^{\infty} \Big[\sum_{n=-\infty}^{\infty} h(n)\, h(n+d) \Big] \frac{1}{N} \Big[\sum_{m=0}^{N-1} \varphi(d,m)\, a((m+d)_M, (m)_M) \Big] \qquad (25)$$

Comparing (25) with (16) it is easily verified that when the matrix A is a circulant matrix (i.e., $a(i,j) = a((i+d)_M, (j+d)_M)$ for $d=0, \ldots, (M-1)$), V coincides with U, provided that the following weight sequence $g(d)$ is used:

$$g(d) = \sum_{n=-\infty}^{\infty} h(n)\, h(n+d)\, a((d)_M, 0) \qquad (26)$$

It is shown in Section V that (26) is equivalent to:

$$G(f) = \frac{1}{M}\sum_{k=0}^{M-1} | H(f + \frac{k}{M})|^2 |\lambda_k|^2 \tag{27}$$

where $H(f)$ is the frequency response of the analysis window and $\{|\lambda_k|^2\}_{k=0}^{M-1}$ are the eigenvalues of the circulant matrix A which are found by applying the IDFT on its zero-th column. Thus, for example, when the transform T used is the DFT, $\lambda_k = 1$, $0 \le k \le M-1$, and since the analysis window is an approximation of an ideal LPF with cut-off frequency $\frac{1}{2M}$ it follows that $G(f) \approx 1$ in (27).

It is also shown in section V that A is a circulant matrix iff the transform T is of the form $T = \Psi\tilde{A}$ where $\tilde{A}$ is a circulant matrix and Ψ is a unitary matrix. In particular, this is guaranteed for all unitary transforms.

Since the error measure V coincides with U for this wide class of transforms, we continue in what follows with U only.

The error measure U given in (16) is suitable for the design of an optimal synthesis filter, given the analysis window. However, when the design of the optimal analysis window is considered, the error measure U should be modified in order to incorporate the low-pass frequency response specification of the analysis window, into the design process. Following the approach in [54], we add the weighted MSE between $H(f)$ and the desired frequency response $D(f)$, to U. Let $W(f) \ge 0$ denote the weight function, and $\hat{U}$ the modified error measure, then:

$$\hat{U} = U + \int_{-0.5}^{0.5} W(f)\,|H(f) - D(f)|^2\, df \tag{28}$$

III. Design of Optimal Synthesis Filters.

The optimality criterion is the minimization of U with respect to the unknown synthesis filter coefficients. Combining (16) and the expressions for $\varphi(\cdot)$ given in Section V, it is easily verified that for both quantization approaches considered here, U is a P.S.D. quadratic form in terms of the synthesis filter vector of coefficients $\underline{f}$, i.e.:

$$U = C + \frac{1}{R}(\underline{f}'Q\underline{f} - \underline{b}'\underline{f}) \tag{29}$$

where the apostrophe denotes transposition. The expressions for the $L_f \times L_f$ matrix Q, the vector $\underline{b}$ of dimension L_f and the constant C are given in section VI.

Thus, the optimal synthesis filter is given by the solution of the following set of linear equations:

$$(Q + Q')\underline{f}_{opt} = \underline{b} \tag{30}$$

If this set of equations is degenerate, each of the infinitely many possible solutions corresponds to the same (minimal) value of U.

We now interpret the general expressions for various particular cases of practical importance, and analyze both the complexity of the solution and the type of data needed, starting with FQ. In this case, $Q=Q_v+Q_h$, where the matrix Q_v reflects the quantization effect and depends on $\Psi_{m,n}(\cdot)$ (defined in (7b)), and the matrix Q_h depends on the analysis window and corresponds to Portnoff's conditions [15]. It can be verified that when a unity system exists (which requires either $R < M$ or $L_f=L_h=M=R$) and no quantization is applied (i.e., $\Psi_{m,n}(d) = 0$ and therefore $Q_v = 0$), any unity system is a solution of the (possibly degenerate) set of equations (30), regardless of the weight function $G(f)$ and the input statistics.

For $R=M$ and $L_f, L_h > M$ (which is the typical situation in waveform coding applications), *no* unity system exists [12]. In this case we derive in section VI sufficient conditions for uniqueness of the solution of (30) even if no quantization is applied. This solution accounts for the statistical properties of the input signal.

If quantization is applied, uniqueness of the solution of (30) is guaranteed when $G(f)=1$ and the noise process $v_{sR}(m)$ is nondegenerate for $0 \le m \le M-1$ (in the sense that $v_{sR}(m)$ cannot be predicted with zero mean square error from the last $\frac{L_f}{R}$ samples $v_{sR-dR}(m)$, $d=1, \ldots, \frac{L_f}{R}$). The proof is given in section VI. Similar conditions can be derived for $G(f) \neq 1$, but are omitted here for simplification.

For $Q_v=0$ there are infinitely many possible solutions of (30), but one can choose among them the appropriate solution assuming a characteristic noise statistic $\Psi_{m,n}(dR)$ as follows:

Use $\varepsilon \, \Psi_{m,n}(dR)$ in the expression for Q_v, where $\varepsilon > 0$ is a small value governing the overall noise level. If the uniqueness condition presented above is satisfied, then $(Q + Q')$ is a Positive Definite Symmetric matrix for every $\varepsilon > 0$. Let $\underline{f}_\varepsilon$ be the unique solution of (30) for $\varepsilon > 0$, then it is shown in section VI that $\lim_{\varepsilon \to 0} (\underline{f}_\varepsilon)$ exists. Furthermore, it is shown there, how this limiting synthesis filter can be determined explicitly.

It is important to note that the above limiting synthesis filter (which guarantees a unity system), *depends* upon *the specific noise and input statistics*, as further elaborated in section VI.

In order to illustrate the effect of the quantization noise, we consider now two simple examples in which a closed form solution of (30) can be obtained.

Example 1:

$G(f)=1,\ L_h=L_f=M,\ \tau_o=1,\ \rho(d)=\sigma_x^2\delta(d),\ \Psi_{m,n}(dR)=\delta(d)\hat{\Psi}((m-n)_M).$

In this case it is shown in section VI that the optimal synthesis filter is:

$$f_{opt}(t) = \frac{h(M-t)}{\sum_{r=-\infty}^{\infty} h(M-t-rR)^2 + \hat{\Psi}(0)/\sigma_x^2} \qquad 0 \le t \le M-1 \tag{31}$$

which extends the results in [16,17], to include the effect of quantization noise.

Example 2:

$G(f)=1,\ R=M$. In this case the set of L_f linear equations given in (30) is decomposable into M sets, of about the same number of equations in each. Let $f_\tau(x)$ denote the τ-th polyphase [6] of the synthesis filter (i.e., $f_\tau(x) \triangleq f(\tau+Mx),\ 0\le\tau\le M-1$) ; $h_\tau(x)$ denotes the properly delayed and inverted in time, τ-th polyphase of the analysis window (i.e., $h_\tau(x) = h(M\tau_0-(\tau+Mx)),\ 0\le\tau\le M-1$), and $\tilde{\rho}(d)$ denotes the decimated, by M, autocorrelation sequence of the input signal. Then, the optimal synthesis filter is the solution of:

$$\sum_{d=-\infty}^{\infty} h_\tau(x-d)\tilde{\rho}(d) = \sum_{y=-\infty}^{\infty} f_\tau(x-y)[R_\tau(y) + \Psi_{\tau,\tau}(yM)] \qquad \begin{array}{l} 0\le\tau\le M-1 \\ 0\le x\le \left\lfloor \frac{L_f-1-\tau}{M} \right\rfloor \end{array} \tag{32}$$

where:

$$R_\tau(y) \triangleq \sum_{r=-\infty}^{\infty} h_\tau(r+y) \sum_{l=-\infty}^{\infty} h_\tau(r-l)\tilde{\rho}(l) \qquad -\infty < y < \infty \tag{33}$$

The derivation of (32),(33) is given in section VI. For $L_f = \infty$ (and non-casuality of the synthesis filter is allowed) these equations have a frequency domain interpretation which resembles the classical Wiener filter. Let $F_\tau(f),\ H_\tau(f),\ \tilde{R}(f),\ \Psi_{\tau,\tau}(f)$ be the frequency responses of $f_\tau(x),\ h_\tau(x),\ \tilde{\rho}(d)$

and $\Psi_{T,T}(dM)$ respectively, then (32), (33) imply that:

$$F_T(f) = \frac{H_T(f)\,\tilde{R}(f)}{\Psi_{T,T}(f) + \tilde{R}(f)\,|H_T(f)|^2} \qquad -0.5 < f \le 0.5 \tag{34}$$

In order to analyze the complexity of the design procedure and the data needed, we assume throughout that $g(d)=0$ for $|d| \ge L_g$. Table I summarizes the complexity of various steps in the design process, and the maximal distance d for which $\rho(d)$ and $\Psi_{m,n}(d)$ still affects the solution (assuming that: $Mr_0 \triangleq (L_f + L_h)/2$, $f(x) = 0$ for $x > L_f - 1$ or $x < 0$, and $h(x)=0$ for $x \le 0$ or $x > L_h$), the details are given in section VI.

Note that at least for $G(f)=1$ and $M=R$ the complexity of the design is small as only $(L_f + L_h)/M$ values of $\rho(\cdot)$, and L_f/M values of $\Psi_{m,m}(\cdot)$, are used (this case refers to Example 2 above).

We turn now to MQ.

In this case, due to the inherent non-linearity of the quantization process, the matrix Q cannot be decomposed into $Q_h + Q_v$ as with FQ. Moreover, the implicit *dependence* of the synthesis filter on the given analysis window coefficients is only via the statistics $F^{(\cdot)}(d)$, $G^{(\cdot)}(d)$ and $P^{(\cdot)}$. Furthermore, since there are only L possible different MDSTT vectors, it is clear that in general no unity system exists (i.e., $U > 0$ even for $\underline{f}_{opt}$ obtained via (30), and no matter what are the values of R, M, L_h, L_f). A sufficient condition for the uniqueness of $\underline{f}_{opt}$ obtained from (30) is (as for FQ): $G(f)=1$, and the process $\hat{x}_{sR}(m)$ is nondegenerate for $0 \le m \le M-1$. The only example in which a closed form solution of (30) is easily obtained is for $G(f)=1$, $L_f = M = R$, in which case:

$$f(m) = \frac{\sum_{d=0}^{B-1} \sum_{l=1}^{\hat{L}} P^{(l)} c_d^{(l)}(m) G_d^{(l)}(m+M(d-r_o))}{\sum_{d=0}^{B-1} \sum_{k=1}^{\hat{L}} \sum_{l=1}^{\hat{L}} F^{(k),(l)}(0) c_d^{(k)}(m) c_d^{(l)}(m)} \qquad 0 \le m \le M-1 \tag{35}$$

Where for ease of presentation we reorder the $\hat{L}$ codewords and thus replace the pairs of indices (l,k) in (11) by a single index.

This result is derived in section VI, as well as the simplified expressions of Q and $\underline{b}$ for the important case of $G(f)=1$, $R=M$.

It is also shown there that the complexity of the design in this case is small, provided that $\hat{L}$ is small enough (typically for $\hat{L} \approx 256$).

IV. Design of Optimal A/S Systems with Fine Quantization.

If fine quantization (FQ) is used, the error measure $\hat{U}$ is given as an explicit function of the coefficients of the analysis window. Combining (16) and (28), and the expression for $\varphi(\cdot)$ given in Section V, it is easily verified that $\hat{U}$ is a P.S.D. quadratic form in terms of the analysis window vector of coefficients $\underline{h}$. Thus, *for a given synthesis* filter, the optimal analysis window which minimizes $\hat{U}$ is the solution of a linear set of equations, as follows:

$$(\tilde{Q}_f + \tilde{Q}'_f + \tilde{Q}_D + \tilde{Q}'_D)\, \underline{h}_{opt} = \underline{\tilde{b}}_D + \underline{\tilde{b}}_f \tag{36}$$

The expressions for the $L_h \times L_h$ matrices $\tilde{Q}_f, \tilde{Q}_D$ and the vectors $\underline{\tilde{b}}_D, \underline{\tilde{b}}_f$ of dimension L_h are given in section VII. It is also shown there that for a weight function $W(f)$ which is positive on a set of non-zero measure, there exists a unique solution of (36). The matrix $\tilde{Q}_D$ and the vector $\underline{\tilde{b}}_D$ reflect the frequency response specification on the analysis window via (28), whereas $\tilde{Q}_f$ and $\underline{\tilde{b}}_f$, which depend on the synthesis filter, reflect the desired unity system specification. The statistics of the quantization noise *do not affect* the optimal analysis window. Furthermore, regardless of the quantization, the optimal analysis window *does not*

correspond in general to a unity system, but is rather a compromise between the frequency response specification and the unity system specification. Unlike the set of linear equations in (30), the equations in (36) are usually irreducible even for $M=R$ and $G(f)=1$.

The design of an optimal A/S system with FQ is done by the following iterative algorithm, similar to [54]:

1. Initialize $r=0$, and let $\underline{h}^{(0)}$, $\underline{f}^{(0)}$ denote the given initial analysis and synthesis filters.

2. Let $\underline{h}^{(r+1)} \in A_1(\underline{f}^{(r)})$ be any solution of (36), for the given synthesis filter $\underline{f}^{(r)}$, with the exception that if $\underline{h}^{(r)}$ is a solution choose this solution.

3. Let $\underline{f}^{(r+1)} \in A_2(\underline{h}^{(r+1)})$ be any solution of (30), for the given analysis window $\underline{h}^{(r+1)}$, with the exception that if $\underline{f}^{(r)}$ is a solution choose this solution.

4. If $\underline{f}^{(r)} = \underline{f}^{(r+1)} \cap \underline{h}^{(r)} = \underline{h}^{(r+1)}$ stop; otherwise $r \leftarrow (r+1)$ and return to *2*.

Here $A_1(\cdot)$ and $A_2(\cdot)$ denote the sets of solutions of (36) and (30) respectively.

This iterative algorithm consists of alternately solving two sets of linear equations. For the special case of: $R=M$, $G(f) = 1$, $\Psi_{m,n}(d) = 0$, and $\rho(d) = \delta(d)$, $\hat{U}$ is similar to the error measure in [54] which is the deterministic MSE between the A/S system unit sample response and the desired ideal response. The only difference is that we incorporated the linear constraint on the gain of the system, used in [54] into the error measure, thus avoiding the need of using Lagrange multipliers and hence simplifying the design algorithm.

The following theorem which is proved in section VII summarizes the convergence properties of the iterative algorithm, assuming that $W(f) > 0$ on a set of non-zero measure (thus, (36) possess a unique solution for every synthesis filter), and denoting the set of fixed points of the iterative algorithm by Γ.

Theorem 2:

(A). $\hat{U}$ is monotonically decreasing from iteration to iteration, unless the algorithm stops at a fixed point in Γ.

(B). Γ is the set of stationary points of $\hat{U}$, and all the stable points in Γ are local minima of $\hat{U}$.

(C). When Q_v is non-singular (i.e., (30) also possess a unique solution for every analysis window), then:

1. Every sequence generated by the algorithm has at least one limit point, and all limit points are in Γ.
2. $\hat{U}$ possess a global minimum which is in Γ.

(D). Assume that (30) possess a unique solution only in some neighborhood $C(\rho_0) = \{\underline{h};(\underline{h}-\underline{\tilde{h}})'Q_D(\underline{h}-\underline{\tilde{h}}) \leq \rho_0\}$ of $\underline{\tilde{h}} \triangleq (Q_D+Q'_D)^{-1}\underline{b}_D$. Then, for any initial condition $(\underline{f}^{(0)},\underline{h}^{(0)})$ satisfying $[\hat{U}(\underline{f}^{(0)},\underline{h}^{(0)}) - \inf U(\underline{f},\underline{h}) - K_0] \leq \rho_0$, property 1 of (C) holds, where $K_0 \triangleq \int_{-0.5}^{0.5} W(f)|D(f)-\tilde{H}(f)|^2 df$ is the frequency response error of the analysis window $\underline{\tilde{h}}$. If such an initial condition exists, then property 2 of (C) holds.

(E). Even when (30) does not possess a unique solution for any $\underline{h}$, the sequence of analysis windows generated by any instance of the algorithm has at least one limit point. Furthermore, any limit point of both $\underline{f}$ and $\underline{h}$ is in Γ.

Assume that $(\underline{f}^*,\underline{h}^*) \in \Gamma$, and that at $(\underline{f}^*,\underline{h}^*)$ both (30) and (36) possess unique solutions, then the iterative algorithm converges linearly to $(\underline{f}^*,\underline{h}^*)$. The rate of convergence is given by:

$$\underline{\delta x}^{(n+1)} = L^2 \underline{\delta x}^{(n)} + o(\underline{\delta x}^{(n)}) \tag{37}$$

where $\underline{\delta x}^{(n)} \triangleq \begin{bmatrix} \underline{f}^{(n)} - \underline{f}^* \\ \underline{h}^{(n)} - \underline{h}^* \end{bmatrix}$ is a vector in $\mathbb{R}^{(L_f+L_h)}$ which measures the convergence error at the (n)-th iteration, $o(\underline{\delta x}^{(n)})$ denotes a vector of functions $\underline{g}(\underline{\delta x}^{(n)})$ such that $\lim_{\|\delta x^{(n)}\|\to 0} \{\underline{g}(\underline{\delta x}^{(n)})/\|\underline{\delta x}^{(n)}\|\} = \underline{0}$, and the matrix L is given by:

$$\mathbf{L} = \begin{bmatrix} 0 & | & -\mathbf{A}^{-1}\mathbf{C} \\ - & | & - \\ -\mathbf{B}^{-1}\mathbf{C}' & | & 0 \end{bmatrix} \tag{38}$$

where:

$$A_{ij} \triangleq \left[\frac{\partial^2 \hat{U}}{\partial f_i \partial f_j}\right]_{(\underline{f}^*\underline{h}^*)} = (Q_{h^*} + Q'_{h^*} + Q_v + Q'_v) \tag{39a}$$

$$B_{ij} \triangleq \left[\frac{\partial^2 \hat{U}}{\partial h_i \partial h_j}\right]_{(\underline{f}^*\underline{h}^*)} = (\tilde{Q}_{f^*} + \tilde{Q}'_{f^*} + \tilde{Q}_D + \tilde{Q}'_D) \tag{39b}$$

$$C_{ij} = \left[\frac{\partial^2 \hat{U}}{\partial f_i \partial h_j}\right]_{(\underline{f}^*\underline{h}^*)} = \left.\frac{\partial\{(Q_h + Q'_h)\underline{f} - \underline{b}\}_i}{\partial h_j}\right|_{(\underline{f}^*\underline{h}^*)} \qquad 1 \le i \le L_f,\ 1 \le j \le L_h \tag{39c}$$

The derivation of (37), (38) and the explicit expression for (39c) are given in section VII. It readily follows from (37), (38) that:

$$\lim_{\|\delta x^{(n)}\|\to 0} \left\{\frac{\|\underline{f}^{(n+1)} - \underline{f}^*\|}{\|\underline{f}^{(n)} - \underline{f}^*\|}\right\} \le \sqrt{\lambda_{MAX}(\mathbf{C}\mathbf{B}^{-1}\mathbf{C}'\mathbf{A}^{-2}\mathbf{C}\mathbf{B}^{-1}\mathbf{C}')} \le \frac{\lambda_{MAX}(\mathbf{C}\mathbf{C}')}{\lambda_{MIN}(\mathbf{A})\lambda_{MIN}(\mathbf{B})} \tag{40a}$$

$$\lim_{\|\delta x^{(n)}\|\to 0} \left\{\frac{\|\underline{h}^{(n+1)} - \underline{h}^*\|}{\|\underline{h}^{(n)} - \underline{h}^*\|}\right\} \le \sqrt{\lambda_{MAX}(\mathbf{C}'\mathbf{A}^{-1}\mathbf{C}\mathbf{B}^{-2}\mathbf{C}'\mathbf{A}^{-1}\mathbf{C})} \le \frac{\lambda_{MAX}(\mathbf{C}'\mathbf{C})}{\lambda_{MIN}(\mathbf{A})\lambda_{MIN}(\mathbf{B})} \tag{40b}$$

V. Derivation of the results presented in section II.

1. Explicit equations for $\varphi(d,m)$ and its properties:

Following (12),(13) and (4) it follows that:

$$\varphi(d,m) = \rho(d) - \sum_{s=-\infty}^{\infty} f(m-sR)E[\hat{x}_{sR}((m)_M)x(m+d-Mr_o)] -$$

$$- \sum_{s=-\infty}^{\infty} f(m+d-sR)E[\hat{x}_{sR}((m+d)_M)x(m-Mr_o)] \qquad \text{(A1)}$$

$$+ \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} f(m-tR)f(m+d-sR)E[\hat{x}_{sR}((m+d)_M)\hat{x}_{tR}((m)_M)]$$

Following (1),(6) and (7) we obtain for the *FQ case:*

$$E[\hat{x}_{sR}((m)_M)x(m+d-Mr_o)] = \sum_{r=-\infty}^{\infty} h(sR-m-rM)\rho(M(r+r_o)-d) \qquad \text{(A2a)}$$

$$E[\hat{x}_{sR}((m+d)_M)x(m-Mr_o)] = \sum_{r=-\infty}^{\infty} h(sR-m-d-rM)\rho(M(r+r_o)+d) \qquad \text{(A2b)}$$

$$E[\hat{x}_{sR}((m+d)_M)\hat{x}_{tR}((m)_M)] = \Psi_{(m+d)_M,(m)_M}((t-s)R) +$$

$$\text{(A2c)}$$

$$+ \sum_{r=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} h(tR-m-rM)h(sR-m-d-nM)\,\rho(d+M(n-r))$$

Following (9-11) we obtain for the *MQ case:*

$$E[\hat{x}_{sR}((m)_M)x(m-d-Mr_o)] = \sum_{l=1}^{L} P^{(l)} c\binom{l}{s}_B ((m)_M) G^{(l)}(m+d-Mr_o-\lfloor \frac{s}{B} \rfloor BR) \text{(A3a)}$$

$$E[\hat{x}_{sR}((m+d)_M)x(m-Mr_o)] = \sum_{l=1}^{L} P^{(l)} c\binom{l}{s}_B ((m+d)_M) G^{(l)}(m-Mr_o-\lfloor \frac{s}{B} \rfloor BR) \text{(A3b)}$$

$$E[\hat{x}_{sR}((m+d)_M)\hat{x}_{tR}((m)_M)] =$$

$$\sum_{l=1}^{\hat{L}} \sum_{k=1}^{\hat{L}} F^{(l),(k)} (\lfloor \frac{t}{B} \rfloor - \lfloor \frac{s}{B} \rfloor) \, c\binom{l}{s}_B ((m+d)_M) c\binom{k}{t}_B ((m)_M) \tag{A4}$$

All the infinite summations in (A1) and (A2) are actually finite since both $f(\cdot)$ and $h(\cdot)$ are of finite length. The periodicity of $\varphi(d,m)$ in the second variable with a period of $N \triangleq BRM/gcd(BR,M)$ can be verified from (A1-A4) using elementary (although tedious) algebra. For the boundness of $|\varphi(d,m)|$ it is enough to check that this sequence is bounded with respect to d, for $0 \le m \le N-1$. This property follows from the fact that $f(\cdot)$ and $h(\cdot)$ are of finite values and length, and from the assumption that the covariance sequences $\rho(\cdot)$ and $\Psi(\cdot)$ are bounded, as well as the conditional expectations sequences $G^{(l)}(\cdot)$. Note that the terms which appear in (A4) are always bounded, due to the boundness of $0 \le F^{(l),(k)}(\cdot) \le 1$. The boundness assumptions on the input signal are quite natural and not very restrictive.

2. Proof of Theorem 1:

(a). Since the summation over n in (17) is finite, all the expectations involved are finite (due to (14a)), and the Fourier coefficients of $G(f)$ are well-defined, it follows that:

$$u_{l,r} = \sum_{n=l-r}^{l+r} \sum_{m=l-r}^{l+r} g(n-m)\, \varphi(n-m,m) \frac{1}{(2r+1)} \tag{A5}$$

Using the periodicity of $\varphi(\cdot)$ with respect to the second variable (see (14b)), we rewrite (A5) as follows:

$$u_{l,r} = \sum_{d=-2r}^{2r} g(d) \sum_{m=0}^{N-1} \varphi(d,m)\, w_{l,r}(d,m) \tag{A6}$$

with

$$w_{l,r}(d,m) \triangleq |\{x; l-r \le x \le l+r \wedge l-r \le x+d \le l+r \wedge x \equiv m (\mathrm{mod}\ N)\}| / (2r+1) .$$

The sequence $w_{l,r}(d,m)$ is a window sequence having the following two properties:

$$w_{l,r}(d,m) = 0 \qquad |d| \ge (2r+1) \tag{A7a}$$

$$\left\lfloor \frac{2r+1-|d|}{N} \right\rfloor \frac{1}{2r+1} \le w_{l,r}(d,m) \le \left\lceil \frac{2r+1-|d|}{N} \right\rceil \frac{1}{2r+1} \qquad |d| \le 2r \tag{A7b}$$

and from (A7b) it follows immediately that:

$$\left| \frac{1}{N}\left(1-\frac{|d|}{2r+1}\right) - w_{l,r}(d,m) \right| \le \frac{1}{(2r+1)} \qquad |d| \le 2r \tag{A7c}$$

Define:

$$v_r \triangleq \sum_{d=-2r}^{2r} g(d) \left[\frac{1}{N} \sum_{m=0}^{N-1} \varphi(d,m) \right] \left(1 - \frac{|d|}{2r+1}\right) \tag{A8}$$

$$u_{2r} \triangleq \sum_{d=-2r}^{2r} g(d) \left[\frac{1}{N} \sum_{m=0}^{N-1} \varphi(d,m) \right] \tag{A9}$$

Then, from (A7a), (A7c), (A8), (A6), (14a), and the triangle inequality for the l_1 norm:

$$|v_r - u_{l,r}| \le \sum_{d=-2r}^{2r} |g(d)| \, C \, \frac{1}{(2r+1)} \tag{A10}$$

Due to the Riemann-Lebesgue lemma [71], for every $G(f) \in L_1[-0.5, 0.5]$, $\lim_{d \to \infty} |g(d)| = 0$. Therefore, from (A10) it follows that $\lim_{r \to \infty} |v_r - u_{l,r}| = 0$. Thus, in order to prove that $\lim_{r \to \infty} (u_{l,r}) = U$ it is sufficient to show that $\lim_{r \to \infty} v_r = U$. Since $v_r = \frac{1}{2r+1} \sum_{k=0}^{2r} u_k$ it follows that $\lim_{r \to \infty} u_r = U$ implies

that $\lim_{r\to\infty} v_r = U$ (although the opposite direction might not hold). Let:

$$\tilde{\varphi}(d) = \frac{1}{N}\sum_{m=0}^{N-1} \varphi(d,m) \tag{A11}$$

It is obvious that if u_r converges to $U < \infty$, then U is given by (16). Furthermore, given that $\sum_{d=-r}^{r} g(d)$ converges absolutely, for $\tilde{\varphi}(d)$ which is bounded (by (14a)), $\sum_{d=-r}^{r} g(d)\,\tilde{\varphi}(d)$ also converges absolutely, and therefore $\lim_{r\to\infty} u_r = U < \infty$. This completes part (a) of the proof.

(b). For $\varepsilon(n)$ which is a wide-sense stationary process, the period N of $\varphi(\cdot)$ is one. Therefore, from (16) $U = \sum_{d=-\infty}^{\infty} g(d)\varphi(d)$. Since $g(d)$ converges absolutely, it follows that $G(f) \in L_\infty[-0.5,0.5] \subseteq L_2[-0.5,0.5]$. If $\varepsilon(n)$ has a spectrum, $S(f)$, which is in $L_2[-0.5,0.5]$, we can apply Parseval's theorem to obtain (18) (due to the realness of $G(f)$ the complex conjugate operation is omitted).

3. The relation between the error measures U and V:

It is well-known that for a circulant matrix A the eigenvalues are the IDFT of the zero-th column [67]. Furthermore, due to (22), A is a P.S.D., Hermitian matrix and therefore it possess real non-negative eigenvalues, i.e.,

$$|\lambda_k|^2 = \sum_{m=0}^{M-1} a(m,0)e^{+j\frac{2\pi}{M}mk} \tag{A12}$$

Substituting (A12) and the definition of $H(\cdot)$ into (27) we obtain that:

$$G(f) = \frac{1}{M}\sum_{m=0}^{M-1}\sum_{k=0}^{M-1} a(m,0)e^{j\frac{2\pi}{M}mk} \sum_{n=-\infty}^{\infty}\sum_{r=-\infty}^{\infty} h(n)\,h(r)e^{j2\pi(f+\frac{k}{M})(n-r)} \tag{A13}$$

Where actually all summations are finite. Using the well-known formula

$$\frac{1}{M}\sum_{k=0}^{M-1} e^{j\frac{2\pi}{M}k(m+n-r)} = \begin{cases} 1 & m \equiv (r-n) \bmod M \\ 0 & \text{otherwise} \end{cases}$$

we obtain:

$$G(f) = \sum_{n=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} h(n)h(r)e^{j2\pi f(n-r)}\, a((r-n)_M, 0)$$

$$= \sum_{d=-\infty}^{\infty} e^{-2\pi f d} \sum_{n=-\infty}^{\infty} h(n)h(n+d)a((d)_M, 0) \qquad \text{(A14)}$$

By evaluating the Fourier coefficients of the trigonometric polynomial $G(f)$ given in (A14), we readily obtain (26). Since $g(d)$ is a sequence of finite length, there is a one to one correspondence between $g(d)$ in (26), and $G(f)$ in (27).

Now if $T = \Psi\tilde{A}$ with Ψ a unitary matrix and $\tilde{A}$ a circulant matrix, then A which is defined in (22) equals $T^*T = \tilde{A}^*\tilde{A}$. The class of circulant matrices is closed under conjugate transpose operation and under multiplications [67]. Therefore, for this class of transforms $A = \tilde{A}^*\tilde{A}$ is a circulant matrix.

Assume that A is a circulant matrix. Due to (22) and the assumption that T is a regular transformation it follows that A is a P.D. Hermitian matrix and therefore:

$$\Lambda = \Omega^* A \Omega = (\Omega^* T \Omega)^* (\Omega^* T \Omega) \qquad \text{(A15)}$$

Where Ω is the normalized DFT matrix which is a unitary matrix, and Λ is a diagonal, P.D. matrix [67]. We define $\tau \triangleq \Omega^* T \Omega$, and thus:

$$\tau^* \tau = \Lambda \qquad \text{(A16)}$$

Since Λ is a non-singular matrix, so is any solution τ of (A16). It is straightforward to verify that if τ_1, τ_2 are two solutions of (A16), then $\tau_1^{-1}\tau_2$ is a unitary matrix. Thus, any solution of (A16) is of the form:

$$\tau = \tilde{\Psi}\Lambda^{1/2} \tag{A17}$$

Where $\tilde{\Psi}$ is an arbitrary unitary matrix, and $\Lambda^{1/2}$ is a diagonal matrix (satisfying $\Lambda^{1/2}\Lambda^{1/2} = \Lambda$). Now, from (A17) and the relation between T and τ, it follows that:

$$T = \Omega\tau\Omega^{*} = (\Omega\tilde{\Psi}\Omega^{*})(\Omega\Lambda^{1/2}\Omega^{*}) \tag{A18}$$

Since $\Lambda^{1/2}$ is a diagonal matrix it is well-known that $\tilde{A} \triangleq (\Omega\Lambda^{1/2}\Omega^{*})$ is a circulant matrix [67], and since $\Omega, \tilde{\Psi}$ and Ω^{*} are unitary matrices so is $\Psi \triangleq \Omega\tilde{\Psi}\Omega^{*}$. Thus, $T = \Psi\tilde{A}$ with Ψ a unitary matrix and $\tilde{A}$ a circulant matrix.

VI. Derivation of the results presented in section III.

1. The expressions of $Q, \underline{b}, C$ in equation (29):

The quadratic form (29) is obtained by substituting (A1-A4) in (16) and rearranging the resulting expression as a quadratic form in terms of the coefficients of the synthesis filter. We present here only the final results, i.e., the expressions for computing the matrix Q, the vector $\underline{b}$ and the scalar C.

$$C = \sum_{d=-\infty}^{\infty} \rho(d) g(d) \tag{B1}$$

For the *FQ case* the t-th element of the vector $\underline{b}$ $(0 \le t \le L_f - 1)$ is:

$$b(t) = \sum_{r=-\infty}^{\infty} h(Mr-t) \sum_{d=-\infty}^{\infty} g(d)(\rho(d-M(r-r_o)) + \rho(-d-M(r-r_o))) \tag{B2a}$$

and the elements of the matrix Q are given by:

$$Q(t,s) = \sum_{r=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} h(Mr-s)\, h(Mn-t) \sum_{d=-\infty}^{\infty} \rho(M(n-r) + (s-t) + Rd) g(s-t+Rd) \,.$$

$$+ \sum_{d=-\infty}^{\infty} g(s-t+Rd) \frac{R}{N} \sum_{\substack{m=0 \\ m \equiv (t)_R}}^{(N-1)} \Psi_{(m+(s-t)+Rd)_M, (m)_M}(-Rd) \; ; \; 0 \le t, s \le L_f - 1 \tag{B2b}$$

Likewise, for the *MQ case* we obtain:

$$b(t) = \sum_{j=0}^{B-1} \sum_{l=1}^{\hat{L}} P^{(l)} \frac{R}{N} \sum_{\substack{m=0 \\ (m-t)_R = 0 \\ (m-t)_{BR} = j}}^{N-1} c_j^{(l)}((m)_M)$$

$$\cdot \sum_{d=-\infty}^{\infty} g(d)(G^{(l)}(d-Mr_o+t+jR) + G^{(l)}(-d-Mr_o+t+jR)) \tag{B3a}$$

$$Q(t,s) = \sum_{j=0}^{B-1} \sum_{l=1}^{\hat{L}} \sum_{k=1}^{\hat{L}} \frac{R}{N} \sum_{\substack{m=0 \\ (m-t)_R=0 \\ (m-t)_{BR}=j}}^{N-1}$$

$$\cdot \sum_{d=-\infty}^{\infty} c^{(l)}_{(j+d)_B}((m+(s-t)+Rd)_M)\, c_j^{(k)}((m)_M)\, F^{(l),(k)}(-\lfloor \frac{j+d}{B} \rfloor) g(s-t+Rd) \tag{B3b}$$

The indices (t,s) in (B2,B3) are limited to a finite interval of length L_f, for which the synthesis filter's coefficients are non-zero. The summations over the (r,n) indices in (B2) are therefore finite, but the summations over d in (B2,B3) may be infinite, if $g(\cdot)$ is infinite. However, these summations can be alternatively evaluated in the frequency domain using Parsevall's theorem. The matrix Q in (B2b) is interpreted throughout as the sum of two matrices $Q_h + Q_v$, where $Q_h(t,s)$ is the first term in the right hand side of (B2b) and $Q_v(t,s)$ is the second term.

2. Unity systems when $\Psi_{m,n}(\cdot) = 0$.

We consider the case of FQ with $\Psi_{m,n}(\cdot) = 0$. In this case $\hat{x}_{sR}(m) = x_{sR}(m)$. A synthesis filter satisfies Portnoff's rules, iff:

$$\sum_{s=-\infty}^{\infty} f(m-sR)\, h(sR-m-M(r-r_0)) = \begin{cases} 1 & r=0 \\ 0 & \text{otherwise} \end{cases} \tag{B4}$$

In that case $\varepsilon(n) \triangleq y(n) - x(n-Mr_0) = 0$ and therefore $U=0$ according to (16). Since $U \geq 0$, as shown in Theorem 1, it follows that any synthesis filter that satisfies (B4) is a solution of (30), i.e. is an optimal filter. Thus, when a unity system exists it is a solution of (30) when no-quantization is applied.

3. Uniqueness of the optimal synthesis filter.

The optimal synthesis filter is the point in which the global minimum of the P.S.D. quadratic form (29) is obtained. Uniqueness of the optimal filter is provided when the matrix Q is a P.D. matrix, i.e., when $\underline{f}'Q\underline{f}>0$ for any vector $\underline{f}\neq\underline{0}$. Furthermore, in the FQ case the matrix Q is the sum of two P.S.D. matrices Q_v and Q_h, thus if either one of them is P.D., then Q is also a P.D. matrix. For simplicity we restrict the discussion to the case in which $G(f)=1$. In this case from (A1), (16) and (29) follows that:

$$\underline{f}'Q\underline{f} = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{s=-\infty}^{\infty}\sum_{t=-\infty}^{\infty} f(m-sR)f(m-tR)E[\hat{x}_{sR}((m)_M)\hat{x}_{tR}((m)_M)]$$
$$= \frac{1}{N}\sum_{m=0}^{N-1} E[(\sum_{s=-\infty}^{\infty} f(m-sR)\hat{x}_{sR}((m)_M))^2] \qquad \text{(B5)}$$

Assume that there exists a non-zero vector $\underline{f}_0\neq\underline{0}$ such that $\underline{f}'_0Q\underline{f}_0=0$. Since $\underline{f}_0\neq\underline{0}$, there is at least one value $0\leq m_0\leq(N-1)$ for which the set $\{f_0(m_0-sR)\}_{s=-\infty}^{\infty}$ contains a non-zero value. Let $f_0(m_0-s_0R)\neq0$ be the last non-zero value in this set. Since the synthesis filter is of length of L_f samples, $f_0(m_0-sR)=0$ for $s\not\in\{s_0-\frac{L_f}{R},\ldots,s_0\}$. Now, due to (B5), since $\underline{f}'_0Q\underline{f}_0=0$, in particular:

$$0 = E[(\sum_{s=-\infty}^{\infty} f_0(m_0-sR)\hat{x}_{sR}((m_0)_M)^2]$$
$$= f_0(m_0-s_0R)^2\, E[(\hat{x}_{s_0R}((m_0)_M)-\sum_{s=s_0-L_f/R}^{s_0-1}\left[\frac{-f_0(m_0-sR)}{f_0(m_0-s_0R)}\right]\hat{x}_{sR}((m_0)_M))^2] \qquad \text{(B6)}$$

And therefore $\hat{x}_{sR}((m_0)_M)$ is a degenerate process.

We have thus proven that if $\hat{x}_{sR}(m)$ is a nondegenerate process for any $0\leq m\leq(M-1)$, then there is a unique optimal synthesis filter. This is the

uniqueness criterion we stated for the MQ case. In the FQ case, by substituting (1),(6) and (7) in (B5) and using the definition of Q_v and Q_h from Section V we obtain:

$$\underline{f}'Q_v\underline{f} = \frac{1}{N}\sum_{m=0}^{N-1} E[\sum_{s=-\infty}^{\infty} f(m-sR)v_{sR}((m)_M)^2] \quad \text{(B7a)}$$

$$\underline{f}'Q_h\underline{f} = \frac{1}{N}\sum_{m=0}^{N-1} E[\sum_{s=-\infty}^{\infty} f(m-sR)x_{sR}((m)_M)^2] \quad \text{(B7b)}$$

A similar argument guarantees that when $v_{sR}(m)$ is a nondegenerate process for any $0 \le m \le (M-1)$, then Q_v is a P.D. matrix and therefore there is a unique optimal synthesis filter. This is one of the two conditions for uniqueness we have stated in Section III. When no quantization is applied, then $Q_v = 0$ and this condition is not satisfied. However, at least for an uncorrelated input signal (i.e., $\rho(d) = 0$, for $d \ne 0$), equation (B7b) is essentially:

$$\underline{f}'Q_h\underline{f} = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{r=-\infty}^{\infty}\left(\sum_{s=-\infty}^{\infty} f(m-sR)h(sR-m-M(r-r_o))\right)^2 \rho(0) \quad \text{(B8)}$$

Thus, $\underline{f}'Q_h\underline{f} = 0$, iff the output of the A/S system is identically zero regardless of the input sequence. For $M=R$, this implies that $\underline{f} = \underline{0}$, provided that non of the polyphases of the analysis filter is identically zero. Thus, for $M=R$, at least for an uncorrelated input signal, there is a unique optimal synthesis filter, provided that all the M polyphases of the analysis filter are non-zero.

4. The limiting synthesis filter for noise level approaching zero:

We consider the case of a characteristic noise statistic $\Psi_{m,n}(dR)$ for which $v_{sR}(m)$ is a nondegenerate process for $0 \le m \le (M-1)$, i.e., the matrix Q_v is a P.D. matrix, and so is the matrix $Q = Q_h + \varepsilon Q_v$, for every $\varepsilon > 0$. Let $\underline{f}_\varepsilon$ be the unique solution of (30) for $\varepsilon > 0$, i.e.:

$$\{(Q_h + Q'_h) + \varepsilon(Q_v + Q'_v)\}\, \underline{f}_\varepsilon = \underline{b} \qquad \text{(B9)}$$

Let A denote the matrix $(Q_h + Q'_h)$ and B denote the matrix $(Q_v + Q'_v)$. Both A and B are P.S.D. symmetric matrices, with B being a P.D. matrix. Thus, B has a non-singular real root C (i.e.: $B = CC'$, where $C = U\,\Lambda^{1/2}$ with the columns of U being an orthonormal basis of the eigenvectors of B, and Λ is the corresponding diagonal P.D. matrix which contains the eigenvalues of B). Equation (B9) becomes (after multiplying from the left by C^{-1}):

$$\{C^{-1}A(C^{-1})' + \varepsilon I\}(C'\underline{f}_\varepsilon) = C^{-1}\underline{b} \qquad \text{(B10)}$$

Now $C^{-1}A(C^{-1})'$ is a symmetric matrix, thus it possess an orthonormal basis of eigenvectors, i.e.: $C^{-1}A(C^{-1})' = VDV'$ with V an orthonormal matrix, and D a diagonal matrix. Therefore:

$$\underline{f}_\varepsilon = (V'C^{-1})'(D + \varepsilon I)^{-1}(V'C^{-1})\underline{b} \qquad \text{(B11)}$$

We now assume that the analysis filter is chosen such that a unity system exists, i.e. there are (many) solutions to the system of linear equations $A\underline{x} = \underline{b}$, or alternatively $\underline{b} \in Image\{A\}$. Using $A = CVDV'C'$ (due to the definition of V and D) it is easily verified that $V'C^{-1}\underline{b} \in Image\{DV'C'\}$. Therefore, since D is a diagonal matrix, $d_{ii} = 0$ implies $\{(V'C^{-1})\underline{b}\}_i = 0$. Thus, $\lim_{\varepsilon \to 0} (D + \varepsilon I)^{-1}(V'C^{-1})\underline{b}$ exists and equals to $D^{\div}(V'C^{-1})\underline{b}$, where $D^{\div}$ denotes the Moore-Penrose (M-P) inverse of D which is obtained by replacing d_{ii} with $1/d_{ii}$ whenever $d_{ii} \neq 0$ [67]. Thus:

$$\lim_{\varepsilon \to 0} \underline{f}_\varepsilon = (C^{-1})'VD^{\div}V'C^{-1}\underline{b} = (C^{-1})'(C^{-1}A(C^{-1})')^{\div}C^{-1}\underline{b} \qquad \text{(B12)}$$

Where the second equality in (B12) follows from the relation between the Moore-Penrose Inverse and the UDV Theorem [67]. Equation (B12) not only shows that $\lim_{\varepsilon \to 0} \underline{f}_\varepsilon$ exists, but also suggests a direct method for its computation. The M-P

Inverse, as well as the real root of a P.D. matrix, are efficiently evaluated by singular value decomposition techniques [60]. Let $\underline{f}_0$ denote $\lim_{\varepsilon \to 0} \underline{f}_\varepsilon$. We shall now interpret $\underline{f}_0$ in terms of the filter design parameters (i.e., the matrices Q_v and Q_h). From the properties of the M-P Inverse follows that $\underline{x}_0 = C'\underline{f}_0$ is the solution with minimal Euclidian norm of the system of linear equations:

$$C^{-1}A(C^{-1})'\underline{x} = C^{-1}\underline{b} \tag{B13}$$

Thus, $\underline{f}_0$ is the solution of the system of linear equations $A\underline{f} = \underline{b}$ which minimizes $\underline{f}'B\underline{f}$. we interpret $\underline{f}_0$ as the synthesis filter that guarantees a unity system, and among all unity systems minimizes the error due to quantization for the given characteristic noise statistic, i.e:

$$\underline{f}_0 = \underset{(Q_h + Q'_h)\underline{f} = \underline{b}}{\text{Argmin}} \quad \{\underline{f}'(Q_v + Q'_v)\underline{f}\} \tag{B14}$$

5. **Example** 1:

We substitute in (B2a) $g(d) = \delta(d)$ (since $G(f) = 1$) and $\rho(d) = \delta(d)\sigma_x^2$ and the summations over d and r collapse into a single element $d=0$, $r=r_0=1$, thus leading to $b(t)=2h(M-t)\sigma_x^2$. Substituting these values and $\Psi_{m,n}(dR) = \delta(d)\,\hat{\Psi}((m-n)_M)$ into (B2b) we notice that $Q(t,s)$ equals zero unless $(t-s)_R=0$. For $s=t+Rr$ it follows that:

$$Q(t,t+rR) = \sigma_x^2 \sum_{n=-\infty}^{\infty} h(Mn-t-rR)h(Mn-t) + \hat{\Psi}(0)\delta(r) \tag{B15}$$

Now, since $L_h = L_f = M$, the summation over n in (B15) can be replaced by a single element $n=1$, and therefore in this example (30) becomes:

$$h(M-t)\sum_{r=-\infty}^{\infty} h(M-t-rR)f(t+rR) + \left[\frac{\hat{\Psi}(0)}{\sigma_x^2}\right]f(t) = h(M-t) \quad 0\le t\le(M-1) \qquad \text{(B16)}$$

It is easily verified that the unique solution of (B16), for $\hat{\Psi}(0) > 0$, is given by (31). For $\hat{\Psi}(0) = 0$, the solution of (31) is the limiting synthesis filter for noise level approaching zero, where the noise is characterized as a white noise.

6. Example 2:

When we substitute into (B2) $R=M=N$ and $g(d) = \delta(d)$, we obtain:

$$b(t) = 2\sum_{x=-\infty}^{\infty} h(Mr_o-(t+Mx))\rho(Mx) \qquad \text{(B17a)}$$

$$Q(t,s) = \{\sum_{r=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} h(Mn-t)h(Mr-s)\rho(M(n-r)) + \qquad \text{(B17b)}$$

$$+ \Psi_{(t)_M,(t)_M}(s-t)\}\ \delta((s-t)_M=0)$$

It follows from (B17b) that the set of L_f linear equations given in (30) is decomposable into M sets, each of them of $\frac{L_f}{M}$ equations. Using the notations of $f_\tau(x), h_\tau(x)$ and $\tilde{\rho}(d) = \rho(Md)$ we can rewrite (30) as following:

$$\sum_{d=-\infty}^{\infty} h_\tau(x+d)\tilde{\rho}(d) = \qquad \text{(B18)}$$

$$= \sum_{y=-\infty}^{\infty} f_\tau(x-y)\frac{1}{2}\{Q(\tau+xM,\tau+(x-y)M) + Q(\tau+(x-y)M,\tau+xM)\}$$

Rearranging (B17b) we obtain:

$$\frac{1}{2}\{Q(\tau+xM,\tau+(x-y)M)+Q(\tau+(x-y)M,\tau+xM)\} =$$

$$\frac{1}{2}[\sum_{r=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} h_\tau(r-y)h_\tau(r-l)\tilde{\rho}(l) + \Psi_{\tau,\tau}(-yM)] \quad \text{(B19)}$$

$$+\frac{1}{2}[\sum_{r=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} h_\tau(r+y)h_\tau(r-l)\tilde{\rho}(l) + \Psi_{\tau,\tau}(yM)]$$

Equations (32),(33) follow from (B18),(B19) using the even symmetry of the sequences $\tilde{\rho}(d)$ and $\Psi_{\tau,\tau}(dM)$. Equation (34) follows directly from (32),(33) using the well-known property that convolution of sequences corresponds to multiplication of their Fourier transforms. However, this equation holds only for $L_f = \infty$ and noncasual synthesis filters since for finite values of L_f, equation (32) holds only for a finite range of x.

7. Complexity of the design for the FQ case:

The design of optimal synthesis filter is composed of two major steps. The first step is the evaluation of the matrix Q and the vector $\underline{b}$ and the second step is the solution of (30) (i.e.: inverting $Q + Q'$). In general the complexity of the solution of (30) is $O(L_f^3)$, since it involves the inversion of an $L_f \times L_f$ matrix. However, when $G(f) = 1$ the set of L_f linear equations in (30) is decomposable into R sets, each of them of about L_f / R equations (as can be easily verified from (B2b)), thus reducing the complexity of the second step of the design to $O((L_f / R)^3 R)$. For $G(f) = 1$ and $R=M$ each set of $\frac{L_f}{M}$ equations is represented by a Toeplitz system of linear equations (as can be verified from (32)), and the complexity is further reduced to $O((\frac{L_f}{M})^2 M)$. The first step is dominated by the complexity of evaluating the matrices Q_h and Q_v. The value $Q_v(t,s)$ depends only on $(t-s)$ and $(t)_R$, thus there are only $2L_f R$ different elements. For each

element the summation over d in (B2b) involves $(2L_g-1)/R$ different values of $g(\cdot)$. The inner summation over m involves $(\frac{N}{R})$ different values, thus the overall complexity of evaluating Q_v is $O\left[(2L_g-1)2L_f\ \frac{N}{R}\right]$. The value $Q_h(t,s)$ depends only on $(t-s)$ and $(t)_M$, thus there are only $2L_f M$ different elements. For each element the summations over r and n in (B2b) involve $(L_h/M)^2$ different pairs of values of $h(\cdot)$, and the summation of d involves $(2L_g-1)/R$ different values of $g(\cdot)$, thus the overall complexity of evaluating Q_h is $O((2Lg-1)2L_f\ \frac{L_h^2}{MR})$. By careful investigation of (B2b), the exact ranges of all the summations can be determined, and as a consequence the total number of different covariance values used in the design process can also be determined, as well as the maximal distance (in samples) for which stationarity is assumed. The details of these results, and of certain possible reductions in the complexity of evaluating the matrices Q_h and Q_v are summarized in Table D.1.

8. Simplified equations for the MQ case when $M=R$, $G(f)=1$:

When $M=R$ and $G(f)=1$, equation (B3) can be significantly simplified and yields:

$$b(t) = \frac{2}{B}\sum_{j=0}^{B-1}\sum_{l=1}^{\hat{L}} P^{(l)}\, c_j^{(l)}((t)_M)\ G^{(l)}(t+(j-r_o)M) \tag{B20a}$$

$$Q(t,s) = \frac{1}{B}\sum_{j=0}^{B-1}\sum_{l=1}^{\hat{L}}\sum_{k=1}^{\hat{L}} c_{(j+\frac{t-s}{M})_B}((t)_M)\ c_j^{(k)}((t)_M)F^{(l),(k)}\left(\left\lfloor\frac{j+\frac{t-s}{M}}{B}\right\rfloor\right)\delta((s-t)_M=0)$$

(B20b)

For $L_f = M$ the matrix Q is a diagonal matrix, and equation (35) follows from substituting (B20) in (30). For $L_f > M$ there is no closed form solution, but

yet the set of L_f linear equations is decomposable into M sets of about (L_f / M) equations each, and each of these sets is represented by a Toeplitz matrix. Therefore, the complexity of solving (30) is $O((L_f / M)^2 M)$. The design process involves $\hat{L}^2(L_f / MB)$ different values of $F^{(l),(k)}(d)$ ($\hat{L}^2$ pairs of $(l),(k)$ and (L_f / MB) values of d), and $\hat{L}(L_f + B)$ different values of $G^{(l)}(d)$ ($\hat{L}$ values of (l) and $(B + L_f)$ values of d). Evaluation of each element of the vector $\underline{b}$ via (B20a) involves $B\hat{L}$ operations, whereas for evaluating each element of Q, $B\hat{L}^2$ operations are needed. Since Q contains only L_f distinct elements, the overall complexity of evaluating Q and $\underline{b}$ is $O(B\hat{L}^2 L_f + B\hat{L} L_f)$. The complexity analysis of the more general cases in which either $R \neq M$ or $G(f) \neq 1$, is omitted here for simplicity.

VII. Derivation of the results presented in section IV.

1. Expressions for $\tilde{Q}_D$, $\tilde{Q}_f$, $\underline{\tilde{b}}_D$, $\underline{\tilde{b}}_f$ in equation (36):

The expressions for the matrices $\tilde{Q}_D$ and $\tilde{Q}_f$, of dimensions $L_h \times L_h$, and for the vectors $\underline{\tilde{b}}_D, \underline{\tilde{b}}_f$, of dimension L_h, are obtained by substituting (A1-A2) into (16), combining it with (28), and rearranging the result as a quadratic form in terms of the analysis window coefficients. For simplification we omit here the details of this easy (although lengthy) derivation, and only state the final results, which are:

$$\tilde{b}_D(t) = 2\,\mathrm{Re}\{ \int_{-0.5}^{0.5} W(f)\, D(f)\, e^{j2\pi f t}\, df \} \tag{C1a}$$

$$\tilde{b}_f(t) = \frac{1}{R} \sum_{r=-\infty}^{\infty} f(Mr-t) \sum_{d=-\infty}^{\infty} g(d)\, (\rho(d-M(r-r_0)) + \rho(-d-M(r-r_0))) \tag{C1b}$$

$$\tilde{Q}_D(t,s) = \{ \int_{-0.5}^{0.5} W(f)\, e^{j2\pi f (t-s)}\, df \} \tag{C1c}$$

$$\tilde{Q}_f(t,s) = \frac{1}{R} \sum_{r=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(Mr-s) f(Mn-t) \cdot \sum_{d=-\infty}^{\infty} \rho(M(n-r)+(s-t)+Rd)\, g(s-t+Rd) \tag{C1d}$$

The indices (t,s) in (C1) are limited to a finite interval of length L_h, for which the coefficients of the analysis window are non-zero. The summations over the indices (r,n) are therefore finite, but the summations over d may be infinite, in which case they can be alternatively evaluated in the frequency domain. When $W(f)$ is positive on a set of non-zero measure it follows from (C1c) (and the assumption that $W(f) \geq 0$) that the matrix $\tilde{Q}_D$ is a P.D. matrix. Since according

to Theorem 1, $U \geq 0$, $\tilde{Q}_f$ is a P.S.D. matrix. Therefore $(\tilde{Q}_D + \tilde{Q}_f)$ is a P.D. matrix, and in particular $(\tilde{Q}_f + \tilde{Q}'_f + \tilde{Q}_D + \tilde{Q}'_D)$ is non-singular, so that existence and uniqueness of the solution of (36) are guaranteed. All the facts mentioned in Section IV in the discussion below (36) are immediate consequence of the expressions given in (C1).

2. Proof of Theorem 2:

In the proof of Theorem 2 we frequently denote by x a pair of filters $(\underline{f},\underline{h})$ corresponding to an A/S system, and use the notation $x_{r+1} \in A_t(x_r)$ to denote the r-th iteration of the iterative algorithm (where $x_r \triangleq (\underline{f}^{(r)},\underline{h}^{(r)})$). The assumption that $W(f) > 0$ on a set of non-zero measure guarantees that Q_D is a P.D. matrix, and that $\underline{h}^{(r+1)} = A_1(\underline{f}^{(r)})$ is defined by a point to point mapping.

(A). Consider: $\hat{U}(x_{r+1}) = \hat{U}(\underline{f}^{(r+1)},\underline{h}^{(r+1)}) \leq \hat{U}(\underline{f}^{(r)},\underline{h}^{(r+1)}) \leq \hat{U}(\underline{f}^{(r)},\underline{h}^{(r)}) = \hat{U}(x_r)$, where the two inequalities follow from the definitions of steps 3 and 2 of the algorithm and the definitions of $A_2(\cdot)$ and $A_1(\cdot)$. Equality is obtained iff $\underline{h}^{(r)}=\underline{h}^{(r+1)} = A_1(\underline{f}^{(r)})$ and $\underline{f}^{(r)} \in A_2(\underline{h}^{(r+1)}) = A_2(\underline{h}^{(r)})$, thus iff the algorithm reaches a fixed point.

(B). According to the definition of $\Gamma(\Gamma = \{x_r; A_t(x_r) = \{x_r\}\})$, $(\underline{f},\underline{h}) \in \Gamma$ iff $\underline{h} = A_1(\underline{f})$ and $\underline{f} \in A_2(\underline{h})$, i.e., iff the A/S system defined by $(\underline{f},\underline{h})$ satisfies both (30) and (36), which are exactly the gradient equations of the error criterion $\hat{U}$. Thus, $\Gamma = \{\, x \; ; \nabla\hat{U}(x) = 0\}$. A point $x^* \in \Gamma$ is a stable point of A_t, iff there exists a ball around x^* such that for any initial condition x_0 contained in this ball there exists an instance of A_t starting from x_0 with x^* as one of its limit points. Let x^* be a stationary point which is not a local minimum of $\hat{U}$, then there is a descent direction of $\hat{U}$ at x^*, and therefore in every ball around x^* there exists a point x_0 such that $\hat{U}(x_0) < \hat{U}(x^*)$. Since $\hat{U}(x)$ is a descent function for A_t, for any instance $\{x_r\}_{r=0}^{\infty}$ starting at x_0, $\hat{U}(x_r) \leq \hat{U}(x_0) < \hat{U}(x^*)$, and any such instance does

not have x^* as one of its limit points.

Before proceding with the proofs of parts $(C)-(E)$ of the theorem we prove the following lemma:

Lemma 1: Define $\Gamma = \{x_r;\ A_t(x_r) = \{x_r\}\}$, i.e., Γ is the set of fixed points of A_t. Then:

(a). $\hat{U}$ is a descent function of A_t with respect to the set Γ, i.e.: $\hat{U}(x)$ is a continuous function and for every $y \in A_t(x)$, $\hat{U}(y) \leq \hat{U}(x)$ with equality only if $x \in \Gamma$.

(b). The point to set mapping A_t is a closed mapping, i.e: for every two convergent sequences $x_r \to x$ and $y_r \to y$ such that for any r: $y_r \in A_t(x_r)$, if follows that $y \in A_t(x)$.

(c). For any initial condition x_0, the sequence $\{\underline{h}^{(r)}\}_{r=0}^{\infty}$ is contained in the compact set $C(\rho)$ with $\rho \triangleq \hat{U}(x_0) - \inf U(x) - K_0$.

Proof of the lemma:

(a). Since $\hat{U}$ is a polynomial in both $\underline{h}$ and $\underline{f}$ it is a continuous function with respect to $x = (\underline{f}, \underline{h})$. We have already shown above, in part (A) of the Theorem, that for $y \in A_t(x)$, $\hat{U}(y) \leq \hat{U}(x)$, with equality iff $x \in A_t(x)$, in which case we choose $y = x$, i.e., $x \in \Gamma$. Thus, $\hat{U}$ is a descent function of A_t with respect to Γ.

(b). The point-to-point mapping $A_1(\cdot)$ is a continuous mapping, since the coefficients of the matrix $\tilde{Q}_f$ and the vector $\underline{\tilde{b}}_f$ are continuous functions of $\underline{f}$, and the determinant of the P.D. matrix $(Q_D + Q'_D + \tilde{Q}_f + \tilde{Q}'_f)$, is bounded away from zero for every $\underline{f} \in \mathbb{R}^{L_f}$. (It is well-known that the solution of a non-degenerate set of linear equations $Ax = b$ with $\det(A)$ bounded away from zero, is a continuous function of the coefficients of A and $\underline{b}$ [72]). Any point to point continuous mapping is a closed mapping, thus in

particular A_1 is a closed mapping. We now show that A_2 is a closed mapping, and this guarantees that A_t is a closed mapping, since $A_t(\underline{f},\underline{h}) = (A_2(A_1(\underline{f})), A_1(\underline{f}))$, where it is well-known that the composition of a point to point continuous mapping and a closed mapping is also a closed mapping [68]. Let $\underline{h}_k \to \underline{h}$ and $\underline{f}_k \to \underline{f}$ where $\underline{f}_k \in A_2(\underline{h}_k)$, i.e.: $\underline{f}_k$ is a solution of (30) for the analysis window $\underline{h}_k$.

Consider the value of $\lim_{k\to\infty} \underline{\Delta}_k$, where $\underline{\Delta}_k \triangleq \{(Q_{h_k} + Q'_{h_k} + Q_v + Q'_v)\underline{f}_k - \underline{b}_k\}$. Since the coefficients of Q_h and $\underline{b}$ are continuous functions of $\underline{h}$, it is easily verified that $\lim_{k\to\infty} \underline{\Delta}_k = \{(Q_h + Q'_h + Q_v + Q'_v)\underline{f} - \underline{b}\}$. However, since $\underline{f}_k \in A_2(\underline{h}_k)$, $\underline{\Delta}_k = \underline{0}$ for every value of $k=1,\ldots,$ and therefore $(Q_h + Q'_h + Q_v + Q'_v)\underline{f} = \underline{b}$, i.e., $\underline{f} \in A_2(\underline{h})$. This proves the closeness of A_2.

(c). Due to (28):

$$\hat{U}(x) = U(x) + (K_0 + (\underline{h}-\underline{\tilde{h}})'Q_D(\underline{h}-\underline{\tilde{h}}) \geq \inf U(x) + K_0 + (\underline{h}-\underline{\tilde{h}})'Q_D(\underline{h}-\underline{\tilde{h}})\,.$$

Since $\hat{U}(x)$ is a descent function of the algorithm A_t, $\hat{U}(x_0) \geq \hat{U}(x_r)$ for any sequence $\{x_r\}_{r=0}^{\infty}$ starting at x_0, and therefore $\hat{U}(x_0) - \inf U(x) - K_0 \geq (\underline{h}^{(r)} - \underline{\tilde{h}})'Q_D(\underline{h}^{(r)} - \underline{\tilde{h}})$, i.e., $\underline{h}^{(r)} \in C(\rho)$. Thus, the whole sequence $\{\underline{h}^{(r)}\}_{r=0}^{\infty}$ is contained in $C(\rho)$, which is a compact set since Q_D is a P.D. matrix.

We shall use the following global convergence theorem [68]:

Theorem C1: If there exists a descent function of a closed mapping A with respect to a set of real vectors Γ, then every limit point of an instance of A which is contained in a compact set, is in the set Γ.

The following lemma is also used in the sequel:

Lemma 2: If (30) possess a unique solution for any $\underline{h} \in C(\rho)$ then the set $S(\rho) = \{(\underline{f};\underline{h}) ; \underline{f} \in A_2(\underline{h}) \cap \underline{h} \in C(\rho)\}$ is contained in a compact set.

Proof: Since $C(\rho)$ is a compact set, it only remains to show that $\underline{h} \in C(\rho)$ implies that $A_2(\underline{h})$ is contained in a compact set which does depend on $\underline{h}$. Let $\lambda_m(\underline{h})$ denote the minimal eigenvalue of $(Q+Q')$, as a function of $\underline{h}$, $\|\underline{f}\|$ denotes the Euclidian norm of a vector $\underline{f} \in A_2(\underline{h})$, and $\|\underline{b}(\underline{h})\|$ denotes the Euclidian norm of the vector $\underline{b}$ which appears in the right hand side of (30) for a given value of $\underline{h}$. Then, since for $\underline{h} \in C(\rho)$ (30) possess a unique solution, $\lambda_m(\underline{h}) > 0$ there, and $\|\underline{f}\| \leq \|\underline{b}(\underline{h})\| / \lambda_m(\underline{h})$. Now, $\lambda_m(\underline{h})$ is a continuous function of $\underline{h}$ (this follows from the continuity of λ_m with respect to the coefficients of Q, and the continuity of these coefficients with respect to $\underline{h}$), thus it possess a global minimum in the compact set $C(\rho)$, which is positive. Let $\lambda_\rho > 0$ denotes this value, then $\|\underline{f}\| \leq \|\underline{b}(\underline{h})\| / \lambda_\rho$. Investigating (B2a) we notice that $\underline{b}$ is a linear function of $\underline{h}$, thus on the compact set $\underline{h} \in C(\rho)$ the value of $\|\underline{b}(\underline{h})\|$ is bounded from above by $B < \infty$. Therefore, $\|\underline{f}\| \leq B / \lambda_\rho < \infty$, and thus the union of $A_2(\underline{h})$ over $\underline{h} \in C(\rho)$ is contained in a ball which is a compact set.

We now continue with the proof of the theorem using both lemma 1 and 2 and theorem C1:

(C). For any initial condition x_0 there exists a value ρ (defined in lemma 1 part (c)), such that $\{\underline{h}^{(r)}\}_{r=0}^{\infty} \subset C(\rho)$. Due to lemma 2 which holds for any value of ρ, the set $S(\rho)$ is contained in a compact set, and in particular any instance of the algorithm A_t which is contained in $S(\rho)$ is also contained in this compact set. Now, combining this result, parts (a),(b) of lemma 1, and Theorem C1 we obtain that any limit point of an instance of A_t is in the set Γ of fixed points of A_t. Furthermore, since every instance of A_t is contained in a compact set, it possess at least one limit point. This completes property 1 of part (C).

Property 1 of part (D) follows from the same arguments, applied only on initial conditions x_0 for which $\rho \le \rho_0$ (thus lemma 2 holds for $S(\rho)$ and in particular $\{\underline{f}^{(r)}\}_{r=0}^{\infty}$ is contained in a compact set). Assuming there exists an initial condition x_0, for which $S(\rho)$ is contained in a compact set, we shall prove that $\hat{U}$ possess a global minimum which is in the set Γ. Thus completing the proof of parts (C) and (D).

Since $\hat{U}$ is a non-negative continuous function its infimum exists, and there exists a sequence $\{x_r\}_{r=0}^{\infty}$ starting on x_0 defined above, such that $\lim_{r\to\infty} \hat{U}(x_r) = \inf \hat{U}(x)$. Furthermore, without loss of generality we can assume that for any value of $r : \hat{U}(x_r) \le \hat{U}(x_0)$. Now, if we replace every element $x_r = (\underline{f}_r, \underline{h}_r)$ of this sequence by $\tilde{x}_r = (\tilde{\underline{f}}_r, \underline{h}_r)$, with $\tilde{\underline{f}}_r \in A_2(\underline{h}_r)$, then $\inf \hat{U}(x) \le \hat{U}(\tilde{x}_r) \le \hat{U}(x_r)$ and therefore $\hat{U}(\tilde{x}_r)$ also converges to $\inf \hat{U}(x)$. Since $\hat{U}(\tilde{x}_r) \le \hat{U}(x_0)$ it follows from lemma 1 part (c) that $\underline{h}_r \in C(\rho_0)$ and therefore $\tilde{x}_r \in S(\rho_0)$. Thus, the sequence $\{\tilde{x}_r\}_{r=0}^{\infty}$ is contained in a compact set, and therefore it has at least one limit point there. Let x^* denote a limit point, and $\tilde{x}_{r_k}$ be the sub-sequence that converges to x^*. Since $\tilde{x}_{r_k} \underset{k\to\infty}{\to} x^*$, $\hat{U}(\tilde{x}_{r_k}) \underset{k\to\infty}{\to} \inf \hat{U}(x)$, and $\hat{U}(x)$ is a continuous function, $\hat{U}(x^*) = \inf \hat{U}(x)$, i.e., there exists a global minimum of $\hat{U}(x)$, which is in Γ since it is a stationary point of $\hat{U}(x)$.

(E). In lemma 1 part (c), we noticed that for any instance of A_t the sequence $\{\underline{h}^{(r)}\}_{r=0}^{\infty}$ is contained in a compact set $C(\rho)$, regardless of the uniqueness of solutions of (30). Therefore, this sequence has at least one limit point. Let $x = (\underline{f}, \underline{h})$ be a limit point of A_t, i.e., there exists an instance $\{x_r\}_{r=0}^{\infty}$ of A_t with a subsequence $x_{r_k} = (\underline{f}^{(r_k)}, \underline{h}^{(r_k)})$ such that $\underline{f}^{(r_k)} \underset{k\to\infty}{\to} \underline{f}$, and $\underline{h}^{(r_k)} \underset{k\to\infty}{\to} \underline{h}$. Since $\hat{U}(x_r)$ is a non-increasing non-negative sequence it converges to a limit $v = \lim_{r\to\infty} \hat{U}(x_r)$. In particular both the sub-sequences $\hat{U}(x_{r_k})$ and $\hat{U}(x_{r_k+1})$ converge to v. Since $x_{r_k} \underset{k\to\infty}{\to} x$, and since $\hat{U}$ is a continuous function, $\hat{U}(x_{r_k}) \underset{k\to\infty}{\to} \hat{U}(x) = v$. Now,

$\underline{f}^{(r_k)} \underset{k\to\infty}{\to} \underline{f}$ and $\underline{h}^{(r_k+1)} = A_1(\underline{f}^{(r_k)})$ is a continuous function of $\underline{f}^{(r_k)}$, therefore $\underline{h}^{(r_k+1)} \underset{k\to\infty}{\to} A_1(\underline{f})$, and since $\hat{U}(x)$ is continuous $\lim_{k\to\infty} \hat{U}((\underline{f}^{(r_k)}, \underline{h}^{(r_k+1)})) = \hat{U}(\underline{f}, A_1(\underline{f}))$. However, by the definition of A_1 and A_2: $\hat{U}(x_{r_k+1}) \le \hat{U}((\underline{f}^{(r_k)}\underline{h}^{(r_k+1)})) \le \hat{U}(x_{r_k})$, and therefore $\hat{U}(\underline{f}, A_1(\underline{f})) = v = \hat{U}(\underline{f}\,\underline{h})$. This implies that $\underline{h} = A_1(\underline{f})$. Now, $\underline{f}^{(r_k)} \in A_2(\underline{h}^{(r_k)})$, and since A_2 is a closed mapping then $\underline{f} \in A_2(\underline{h})$. Therefore, $A_t(x) = \{x\}$ i.e., $x \in \Gamma$.

3. The rate of convergence:

Since $(\underline{f}^* \underline{h}^*) \in \Gamma$ they satisfy equations (30) and (36). Furthermore, $\underline{h}^{(n+1)}$ is a solution of (36), with respect to $\underline{f}^{(n)}$, and $\underline{f}^{(n+1)}$ is a solution of (30), with respect to $\underline{h}^{(n+1)}$ (due to the definition of A_1 and A_2). Substructing these two pairs of equations one from the other and rearranging the result we obtain:

$$\mathbf{B}(\underline{h}^{(n+1)} - \underline{h}^*) = (\underline{\tilde{b}}_{f^{(n)}} - \underline{\tilde{b}}_{f^*}) - (\tilde{Q}_{f^{(n)}} + \tilde{Q}'_{f^{(n)}})\underline{h}^{(n+1)} + (\tilde{Q}_{f^*} + \tilde{Q}'_{f^*})\underline{h}^{(n+1)} \quad \text{(C2a)}$$

$$\mathbf{A}(\underline{f}^{(n+1)} - \underline{f}^*) = (\underline{b}_{h^{(n+1)}} - \underline{b}_{h^*}) - (Q_{h^{(n+1)}} + Q'_{h^{(n+1)}})\underline{f}^{(n+1)} + (Q_{h^*} + Q'_{h^*})\underline{f}^{(n+1)} \quad \text{(C2b)}$$

We now expand the right hand side of (C2) in a Taylor series around $(\underline{f}^* \underline{h}^*)$ and obtain:

$$\mathbf{B}(\underline{h}^{(n+1)} - \underline{h}^*) = \; -\mathbf{C}' \; (\underline{f}^{(n)} - \underline{f}^*) + g_1(\underline{f}^{(n)} - \underline{f}^*, \underline{h}^{(n+1)} - \underline{h}^*) \quad \text{(C3a)}$$

$$\mathbf{A}(\underline{f}^{(n+1)} - \underline{f}^*) = \; -\mathbf{C} \; (\underline{h}^{(n+1)} - \underline{h}^*) + g_2(\underline{f}^{(n+1)} - \underline{f}^*, \underline{h}^{(n+1)} - \underline{h}^*) \quad \text{(C3b)}$$

Where both g_1 and g_2 are $o(\underline{\delta x}^{(n)})$. Since we assumed that (30) and (36) possess a unique solution at $(\underline{f}^* \underline{h}^*)$,$\mathbf{A}^{-1}$,$\mathbf{B}^{-1}$ exists and (37-39) follow immediately from (C3). The left inequality in (40) is obtained by explicitly evaluating $\|\underline{f}^{(n+1)} - \underline{f}^*\|^2$ and $\|\underline{h}^{(n+1)} - \underline{h}^*\|^2$ using (37,38), ignoring all the $o(\underline{\delta x}^{(n)})$ terms, and using the well-known inequality $\|\mathbf{D}\underline{x}\|^2 / \|\underline{x}\|^2 \le \lambda_{MAX}(\mathbf{D}'\mathbf{D})$ which holds for any

matrix

D [69]. The right inequality in (40) is due to the fact that both **A** and **B** are P.D. symmetric matrices, and to certain elementary inequalities regarding the eigenvalues of these matrices [69].

The explicit expression for C_{ij} which is obtained from (39c) and (B2) after some simple (although tedious) algebraic manipulations is:

$$C_{ij} = 2 \sum_{r=-\infty}^{\infty} f^*(Mr-j) \sum_{n=-\infty}^{\infty} h^*(Mn-i) \sum_{k=-\infty}^{\infty} g(kR-Mn+i+j)\rho(kR-Mr+i+j)$$

$$+ 2\delta((i+j)_M = 0) \sum_{d=-\infty}^{\infty} g(d) \cdot \tag{C4}$$

$$\cdot \{ \sum_{n=-\infty}^{\infty} \rho(d+Mn-i-j) \sum_{s=-\infty}^{\infty} f^*(i-d-sR)h^*(sR-i+d-Mn) - \rho(d+Mr_0-i-j)\}$$

Specifications	Complexity of Inverting $Q+Q'$	Complexity of Evaluating $\underline{b}$ and Q	Maximal Distance for $\rho(\cdot)$	Maximal Distance for $\psi(\cdot)$	Total Number of Values of ρ Used	Total Number of Values of ψ Used
(1) The General Case	L_f^3	$\frac{(2L_g-1)2L_f}{MR}(L_h^2+NM)$	$L_g+L_f+L_h-3$	$(L_g+L_f-2)/R$	$L_g+L_f+L_h-2$	$M^2(L_g+L_f-1)/R$
(2) $G(f) = 1, (L_g = 1)$	$(L_f/R)^3R$	$\frac{2L_f}{MR}(L_h^2+NM)$	L_f+L_h-2	$(L_f-1)/R$	$(L_f+L_h-1)/M$	$M\ L_f/R$
(3) $\rho(d) = \delta(d)$	L_f^3	$\frac{2L_f}{NR}(L_h^2R+N^2(2L_g-1))$	0	$(L_g+L_f-2)/R$	1	$M^2(L_g+L_f-1)/R$
(4) $\forall t,s$, $\underline{V}_{tR}$, $\underline{V}_{sR}$ uncorrelated	L_f^3	$\frac{2L_f}{MR}(L_h^2(2L_g-1)+MNR)$	$L_g+L_f+L_h-3$	0	$L_g+L_f+L_h-2$	M^2
(5) T = DFT; $\forall_{k,\ell}$, $\underline{V}_{sR}(k)$, $\underline{V}_{sR}(\ell)$ uncorrelated	L_f^3	$\frac{(2L_g-1)2L_f}{MR}(L_h^2+M)$	$L_g+L_f+L_h-3$	$(L_g+L_f-2)/R$	$L_g+L_f+L_h-2$	$M(L_g+L_f-1)/R$
(6) M = R	L_f^3	$(2L_g-1)2L_f\left(\frac{L_h}{M}\right)^2$	$L_g+L_f+L_h-3$	$(L_g+L_f-2)/M$	$L_g+L_f+L_h-2$	$M(L_g+L_f-1)$
(2) + (6)	$(L_f/M)^2M$	$2\left(\frac{L_h}{M}\right)\max(L_f,L_h)$	L_f+L_h-2	$(L_f-1)/M$	$(L_f+L_h-1)/M$	L_f
(2) + (3) + (6)	$(L_f/M)^2M$	$2L_fL_h/M$	0	$(L_f-1)/M$	1	L_f
(2) + (3) + (4) + (6)	$(L_f/M)^2M$	$2L_fL_h/M$	0	0	1	M
(2) + (3) + (4) + (5) + (6)	$(L_f/M)^2M$	$2L_fL_h/M$	0	0	1	1

Table D-1: Complexity Analysis for FQ

נספח ה': סינתזה אופטימלית של טרנספורם לזמן קצר שעבר מודיפיקציה

I. הגדרת מושגים

תהא נתונה סדרה חצי-אינסופית של דגימות $\{x(n)\}_{n=0}^{\infty}$ (קומפלקסיות **או** ממשיות), ומסנן אנליזה סיבתי שתגובת דגם היחידה שלו היא $\{h(n)\}_{n=0}^{\infty}$. מערכת האנליזה מוגדרת על ידי מסנן האנליזה, מטריצה רגולרית T, $M \times M$ ממדית ושני פרמטרים R ו-M, לאורך כל הדיון נניח ש- $M \geq R$ (זהו המקרה המעניין), וכן ש- $h(0) = 0$ (נחוץ למטרות סינכרון של מוצא המערכת לוקטורים שלמים).

<u>הגדרה</u>: ה-DSTV (Discrete-Short-Time-Vectors) של הסדרה $x(n)$ יהא סדרת הוקטורים $\{\underline{x}_{sR}\}_{s=1}^{\infty}$ מאורך M המוגדרים על ידי:

$$\mathbf{x}_{sR}(n) = \sum_{r=-\infty}^{\infty} h(sR-n-Mr) \quad x(n+Mr) \qquad 0 \leq n \leq M-1 \tag{1}$$

כשב-(1) מתייחסים לסדרות $x(\cdot)$, $h(\cdot)$ כאל סדרות דו-צדדיות שנוצרו מהסדרות החד-צדדיות על ידי הוספת אפסים.

את ההעתקה מהסדרה $x(\cdot)$ ל-DSTV נסמן על ידי $\tilde{H}$, ונכנה אותה העתקת האנליזה.

<u>הגדרה</u>: ה-DSTT (Discrete-Short-Time-Transforms) של הסדרה $x(n)$ יהא סדרת הוקטורים $\{\underline{X}_{sR}\}_{s=1}^{\infty}$ מאורך M, הנוצרת מה- DSTV על ידי:

$$\underline{X}_{sR} = T\underline{x}_{sR} \tag{2}$$

כש-T הוא אופרטור ליניארי רגולרי המיוצג על ידי מטריצה $M \times M$ ממדית.

את ההעתקה מה-DSTV ל-DSTT נסמן על ידי $\tilde{T}$, ונכנה אותה העתקת הטרנספורם.

<u>הגדרות</u>: נכנה בשם MDSTV(Modified DSTV), ובשם MDSTT(Modified DSTT) , סדרות של וקטורים $\{\hat{\underline{x}}_{sR}\}_{s=1}^{\infty}$ ו- $\{\hat{\underline{X}}_{sR}\}_{s=1}^{\infty}$ (בהתאמה) של וקטורים M ממדיים שלאו דוקא נוצרו ע"י (1) ו/או (2) מסדרת דגימות. בדרך כלל ה-MDSTT (וה-MDSTV) יתקבלו מ-DSTT על ידי מודיפיקציה כלשהי (ראה תאור מערכות A/S בפרק 1).

<u>הגדרה</u>: "כמעט מערכת יחידה" aUS (almost Unity System) , זו מערכת שמקבלת כקלט את סדרת ה-DSTT ועוד מספר סופי של דגימות מהסדרה המקורית (שמקומן אינו תלוי בערכי הסדרה אלא רק במסנן האנליזה) ומשחזרת את כל הסדרה המקורית $x(\cdot)$ מתוכן (בידיעת מסנן האנליזה).

<u>הגדרה</u>: מערכת יחידה US (Unity System), זו מערכת שמקבלת כקלט את סדרת ה-DSTT ומשחזרת מתוכה את הסדרה המקורית $x(\cdot)$ (בידיעת מסנן האנליזה).

<u>הערה 1</u>: בהגדרת aUS ו-US אנו מניחים שהמערכת מפיקה כל דגם של $x(\cdot)$ בזמן סופי התלוי רק באינדקס שלו בסידרה (בפרט זה מכיל את כל המערכות בעלות השהייה סופית). ברור שמערכות שאינן מפיקות פלט תוך זמן סופי אינן מעניינות מבחינה מעשית.

<u>הגדרה</u>: מערכת יחידה בזמן סופי ftUS (finite time Unity System), זו מערכת US שמקיימת גם את הכלל הבא: מתוך רישא סופי של IL וקטורי ה-DSTT $\{\underline{x}_{sR}\}_{s=1}^{IL}$, ניתן כבר לשחזר את הרישא של דגמי הסדרה $\{x(n)\}_{n=o}^{ILR-1}$, וזאת לכל $L \geq 1$.

<u>כאשר</u>: $N \triangleq RM/g$, $I \triangleq M/g$, $J \triangleq R/g$ ו-$g = \gcd(R,M)$.

<u>הערה 2</u>: בברור {קבוצת ה-ftUS} $\subseteq$ {קבוצת ה-US} $\subseteq$ {קבוצת ה-aUS}.

<u>הגדרה</u>: מערכת יחידה דואלית DUS (Dual Unity System), זו מערכת שלכל סדרת MDSTT מפיקה סדרת דגימות $\{x(n)\}_{n=o}^{\infty}$, כך שה-DSTT של סדרה זו מזדהה עם ה-MDSTT הנתון.

<u>II. תנאים לקיום מערכות יחידה לסוגיהן השונים</u>

<u>טענה 1</u>: קיומן של ftUS, aUS, US ו-DUS הוא ב"ת בטרנספורם T ולכן ללא הגבלת הכלליות ניתן להניח $T = I$ (קרי, לדון בשחזור מתוך DSTV).

<u>הוכחה</u>

הטרנספורם T הוא 1 - 1 ועל ופועל מוקטור $\underline{x}_{sR}$ לוקטור $\underline{X}_{sR}$, ולכן קיים שחזור מ-DSTT אם"ם קיים שחזור מ-DSTV. בנוסף ההעתקה $\tilde{T}$ מעבירה רישא סופית של DSTV לרישא מאותו אורך של ה-DSTT. מתוך ההגדרות של ה-ftUS, aUS, US ו-DUS והתכונות שצוינו לעיל נובעת הטענה.

<u>הערה 3</u>: ההעתקה $\tilde{H}$ המתוארת ב-(1) ניתנת לפרוק ל-M העתקות שונות שיסומנו על ידי $\tilde{H}_p$ $1 \leq p \leq M$. על מנת להגדיר את ההעתקות אלו, נסמן:

$$(3) \quad \begin{cases} x_p(r) = x(Mr-p) & \infty > r \geq 1 \\ h_p(x) = h(gx+p) & \infty > x > -\infty \\ z_p(s) = x_{sR}(M-P) & \infty > s \geq 1 \end{cases}$$

אזי (1) תיכתב כ:

$$(4) \quad z_p(s) = \sum_{r=1}^{\infty} h_p(sJ-Ir) \quad x_p(r) \qquad 1 \leq S < \infty \quad ; \quad 1 \leq p \leq M$$

ההעתקה $\tilde{H}_p$ מעתיקה את הסדרה $\{x_p(r)\}_{r=1}^{\infty}$ לסדרה $\{z_p(s)\}_{s=1}^{\infty}$ על פי (4), ועתה ההעתקה $\tilde{H}$ היא סכום ישר של העתקות אלו, קרי: $\tilde{H} \triangleq \tilde{H}_1 \oplus \dots \oplus \tilde{H}_M$.

נכנה את $\{h_p(x)\}_{x=-\infty}^{\infty}$ כ-GP (Generalized Polyphase) ה-p-י של מסנן האנליזה. ישנם M GP-ים , אך מתוכם רק g שונים והיתר הם הזזות שלהם.

<u>הגדרה</u>: מערכת תיקרא aUS_p אם מתוך דגמי הסדרה $\{z_p(s)\}_{s=1}^{\infty}$ ומספר סופי (וקבוע) של דגמי $x_p(\cdot)$, היא משחזרת את יתר הסדרה $\{x_p(r)\}_{r=1}^{\infty}$, בידיעת ה-GP ה-p-י . באופן דומה מערכת תיקרא US_p אם מתוך דגמי הסדרה $\{z_p(s)\}_{s=1}^{\infty}$ היא משחזרת את הסדרה $\{x_p(r)\}_{r=1}^{\infty}$ בידיעת ה-GP ה-p-י.
באופן דומה מערכת תיקרא $ftUS_p$ אם לכל $L \geq 1$ מתוך $\{z_p(s)\}_{s=1}^{IL}$ היא משחזרת את $\{x_p(r)\}_{r=1}^{JL}$ ומערכת תיקרא DUS_p אם לכל סדרה $\{z_p(s)\}_{s=1}^{\infty}$ קיימת סדרה $\{x_p(r)\}_{r=1}^{\infty}$ כך שההעתקה $\tilde{H}_p$ של $x_p(\cdot)$ נותנת בדיוק את $z_p(\cdot)$.

<u>טענה 2</u>: קיים aUS אם"ם קיימים aUS_p $1 \leq p \leq M$.
קיים US אם"ם קיימים US_p $1 \leq p \leq M$.
קיים ftUS אם"ם קיימים $ftUS_p$ $1 \leq p \leq M$.
קיים DUS אם"ם קיימים DUS_p $1 \leq p \leq M$.

יתר על כן, כל aUS הוא M-יה סדורה של מערכות aUS_p,

כל US הוא M-יה סדורה של מערכות US_p,

כל ftUS הוא M-יה סדורה של מערכות $ftUS_p$,

וכל DUS הוא M-יה סדורה של מערכות DUS_p.

<u>הוכחה:</u> ההוכחה נובעת מיידית מתוך טענה 1, מהעובדה ש- $\tilde{H}$ הוא סכום ישר של ההעתקות $\tilde{H}_p$ ומההגדרות של aUS_p, US_p, $ftUS_p$ ו- DUS_p.

<u>טענה 3:</u> קיים DUS_p אם"ם $I = J = 1$ ו- $h_p(0) \neq 0$.

<u>משפט 1:</u> קיים DUS אם"ם $M = R$ ו-$h(n) \neq 0$ עבור $1 \leq n \leq M$.

המשפט נובע מיידית מהטענה לעיל, על פי טענה 2 והסימונים שניתנו ב-(3).

<u>הוכחת טענה 3:</u> נתבונן ב-$z_p(1),\ldots,z_p(I)$. מאחר ו-

$$\begin{cases} x_p(r) = 0 & r \leq 0 \\ h_p(x) = 0 & x < 0 \end{cases}$$

הרי וקטור I הדגמים הראשונים של $z_p(\cdot)$ הוא פונקציה לינארית של $x_p(1),\ldots,x_p(J)$. הנחנו ש- $R \geq M$ ולכן $J \geq I$. עבור $J > I$ הרי יש תלויות לינאריות בין I הדגמים הללו.

אם נבחר סדרת I דגמים $\{\hat{z}_p(1),\ldots,\hat{z}_p(I)\}$ שאינה מקיימת את אחת התלויות הללו, הרי לא קיימת סדרה $x_p(\cdot)$ ש-$\tilde{H}_p$ שלה יזדהה עם הסדרה הנ"ל <= לא תיתכן מערכת DUS_p.

מכאן שתנאי הכרחי ל-DUS_p הוא $I = J = 1$ ואזי $M = g$ (השיוויון האחרון מיידי כי I ו- J זרים).

עבור $I = J = 1$ הרי $z_p = h_p * x_p$ וכן $h_p(x) = 0$, $x < 0$ לכל p, ואזי $z_p(1) = x_p(1)\, h_p(0)$. לכן עבור $h_p(0) = 0$ לא תיתכן מערכת DUS_p (שכן לא ניתן לקבל סדרה $\hat{z}_p(\cdot)$ המתחילה ב-$\hat{z}_p(1) \neq 0$ על ידי ההעתקה $\tilde{H}_p$). מאידך כאשר $h_p(0) \neq 0$, הרי המערכת:

$$(5) \quad x_p(r) = \frac{1}{h_p(o)} \{\hat{z}_p(r) - \sum_{t=1}^{r-1} h_p(r-t)x_p(t)\} \qquad r \geq 1$$

היא מימוש של מערכת DUS_p.

טענה 4: כל מערכת aUS_p היא מערכת לינארית (ביחס לצמד כניסותיה).

משפט 2: כל מערכת aUS, US או $ftUS$ היא מערכת לינארית.

המשפט נובע ישירות מהטענה לעיל, על פי טענה 2 והערה 2.

הוכחת טענה 4:

נתבונן במערכת aUS_p. נשחזר מהצמד $\{z_p^{(1)}(\cdot),\ \tilde{x}_p^{(1)}(\cdot)\}$ (כאשר $\tilde{x}_p^{(1)}(\cdot)$ מייצג אוסף סופי וקבוע של דגימות $x_p^{(1)}(\cdot)$) את $x_p^{(1)}(\cdot)$ על ידי המערכת. ובאופן דומה מהצמד $\{z_p^{(2)}(\cdot),\ \tilde{x}_p^{(2)}(\cdot)\}$ את הסדרה $x_p^{(2)}(\cdot)$ על ידי אותה המערכת.

עתה נתבונן בהעתקה $\tilde{H}_p$ של הסדרה: $\alpha x_p^{(1)}(\cdot) + \beta x_p^{(2)}(\cdot)$, $\alpha,\beta \in ¢$.

מאחר ו-$\tilde{H}_p$ (שניתנה ב-(4)) היא העתקה ליניארית הרי:

$$\tilde{H}_p(\alpha x_p^{(1)}(\cdot) + \beta x_p^{(2)}(\cdot)) = \alpha z_p^{(1)}(\cdot) + \beta z_p^{(2)}(\cdot)$$

מכאן שלכל $\alpha,\beta \in ¢$ הצמד $\{\alpha z_p^{(1)}(\cdot) + \beta z_p^{(2)}(\cdot),\ \alpha\tilde{x}_p^{(1)}(\cdot) + \beta\tilde{x}_p^{(2)}(\cdot)\}$ הוא קלט חוקי של המערכת, והפלט שחייב להתאים לו הוא $\alpha x_p^{(1)}(\cdot) + \beta x_p^{(2)}(\cdot)$.

זו בדיוק ההגדרה של מערכת לינארית.

טענה 5: קיים $ftUS_p$ אם"ם ניתן לשחזר את $\{x_p(J),\ldots,x_p(1)\}$ מתוך $\{z_p(I),\ldots,z_p(1)\}$.

הוכחה: שזהו תנאי הכרחי ברור מתוך הגדרת ה-$ftUS_p$ (הצבת $L = 1$ בהגדרה).

נוכיח עתה שזהו גם תנאי מספיק.

נרשום את ההעתקה $\tilde{H}_p$ על ידי מטריצה אינסופית, דהיינו:

$$\infty > s \geq 1 \quad , \quad z_p(s) = \sum_{r=1}^{\infty} \tilde{H}_p(s,r)x_p(r)$$

ואזי על-פי (4) ברור ש:

$$(6) \quad \tilde{H}_p(s,r) = h_p(sJ-rI) = \tilde{H}_p(s+kI,r+kJ) \quad \forall k$$

לכן המטריצה $\tilde{H}_p$ היא מטריצת טואפליץ של בלוקים מגודל $J \times I$. יתר על כן, זו מטריצה משולשת תחתונה בבלוקים שלה (קרי, הבלוקים שמעל "האלכסון הראשי" הם

אפס זהותית), כי עבור $r \geq Jk + 1$, $s \leq Ik$ הרי $(sJ-rI) \leq -I$ ואזי
$h_p(sJ-rI) = 0$ מהסיבתיות של מסנן האנליזה ונוסחא (3). להלן תאור גרפי של $\tilde{H}_p$:

$$\tilde{H}_p = \begin{array}{c} \\ I\updownarrow \\ I\updownarrow \\ I\updownarrow \\ I\updownarrow \\ \vdots \end{array} \left[\begin{array}{cccc} \overset{J}{\leftrightarrow} & \overset{J}{\leftrightarrow} & \overset{J}{\leftrightarrow} & \\ H_o^{(p)} & 0 & 0 & \cdots \\ H_1^{(p)} & H_o^{(p)} & 0 & \ddots \\ H_2^{(p)} & H_1^{(p)} & H_o^{(p)} & \ddots \\ H_3^{(p)} & H_2^{(p)} & \ddots & \\ \vdots & \ddots & \ddots & \end{array}\right] \tag{7}$$

התנאי של הטענה מבטיח שלתת-המטריצה $H_o^{(p)}$ יש דרגה J (ונסמן זאת על ידי
$\deg H_o^{(p)} = J$).

נסמן על ידי $\underline{u}_p(t)$ את הוקטור שאבריו הם $\{z_p(tI+1),\ldots,z_p(tI+I)\}$ ועל ידי
$\underline{v}_p(t)$ את הוקטור שאבריו הם $\{x_p(tJ+1),\ldots,x_p(tJ+J)\}$. עתה (4) ניתנת לרישום
הוקטורי הבא:

$$\underline{u}_p(t) = \sum_{q=o}^{t} H_q^{(p)}\, \underline{v}_p(t-q) \qquad t \geq 0 \tag{8}$$

קיום מערכת $ftUS_p$ שקול ליכולת לשחזר את $\{\underline{v}_p(t)\}_{t=o}^{L-1}$ מתוך $\{\underline{u}_p(t)\}_{t=o}^{L-1}$.
כאשר $\deg H_o^{(p)} = J$, הרי קיים הפיך משמאל של $H_o^{(p)}$ שנסמנו $H_o^{(p)\#}$ (ומימדו
$I \times J$), ואזי המערכת הבאה היא $ftUS_p$:

$$\underline{v}_p(t) = H_o^{(p)\#} \left\{\underline{u}_p(t) - \sum_{q=1}^{t} H_q^{(p)}\, \underline{v}_p(t-q)\right\} \qquad (L-1) \geq t \geq 0 \tag{9}$$

<u>טענה 6</u>: קיימת מערכת ftUS אם"ם לכל אחת מ-M המטריצות $A^{(p,k)}$, $0 \leq k \leq I-1$,
$1 \leq p \leq g$ שמימדן $J \times I$ יש דרגה J , כאשר:

$$A^{(p,k)}(i,j) = h(p+gk + iR - jM) \qquad 1 \leq i \leq I \tag{10}$$

$$1 \leq j \leq J$$

<u>משפט 3</u>: קיימת מערכת ftUS אס"ם $h(n) \neq 0$ $1 \leq n \leq R$.

<u>הוכחת טענה 6</u>: נובעת מיידית מטענה 5, מטענה 2 ומהסימונים של (3), כאשר המטריצה $A^{(p,k)}$ היא למעשה $H_o^{(p+kg)}$ של טענה 5.

<u>הוכחת המשפט</u>: מספיק להראות שהתנאים של משפט 3 וטענה 6 שקולים.

(א) נראה שתנאי טענה 6 <= תנאי משפט 3. לשם כך נתבונן בעמודה ה-J-ית של המטריצות $A^{(p,k)}$ עבור $1 \leq p \leq g$, $0 \leq k \leq J-1$. מהסיבתיות של $h(\cdot)$ נובע שזו עמודת אפסים, פרט לאבר האחרון שלה שהוא

$$0 \neq h(p+gk) \triangleq A^{(p,k)}(I,J) \Leftarrow 0 \neq h(n), \quad \text{לכל} \quad 1 \leq n \leq R.$$

(ב) נראה את הכיוון ההפוך. תמיד קיים $i_o(j)$ המקיים $jM < i_o(j)R + p+gk \leq R+jM$, וקל לראות שעבור $1 \leq j \leq J$, $1 \leq p + gk \leq M$ הרי $1 \leq i_o(j) \leq I$.
ברור ש-$A^{(p,k)}(i,j) = 0$ עבור $i < i_o(j)$, ומאחר ו-$M \geq R$ הרי $i_o(j+1) > i_o(j)$.
אם $h(n) \neq 0$ לכל $1 \leq n \leq R$, הרי $A^{(p,k)}(i_o(j),j) \neq 0$ ולכן לכל עמודה ב-$A^{(p,k)}$ ישנו איבר ראשון שונה מאפס במקום אחר, ולכן כל העמודות של מטריצות אלו בת"ל.

מ.ש.ל.

<u>טענה 7</u>: קיימת aUS_p אס"ם ישנו $1 \leq L$ סופי כך שמתוך $\{\underline{u}_p(k)\}_{k=o}^{L-1}$ ניתן לשחזר את $\underline{v}_p(0)$.

<u>הוכחה</u>: ברור שזהו תנאי מספיק (זהו למעשה אפילו תנאי מספיק לקיום US_p) לאור המבנה של $\tilde{H}_p$ שתואר ב-(7). הסיבה לכך היא פשוטה ← הקשר בין $\underline{u}_p(0),\ldots,\underline{u}_p(L-1)$, ל-$\underline{v}_p(0),\ldots,\underline{v}_p(L-1)$ זהה לקשר בין $\underline{u}_p(1),\ldots,\underline{u}_p(L)$ ל-$\underline{v}_p(1),\ldots,\underline{v}_p(L)$ וכו' (כי $\tilde{H}_p$ היא טואפליץ בבלוקים שלה). לכן ניתן ליצור את המקביל של (9) כדלקמן:

$$\underline{v}_p(t) = \sum_{k=o}^{L-1} H_p^{\#}(k) \left\{\underline{u}_p(t+k) - \sum_{q=1}^{t} H_{q+k}^{(p)} \underline{v}_p(t-q)\right\} \qquad t \geq 0 \tag{10}$$

כאשר וקטור L המטריצות $\{H_p^{\#}(k)\}_{k=o}^{L-1}$ שמימדן $J \times I$ מייצג את המשחזר הלינארי של $\underline{v}_p(0)$ מתוך $\{\underline{u}_p(k)\}_{k=o}^{L-1}$, וזו בברור מערכת US_p (ולכן גם aUS_p).

נראה שזהו גם תנאי הכרחי. נניח שקיימת aUS_p , אזי קיימת J-יה עוקבת (ה-t_o-ית) כך שניתן לשחזר את כולה מתוך כל הסדרה $z_p(\cdot)$, וכל דגמי $x_p(\cdot)$ שקדמו לה (מתוך הגדרת aUS_p). יתר על כן מאחר וכל דגם ב-J-יה משוחזר תוך זמן סופי (ראה הערה 1), הרי השחזור של J-יה זו מבוסס רק על $\{z_p(s)\}_{s=1}^{I\hat{L}}$ עבור $\hat{L}$ גדול מספיק. מהמבנה הטואפליצי של $\tilde{H}_p$ נובע שבאותה צורה (דהיינו אותם מטריצות $(H_p^{\#}(\cdot)$, ניתן לשחזר את ה-J-יה הראשונה (קרי $\underline{u}_p(0)$) מתוך $\{z_p(s)\}_{s=1}^{I(\hat{L}+1-t_o)}$.

מכאן שהתנאי של הטענה מתקיים.

מ.ש.ל.

<u>משפט 4</u>: כל aUS הוא גם US.

<u>הוכחה</u>: המשפט נובע מיידית מטענה 7 ומטענה 2.

<u>משפט 5</u>: עבור $J = 1$ קיים US (aUS) אם"ם $\{h(p+Rk)\}_{k=o}^{\infty}$ אינה זהותית אפס, לכל $1 \leq p \leq R$, ותנאי זה אקויולנטי ל-

$$0 \leq p \leq R-1, \sum_{n=o}^{R-1} H(Z\, \omega_R^n)\omega_R^{np} \neq 0$$

כש-$H(Z)$ היא התמרת Z של מסנן האנליזה, ו- $\omega_R \overset{\Delta}{=} e^{j(2\pi/R)}$.

<u>הוכחה</u>:

האקויולנטיות של התנאים נובעת מיידית מהגדרת התמרת Z והנוסחא

$$\frac{1}{R}\sum_{n=o}^{R-1} \omega_R^{np} = \delta(p \equiv 0 \bmod R)$$

(א) עבור $J = 1$ הרי העמודה הראשונה של $\tilde{H}_p$ היא $\{h(p+Rs-M)\}_{s=1}^{\infty}$

וברור שכאשר עבור p מסוים, $h(p+Rk) = 0 \;\; \forall k$, הרי לאותו p העמודה הראשונה של $\tilde{H}_p$ היא אפס זהותית => נוצרת סתירה ברורה לתנאי של טענה 6 ולא קיים aUS_p<= על פי טענה 2, לא קיים aUS (ולכן גם לא US).

(ב) על פי טענה 2 מספיק להוכיח שהתנאים של המשפט הם תנאים מספיקים לקיום US_p לכל p. זה בדיוק שקול להוכחה שאם $h_p(x)$ אינו זהותית אפס ו-$J = 1$ אזי קיים US_p.

יהא $h_p(x)$ שאינו זהותית אפס, דהיינו קיים x_o כך ש-$h_p(x_o) \neq 0$ ו-$h_p(x) = 0$, $x<x_o$. יתר על כן ברור ש-$x_o \geq -(I-1)$ כי $h(\cdot)$ סיבתי.
נתבונן ב- $\{z_p(x_o+rI)\}_{r=1}^{\infty}$, בברור $x_o + rI \geq 1$ ולכן אלו דגמים של הסדרה הנתונה לצורך שחזור $x_p(\cdot)$. עתה מתוך (4) נקבל:

$$x_p(r) = \frac{1}{h_p(x_o)} \left\{ z_p(x_o+rI) - \sum_{q=1}^{r-1} h_p(x_o+qI)\, x_p(r-q) \right\} \tag{11}$$

וזו נוסחאת מערכת US_p.

<u>הערה</u>: תנאי משפט 5 הוא תנאי הכרחי גם עבור $J \neq 1$, אך אזי אינו בהכרח תנאי מספיק. ההכרחיות ברורה מ-(1). אם קיים $1 \leq p_o \leq R$ כך ש- $\{h(p_o+Rk)\}_{k=o}^{\infty}$ הוא זהותית אפס, הרי $x(p_o)$ לא משפיע כלל על סדרת ה-DSTV ולכן לא ניתן לשחזור מתוכה <= אין מערכת US.

הדוגמא הבאה מראה שזהו אינו תנאי מספיק: $J = R = 2$, $I = M = 3$, $g = 1$ ומקדמי המסנן $h(n)$ הם אפס פרט ל-$h(1) \neq 0$, $h(4)$. במקרה זה נתבונן בהעתקה $\tilde{H}_2$, המטריצה שלה תיראה כדלקמן:

$$\tilde{H}_2 = \left[\begin{array}{cc|cc|cc}
h(1) & 0 & & \cdots & & \\
0 & 0 & & & & \\
0 & 0 & & & & \\
\hline
0 & h(4) & h(1) & 0 & \cdots & \\
0 & 0 & 0 & 0 & & \\
0 & 0 & 0 & 0 & & \\
\hline
0 & 0 & 0 & h(4) & h(1) & 0 \\
\vdots & & \vdots & & \vdots &
\end{array}\right]$$

וברור שניתן לשחזר רק את $h(4)x_2(2) + h(1)x_2(3)$ ללא ידיעת $x_2(2)$ ו-$x_2(3)$ בנפרד.

III. סינתזה אופטימלית (במובן של WMMSE) מתוך MDSTT סופי

מכאן ואילך נניח שקיימת מערכת ftUS, קרי ש-0 $h(n) \neq 0$ $R \geq n \geq 1$.

עבור $R = M$ הסינתזה האופטימלית ברורה מתוך משפט 1. ישנה מערכת DUS ולכן לכל MDSTT סופי נתאים את ה-MDSTV המתאים לו וממנו נייצר סדרה $x(n)$ בעלת אותו DSTV על ידי שימוש במשוואה (5) ל-$M \geq p \geq 1$.

מאידך עבור $R < M$ הסינתזה האופטימלית היא בעלת שגיאה חיובית (כי אין DUS) ולא ברור איך מבצעים אותה.

ראשית נגדיר שתי בעיות סינתזה שונות ונראה שהן אקויולנטיות מבחינה מתמטית. נפתח מערכת משוואות ליניאריות שפתרונן (היחיד) נותן את סדרת הדגימות המתאימה לסינתזה האופטימלית, ונדון אחר כך במימוש יעיל של הפתרון קרי בסכמת הסינתזה.

<u>בעיה 1</u>: נתון MDSTT באורך סופי $\{\underline{Y}_{sR}\}_{s=1}^{IL}$, נחפש את סדרת הדגימות $\{x(n)\}_{n=o}^{ILR-1}$ שהרישא הסופי של ה-DSTT שלה $\{\underline{X}_{sR}\}_{s=1}^{IL}$ קרוב ביותר ל-MDSTT הנתון בנורמת ℓ_2 משוקללת, קרי ממזער את:

$$D_1 \triangleq \frac{1}{IL} \sum_{s=1}^{IL} (\underline{X}_{sR} - \underline{Y}_{sR})^* \, G(\underline{X}_{sR} - \underline{Y}_{sR}) \tag{12}$$

כש-G זו מטריצה .P.D הרמיטית ממימד $M \times M$.

<u>בעיה 2</u>: נתון MDSTT באורך סופי $\{\underline{Y}_{sR}\}_{s=1}^{IL}$ שהתקבל מתוך DSTT חוקי על ידי הפעלת מודיפיקציה ליניארית זהה, והפיכה, על כל אחד מהוקטורים שבסדרה זו. תהא המודיפיקציה מיוצגת על ידי המטריצה הרגולרית P ממימד $M \times M$. זהו התהליך המבוצע למשל בערבול של דיבור. נחפש סדרת דגימות כך שהרישא הסופי של ה-DSTT שלה $\{\underline{X}_{sR}\}_{s=1}^{IL}$ לאחר שמפעילים עליו מודיפיקציה הפוכה (P^{-1}) קרוב ביותר למודיפיקציה ההפוכה של ה-MDSTT הנתון (קרי ל-DSTT המקורי). מכאן שנחפש את הממזער של הפונקציה:

$$D_2 \triangleq \frac{1}{IL} \sum_{s=1}^{IL} (P^{-1}(\underline{X}_{sR} - \underline{Y}_{sR}))^* \, G(P^{-1}(\underline{X}_{sR} - \underline{Y}_{sR})) \tag{13}$$

כששוב G זו מטריצה .P.D הרמיטית ממימד $M \times M$.

בעיה אקויולנטית: נציב ב-(13) וב-(12) את הקשר שבין ה-DSTV ל-DSTT (ובין ה-MDSTV ל-MDSTT), ונקבל תאור אחיד של שתי בעיות האופטימיזציה, כבעית המזעור של:

$$(14) \quad D = \frac{1}{IL} \sum_{s=1}^{IL} (\underline{x}_{sR}-\underline{y}_{sR})^* A(\underline{x}_{sR}-\underline{y}_{sR})$$

כאשר $\{\underline{y}_{sR}\}_{s=1}^{IL}$ הוא ה-MDSTV ($\underline{y}_{sR} \triangleq T^{-1}\underline{Y}_{sR}$), ו-$\{\underline{x}_{sR}\}_{s=1}^{IL}$ הוא ה-DSTV של הסידרה שמסונתזת. המטריצה A היא P.D. הרמיטית ממימד M × M, וניתנת על ידי $A = T^*GT$ עבור בעיה 1, ועל ידי; $A = (P^{-1}T)^* \; G(P^{-1}T)$ עבור בעיה 2.
בפרט, כאשר T ו-P יוניטריות ו- G = I (וזהו המקרה המקובל), אזי A = I והסינתזה האופטימלית פשוטה יותר במקרה זה (כפי שנראה בהמשך).

תכונות הסינתזה האופטימלית

משפט 6: D היא תבנית-ריבועית P.D. בנעלמים $\{x(n)\}_{n=o}^{ILR-1}$, ולכן $D \geq 0$ וקיימת סינתזה אופטימלית יחידה.

הוכחה: מאחר ו-A היא P.D. הרמיטית, הרי ברור ש-$D \geq 0$. כמוכן, ברור ש-D היא תבנית ריבועית בערכי ה-DSTV $\{\underline{x}_{sR}\}_{s=1}^{IL}$ ומאחר וה-DSTV תלוי לינארית בנעלמים הרי D היא תבנית ריבועית P.S.D. בנעלמים. נותר להראות ש-D היא תבנית P.D. בנעלמים. לשם כך מספיק להראות שעבור $\underline{y}_{sR} = \underline{0}$, $D = 0 \Rightarrow$ $x(n) = 0$ זהותית. ממשוואה (14) והעובדה ש-A רגולרית ברור ש-$D = 0 \Rightarrow$ $\underline{x}_{sR} = \underline{0}$ זהותית ב-s (כאשר $\underline{y}_{sR} = \underline{0}$). מהגדרת ה-DSTV ב-(1), ברור שהסדרה $x(n) = 0$ מביאה ל-DSTV שהוא 0 זהותית. בנוסף, ההנחה שקיימת ftUS משמעה שמתוך $\{\underline{x}_{sR}\}_{s=1}^{IL}$ ניתן לשחזר חד ערכית את $\{x(n)\}_{n=o}^{ILR-1}$ ולכן $D = 0 \Rightarrow x(n) = 0$ זהותית.

מ.ש.ל.

מסקנה: הסינתזה האופטימלית היא מערכת ftUS, ולכן כאשר $\{\underline{Y}_{sR}\}_{s=1}^{IL}$ מייצג DSTT חוקי, יווצר שחזור ללא שגיאה.

<u>הוכחה</u>: המסקנה נובעת ישירות מהמשפט. כאשר $\{\underline{y}_{sR}\}_{s=1}^{IL}$ מייצג DSTT חוקי, הרי קיימת סדרה שמשיגה $D = 0$, ומאחר וזהו החסם התחתון על D הרי הסינתזה האופטימלית תזדהה עם סדרה זו, ועל פי ההגדרה משמעות עובדה זו היא שהסינתזה האופטימלית מייצגת מערכת ftUS.

<u>פתרון בעית הסינתזה האופטימלית</u>

לאור משפט 6, ברור שהסינתזה האופטימלית מתקבלת על ידי הפתרון של מערכת המשוואות הלינאריות שנוצרות על ידי $\nabla D = \underline{0}$.

בהצבת (1) לתוך (14) ושימוש בהרמיטיות של $\underline{A}$ נקבל ש:

$$\frac{\partial D}{\partial x(n)} = \frac{1}{IL} \sum_{s=1}^{IL} 2h^*(sR-n) \sum_{k=o}^{M-1} a((n)_M,k) \; [x_{sR}(k)-y_{sR}(k)] \tag{15}$$

לכן משוואות הגרדיאנט יהיו (שוב בהצבת (1) לתוך (15)

$$\begin{aligned} &\sum_{k=o}^{ILR-1} \{ \sum_{s=1}^{IL} a((n)_M,(k)_M) h^*(sR-n) h(sR-k) \} \; x(k) = \\ &= \sum_{s=1}^{IL} \sum_{k=o}^{M-1} h^*(sR-n) \; a((n)_M,k) \; y_{sR}(k) \qquad ILR-1 \geq n \geq 0 \end{aligned} \tag{16}$$

לכן הסינתזה האופטימלית תהא מהצורה:

$$\underline{x} = S^{-1}\underline{z} \tag{17}$$

כשהוקטור $\underline{x}$ ממימד ILR מכיל את הדגמים $\{x(n)\}_{n=o}^{ILR-1}$. הוקטור $\underline{z}$ ממימד ILR מוגדר על ידי:

$$z(n) = \sum_{s=1}^{IL} h^*(sR-n) \sum_{k=o}^{M-1} a((n)_M,k) \quad y_{sR}(k) \qquad ILR-1 \geq n \geq 0 \tag{18}$$

ואילו המטריצה S ממימד ILR× ILR מוגדרת על ידי:

$$S_{nk} = \sum_{s=1}^{IL} h^*(sR-n) \; a((n)_M,(k)_M) \; h(sR-k) \tag{19}$$

IV. ממוש הסינתזה האופטימלית

מאחר וקצב הדגימה האופייני של וקטורי DSTT הוא של 100Hz, ומאחר ואות דיבור אופייני מתאים למשפט של מס' שניות ומעלה, הרי ש-ILR יהא בסדר גודל של מספר אלפים. לאור האמור לעיל, הפתרון המתואר ב-(17) נותר בחזקת פתרון תאורטי בלבד, וכדאי לדון במימוש יעיל שלו.

(א) חישוב הוקטור $\underline{z}$ = סינתזה בשיטת WOLA.

הוקטור $\underline{z}$ ממימד ILR מוגדר על ידי (18). ניתן לתאר את מימוש משוואה זו כדלקמן. חשב את סידרת הוקטורים $\{\underline{z}_{sR}\}_{s=1}^{IL}$ על ידי:

$$\underline{z}_{sR} = AT^{-1}\underline{Y}_{sR} \qquad (20a)$$

ועתה חשב את הסידרה $z(n)$ על ידי:

$$z(n) = \sum_{s=-\infty}^{\infty} h^*(sR-n)\, z_{sR}((n)_M) \qquad ILR-1 \geq n \geq 0 \qquad (20b)$$

כשבמשוואה (20) משלימים את הסדרה $\{\underline{z}_{sR}\}_{s=1}^{IL}$ על ידי וקטורי אפסים.
עבור $A = I$, משוואה (20) מתארת סינתזה בשיטת WOLA עם מסנן סינתזה לא סיבתי בעל תגובה להלם $f(n) = h^*(-n)$. כשמסנן האנליזה הוא מסנן FIR, אין בעיה בממוש יעיל של חלק זה (ראה, למשל, בפרק 1). עבור $A \neq I$ זו וריאציה קלה של שיטת WOLA.

(ב) היפוך ריקורסיבי של המטריצה S

על מנת לתאר היפוך יעיל של המטריצה S, נניח מכאן ואילך שהמסנן $h(\cdot)$ הוא מסנן FIR כך ש-$h(n) \neq 0$ עבור $1 \leq n \leq IRK_h$.
במקרה זה המטריצה S שהיא מטריצה הרמיטית, הינה Banded ברוחב של IRK_h, קרי $S_{nk} = 0$ עבור $|n-k| \geq IRK_h$.
יתר על כן כאשר $n = tR + \theta$, $0 \leq \theta \leq R-1$, הרי $S_{nk} \neq 0$ רק עבור:

$$(t-IK_h+1)R \leq k < (t+IK_h)R$$

ולכן המטריצה S תהא מטריצת בלוקים ממימד $R \times R$ שהיא Banded ברוחב IK_h.
או:

$$S = \begin{bmatrix} B_{00} & B_{01} & \cdots & B_{0(IK_h-1)} & 0 & \cdots & 0 & 0 \\ B_{10} & B_{11} & \ddots & B_{1(IK_h-1)} & B_{1IK_h} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ B_{(IK_h-1)} & \ddots & \ddots & \ddots & \ddots & & & \\ 0 & \ddots & \ddots & & & & & \\ \vdots & \ddots & & & & & & \\ 0 & & & & & & & \end{bmatrix}$$

בנוסף הערך בכל בלוק תלוי במרחקו מהאלכסון ובערך של $(n/R) \bmod I$, ולכן זו השתרגות של I מטריצות Toeplitz בבלוקים (מלבד IK_h הבלוקים האחרונים שתלויים מפורשות ב-L, עקב אפקטי קצה של הסכומים). לסיכום ניתן לתאר את המטריצה S כדלקמן:

$$S = \begin{bmatrix} B_0^{(o)} & \cdots & B_{(IK_h-1)}^{(o)} \\ \vdots \ \ddots & B_0^{(I-1)} & \ddots \\ B_{(IK_h-1)}^{(o)*} & \ddots & B_{(IK_h-1)}^{(I-1)} \\ 0 \ \ddots & & \text{"זנב"} \end{bmatrix} \Big\} IK_h$$

כאשר: $B_d^{(p)}$, $0 \le d \le IK_h-1$, $0 \le p \le I-1$ זו מטריצה $R \times R$ המוגדרת על ידי:

$$0 \le i,j \le R-1,\quad B_d^{(p)}(i,j) = a((pR+i)_M, (pR+dR+j)_M) \sum_{s=1}^{IK_h} h^*((s+d)R-i)\,h(sR-j) \qquad (21)$$

ובלוקים אלו בלתי תלויים בערך של L. החלק של S המסומן כ"זנב", שמימדו $IK_hR \times IK_hR$, מכיל בלוקים הדומים לבלוקים $B_d^{(p)}$, למעט הסכום על s שמוגבל כאן לסכום עבור $1 \le s \le IL - \lfloor \frac{k}{R} \rfloor$.

נתאר להלן שיטה להשגת פתרון של (17) בסיבוכיות שהיא לינארית באורך L, תוך ניצול המבנה המיוחד של S . לשם כך נניח שהבלוקים $B^{(p)}_{(IK_h-1)}$ הפיכים ל- $0 \le p \le (I-1)$ (או לחילופין, שהם אפס זהותית ואזי $B^{(p)}_{(IK_h-2)}$ הפיכים כולם, וכו'). נפרק את S לארבעה בלוקים כדלקמן:

$$S = \left[\begin{array}{c:c} S_1 & S_2 \\ \hline S_4 & S_3 \end{array}\right] \qquad (\text{width of } S_1: IK_hR;\ \text{height of } S_3: IK_hR)$$

כשהמטריצה S_2 ממימד $I(L-K_h)R$ היא משולשת תחתונה בבלוקים. מערכות משוואות מהצורה $S_2\underline{v} = \underline{u}$, ניתן לכן לפתור בסיבוכיות של $(IK_hR) \approx$ פעולות לאיבר. את הסיפא של הפתרון של (17) נתאר כקומבינציה לינארית של IK_hR פתרונות הומוגניים ופתרון פרטי אחד, קרי:

$$\underline{x}_2 = \underline{x}_{2_p} + \sum_{i=o}^{IK_hR-1} \alpha_i\, \underline{x}_{2h_i}$$

כאשר $S_2\underline{x}_{2_p} = \underline{z}_1$ ($\underline{z}_1$ ו-$\underline{x}_2$ הם הרישא/הסיפא של $I(L-K_h)R$ אברים ראשונים/אחרונים ב-$\underline{z}$ ו-$\underline{x}$ בהתאמה) ואילו $S_1\underline{e}_i + S_2\underline{x}_{2h_i} = \underline{0}$, כש-$\underline{e}_i$ הוא וקטור היחידה ה-i-י ב-$R^{(IK_hR)}$.

לכן קבלת IK_hR+1 הוקטורים הנ"ל כרוכה ב-$O((IK_hR)^2)$ פעולות לאיבר, וכל היפוכי המטריצות הם ב"ת בערך הספציפי של L (רק אורכי הוקטורים תלויים בו). עתה IK_hR המשוואות האחרונות שעל הוקטור $\underline{x}$ לקיים, יפתרו על ידינו על מנת לקבוע את המקדמים $\{\alpha_i\}_{i=o}^{IK_hR-1}$ וכך להשלים את הפתרון. דהיינו נפתור את מערכת המשוואות:

$$\left(\begin{bmatrix} \underline{x}^T_{2h_o} \\ \vdots \\ \underline{x}^T_{2h_{(IK_hR-1)}} \end{bmatrix} S_3 + S_4\right)\underline{\alpha} = \underline{z}_2 - S_3^T \cdot \underline{x}_{2_p}$$

וזה יגזול עוד לערך $O((IK_hR)^2)$ פעולות לאיבר. לכן הסיבוכיות הכוללת של שיטה זו לפתרון (17) היא $O((IK_hR)^2(ILR))$, דהיינו סיבוכיות לינארית באורך.

לשיטת סינתזה זו נותרו שלוש מגבלות והן כדלקמן:

(א) פעולה Batch, מחייבת לחכות עד סוף ה-MDSTT, עד להפקת פלט הסינתזה.

(ב) חייבים לשמור בזכרון $(IK_h R)$ סדרות באורך ILR דגמים => נדרש זכרון גדול (בסדרי גודל של מליוני ערכים).

(ג) מאחר ולמעשה אנו הופכים את המטריצה S, לא מספיק שאנו בטוחים ש-S רגולרית, אלא נדרש לבדוק את היציבות הנומרית של תהליך ההפוך (קרי את ה-condition-number של S כאשר $L \to \infty$).

V . <u>סינתזה אופטימלית עבור MDSTT אינסופי (פתרון במצב יציב)</u>

כפי שראינו בסעיף הקודם, חישוב הוקטור $\underline{z}$ (משואה (20)) נעשה על ידי סינתזת WOLA המקובלת ואינו תלוי מפורשות באורך ה-MDSTT הנתון. מאידך היפוך המטריצה S תלוי ב-L ולכן מחייב פעולה ב-Batch. נחפש מימוש רקורסיבי של ההפכי של S עבור $L \to \infty$. למטרות הפשטות נניח ראשית כל ש- $A = I$ (וזהו בדרך כלל המקרה המעניין). במקרה זו מערכת המשוואות (17) היא פריקה ל-M מערכות משוואות ממימד JL כל אחת, המתאימות ל-M ההעתקות $\tilde{H}_p$ $1 \le p \le M$ שתוארו בסעיף II. זה נובע מכך ש-0 = S_{nk} כאשר $(n-k)\bmod M \neq 0$. לפיכך נגדיר את הסדרות $\{z_p(r)\}_{r=1}^{JL}$ ו- $\{x_p(r)\}_{r=1}^{JL}$ על ידי

$$\begin{cases} z_p(r) = z(rM-p) & JL \ge r \ge 1 \\ x_p(r) = x(rM-p) & M \ge p \ge 1 \end{cases}$$

ואזי משואה (17) תפושט ל-

$$S^{(p)}\underline{x}_p = \underline{z}_p \qquad M \ge p \ge 1 \tag{21}$$

כשאברי המטריצה $S^{(p)}$ ממימד JLxJL ניתנים על ידי:

$$S_{ij}^{(p)} = \sum_{s=1}^{IL} h_p^*(sJ-iI)h_p(sJ-jI) \qquad 1 \le i,j \le JL \tag{22}$$

ו-$h_p(x)$ הוא ה-GP ה-p-י של מסנן האנליזה שהוגדר ב-(3).

כאשר המסנן $h(\cdot)$ הוא מאורך של $IR=MJ$ דגמים או פחות (קרי $K_h = 1$) הרי קל לראות מ-(22) שהמטריצה $S^{(p)}$, היא מטריצה אלכסונית בבלוקים (שממדם $J \times J$). במקרה זה סינתזת ה-WOLA משתנה רק בהכפלת מסנני הסינתזה במטריצת משקולות מתאימה ולכן גם הסינתזה האופטימלית עבור L קבוע, תהא מאוד פשוטה וישימה באופן רקורסיבי, ופרט לבלוק האחרון של $JK_h \times JK_h$ דגמים תהא בלתי-תלויה מפורשות בערך של L. עבור $J = 1$ זהו בדיוק הפתרון של [16,17] שתואר בפרק 2, אך ניתן להכלילו (כך מסתבר כאן) למסנני אנליזה מאורך גדול מ-M דגמים על ידי בחירת ערך גבוה של J (למשל על ידי שימוש ב-R שהוא ראשוני).

את הסינתזה האופטימלית עבור $L \to \infty$ ניתן לפתח בשתי דרכים שונות. היא מתבססת על התכונות הבאות של $S^{(p)}$.

<u>טענה 8</u>: המטריצה $S^{(p)}$ היא (פרט לזנב ממימד $J(K_h-1) \times J(K_h-1)$) מטריצת טואפליץ בבלוקים (שמימדם $J \times J$), הרמיטית והיא Banded לרוחב של K_h בלוקים, וזאת עבור מסנן אנליזה שהוא FIR מאורך $R\cdot I\cdot K_h$.

<u>הוכחה</u>: בדומה לתכונות של S שתוארו בסעיף III, קל לראות ש-$S^{(p)}$ מקיימת את הקשר $S^{(p)}(i+J,j+J) = S^{(p)}(i,j)$ עבור $1 \le i,j \le J(L-K_h)$ וכן ש- $S^{(p)}(i,j) = 0$ עבור $|i-j| \ge JK_h$. זה ש-$S^{(p)}$ היא הרמיטית ברור מהגדרתה.

לפיכך המטריצה $S^{(p)}$ תהא מהצורה:

$$S^{(p)} = \left[\begin{array}{cccccc} B_0 & \cdots & B_{-(K_h-1)} & 0 \cdots & 0 \cdots & 0 \\ B_{+1} & \ddots & & \ddots & & \vdots \\ \vdots & & \ddots & & \ddots & \vdots \\ B_{+K_h-1} & & & \ddots & & \vdots \\ 0 & \ddots & & & \ddots & 0 \\ \vdots & & \ddots & B_0 & & B_{-(K_h-1)} \\ 0 & \cdots & 0 & B_{+K_h-1} & & \underbrace{C}_{J(K_h-1)} \end{array}\right] \Big\} J(K_h-1)$$

הן המטריצות $B_{(K_h-1)} \cdots B_{-(K_h-1)}$ והן C ב"ת מפורשות בערך של L.

כמוכן:

$$(23)\quad B_k(i,j) = \sum_{s=-\infty}^{\infty} h_p^*(sJ-iI)\, h_p(sJ+kIJ-jI) \qquad 1 \le i,j \le J, \quad 0 \le k \le (K_h-1)$$

ו-

$$(24)\quad B_{-k} = B_k^*$$

הדרך הראשונה לפתח את הסינתזה האופטימלית הינה למעשה הרחבה של הגישה האלגברית שתוארה ב-[12]. בדומה לתוצאות של [73], ניתן להראות שהמטריצה $S^{(p)}$ שהיא כמעט טואפליץ בבלוקים, היא אקויולנטית אסימפטוטית (עבור $L \to \infty$) למטריצה שהיא Block-Circulant (וזאת תחת תנאים מסוימים על המטריצות B_k שממילא יתוארו אח"כ). ההפכי של מטריצות כאלה גם הוא Block-Circulant וניתן לקבלו ביעילות על ידי שימוש ב-DFT איבר-איבר ואחר כך היפוך מטריצות "Pointwise" בתדר (זו הרחבה מיידית של ההפוך של מטריצות Circulant המתואר ב-[67]).

עתה ניתן להראות שבטעות קטנה כרצוננו ניתן לקטום את המטריצה ההפכית ולמעשה לבצע את הסינתזה עם מטריצת Block-Toeplitz שהיא גם Banded (שזו ההרחבה המיידית של מסנן FIR, שמתקבל ב-[12] עבור $J = 1$). עם זאת נראה לנו שדרך זו היא מסורבלת להבנה (ההוכחות של כל האמור לעיל, הן די ארוכות ומייגעות), וניתן להגיע לאותן תוצאות תוך שימוש בהתמרת פורייה, ואכן בהתאם נציג את הפיתוח. נתבונן במערכת המשוואות (21) עבור $L \to \infty$, נחלק את הוקטורים $\underline{z}_p$ ו-$\underline{x}_p$ לסדרה אינסופית של וקטורים קצרים באורך J כל אחד. דהיינו:

$$(25)\quad \underline{u}_n \triangleq \begin{vmatrix} x_p(nJ+1) \\ \vdots \\ x_p(nJ+J) \end{vmatrix} \quad ; \quad \underline{v}_n \triangleq \begin{vmatrix} z_p(nJ+1) \\ \vdots \\ z_p(nJ+J) \end{vmatrix} \qquad n = 0,1,2,\ldots$$

ועתה עבור $L \to \infty$ (21) היא למעשה:

$$(26)\quad \sum_{k=-(K_h-1)}^{(K_h-1)} B_k\, \underline{u}_{n-k} = \underline{v}_n \qquad n \ge 0$$

כאשר תנאי ההתחלה הם $\underline{u}_k = 0$ עבור $k < 0$. ברור שלכל ערך של $\underline{u}_{(K_h-2)}, \ldots, \underline{u}_1, \underline{u}_0$ ישנו פתרון אחר ל-(26), לכן נתעלם כליל מתנאי ההתחלה ונחפש את הפתרון היציב של המשוואות הנ"ל.

נניח עתה שסדרת ה-MDSTT שייכת ל-ℓ_1 קרי $\sum_{s=1}^{\infty} ||\underline{Y}_{sR}||_1 < \infty$.

מאחר ומסנן האנליזה $h(\cdot)$ הוא FIR, הרי על פי (20) ברור שגם הסדרה $\{\underline{v}_n\}_{n=o}^{\infty}$ שייכת ל-ℓ_1, ולכן התמרת פורייה הדיסקרטית שלה מוגדרת לכל $f \in [-0.5,0.5]$ והיא וקטור J הפונקציות הרציפות הבא:

$$(27) \qquad \underline{V}(f) \triangleq \sum_{n=o}^{\infty} \underline{v}_n \, e^{-j2\pi fn}$$

נגדיר את התמרת Z של סידרת המטריצות $\{B_k\}_{k=-(K_h-1)}^{(K_h-1)}$ כמטריצה ממימד $J \times J$ שכל אבריה הם פולינומים ב-Z ממעלה של לכל היותר $(2K_h-1)$ המתקבלת על ידי:

$$(28) \qquad B(Z) \triangleq \sum_{k=-(K_h-1)}^{(K_h-1)} B_k Z^{-k}$$

עתה $\det B(Z)$ הוא פולינום ממעלה של לכל היותר $J(2K_h-1)$ ב- Z ולכן יש לו מספר סופי של אפסים. נניח שאף אחד מהם אינו על מעגל היחידה (קרי $B(e^{j2\pi f})$ היא רגולרית לכל $f \in [-0.5,0.5]$). ואזי קיימת מטריצה של פונקציות רציפות:

$$(29) \qquad Y(f) \triangleq B(Z)^{-1}\Big]_{Z=e^{j2\pi f}}$$

(הרציפות של אברי $Y(f)$ נובעת מכך שאלו פולינומים טריגונומטריים רציונלים שמכניהם אינו מתאפס על מעגל היחידה). נגדיר את סדרת המטריצות הבאה:

$$(30) \qquad Y_k \triangleq \int_{-0.5}^{0.5} Y(f) e^{j2\pi fk} df \qquad -\infty < k < \infty$$

שמוגדרת היטב מכיון שאברי $Y(f)$ רציפים, בנוסף מאחר ו- $B_k^*=+B_{-k}$ הרי קל לראות שגם $Y_k^* = Y_{-k}$.

<u>משפט 7:</u>

(א)
$$\tilde{\underline{u}}_n = \int_{-0.5}^{0.5} Y(f) \, \underline{V}(f) e^{j2\pi fn} \, df$$

הוא פתרון של (26), אולם לאו דוקא עם תנאי התחלה אפס.

(ב)
$$\tilde{\underline{u}}_n = \sum_{k=-\infty}^{\infty} Y_k \, \underline{v}_{n-k}$$
כאשר $\underline{v}_k = 0$ עבור $k < 0$.

(ג) $||\tilde{\underline{u}}_n - \sum_{k=-(T-1)}^{(T-1)} Y_k \underline{v}_{n-k}||_2 \underset{T\to\infty}{\to} 0$ במידה שווה ב-n.

<u>הוכחה:</u>

(א) $\tilde{\underline{u}}_n$ מוגדר היטב, שכן כל אבר בוקטור $Y(f)\ \underline{V}(f)$ הוא מכפלת שתי פונקציות רציפות ולכן בודאי זו פונקציה ב-$L_1[-0.5,0.5]$, כך שהאינטגרל מוגדר היטב.
נתבונן ב-

$$\sum_{k=-(K_h-1)}^{(K_h-1)} B_k \tilde{\underline{u}}_{n-k} = \int_{-0.5}^{0.5} \sum_{k=-(K_h-1)}^{(K_h-1)} B_k e^{-j2\pi fk}\, Y(f)\ \underline{V}(f) e^{j2\pi fn} df =$$

$$= \int_{-0.5}^{0.5} \underline{V}(f)\ e^{j2\pi fn} df = \underline{v}_n$$

השיוויון הראשון נובע מהגדרת $\tilde{\underline{u}}_n$ והחלפת סדר הסכום סופי והאינטגרציה, השני מהגדרת $Y(f)$ (משוואות (28), (29)) והשלישי מהגדרת $\underline{V}(f)$ (משוואה (27)).
לכן (26) מתקיימת, אולם יתכן ש-$\tilde{\underline{u}}_k$ $-(K_h-1) < k < 0$ איננו אפס.

(ב) מאחר ו-$\underline{v}_n$ שייכת ל-ℓ_1 ו-$Y(f)$ שייכת ל-$L_1[-0.5,0.5]$, הרי כל איבר בוקטור $Y_k \underline{v}_{n-k}$ שייך ל-ℓ_1 ולכן הסכום באגף ימין של (ב) מוגדר היטב. עתה:

$$\sum_{k=-\infty}^{\infty} Y_k \underline{v}_{n-k} \triangleq \sum_{k=-\infty}^{\infty} \int_{-0.5}^{0.5} Y(f) e^{j2\pi fk} df\ \underline{v}_{n-k} = \sum_{k=-\infty}^{\infty} \int_{-0.5}^{0.5} Y(f) e^{j2\pi fn} \cdot$$

$$\cdot\ [\underline{v}_{n-k} e^{-j2\pi f(n-k)}] df = \int_{-0.5}^{0.5} Y(f) e^{j2\pi fn}\ \underline{V}(f) df \triangleq \tilde{\underline{u}}_n$$

כשהשיוויון האחרון מתקבל מהחלפת סדר האינטגרציה וסכימה שמותרת כי $\underline{v}_n$ שייכת ל-ℓ_1.

(ג) $$||\tilde{\underline{u}}_n - \sum_{k=-(T-1)}^{(T-1)} Y_k\ \underline{v}_{n-k}||_2 = ||\sum_{|k|\geq T} Y_k \underline{v}_{n-k}||_2 \leq \sum_{|k|\geq T} ||Y_k \underline{v}_{n-k}||_2 \leq$$

$$\leq \max_{|k|\geq T} ||Y_k||_2 \sum_{n=o}^{\infty} ||\underline{v}_n||_2$$

כשהאי-שיוויון הראשון הוא אי-שיוויון המשולש לנורמה $||\cdot||_2$, והאי-שיוויון השני מתקבל מכך ש-$||A\underline{v}||_2 \le ||A||_2 \cdot ||\underline{v}||_2$, כש- $||A||_2 \overset{\Delta}{=} \sqrt{\lambda_{max}(A'A)}$.

עתה מאחר ו-$\underline{v}_n$ שייכת ל-ℓ_1 הרי $\sum_{n=0}^{\infty} ||\underline{v}_n||_1 < \infty$ ולכן בברור גם $\sum_{n=0}^{\infty} ||\underline{v}_n||_2 < \infty$ (כי $\underline{v}_n$ וקטור סופי, ולכן עבורו הנורמות אקויולנטיות). לכן להוכחת (ג) מספיק שנראה כי $\lim_{T\to\infty} \max_{|k|\ge T} ||Y_k||_2 = 0$. מהלמה של רימן-לבג [71] נובע שכל איבר של Y_k שותף לאפס עבור $|k|\to\infty$ (כי $Y(f)$ שייכת ל-$L_1[-0.5,0.5]$), ומאחר שהמטריצה Y_k ממימד סופי הרי גם $||Y_k||_2 \underset{k\to\infty}{\to} 0$, ובזאת הושלמה ההוכחה.

<u>מסקנה:</u> לכל $\varepsilon > 0$ קטן כרצוננו, קיים T_ε גם ש-$\sum_{k=-(T_\varepsilon-1)}^{(T_\varepsilon-1)} Y_k \underline{v}_{n-k}$ מהווה קירוב ε של הסידרה $\underline{\tilde{u}}_n$. הסידרה $\underline{\tilde{u}}_n$ היא פתרון של (26), אך תנאי ההתחלה שלה אינם אפס.

<u>טענה 9:</u> $\underline{\tilde{u}}_n$ הוא הפיתרון היציב היחיד של (26).

<u>הוכחה:</u> ראשית נגדיר במדויק מהו פתרון יציב: יהיו $\underline{v}_n^{(1)}$, $\underline{v}_n^{(2)}$ שתי סדרות השונות זו מזו באיברים הראשונים שלהן בלבד, אזי עבור פתרון יציב של (26) נצפה שלשתי סדרות הפתרונות המתאימים $\underline{u}_n^{(1)}$, $\underline{u}_n^{(2)}$ מתקיים $\lim_{n\to\infty} ||\underline{u}_n^{(1)} - \underline{u}_n^{(2)}||_2 = 0$.

זה ש-$\underline{\tilde{u}}_n$ הוא פתרון יציב נובע מיידית מ-(ב) ומהעובדה ש-$\lim_{k\to\infty} ||Y_k||_2 = 0$.

על מנת להוכח שזהו הפתרון היציב היחיד של (26), נבחין בכך ש-$B^*(Z) = B(\bar{Z}^{-1})$ ולכן האפסים של $\det B(Z)$ מסודרים בזוגות המקיימים $Z_1\bar{Z}_2 = 1$, ולכן מחציתם בתוך מעגל היחידה ומחציתם מחוצה לו. אפסים אלו הופכים לקטבים של המערכת המאופיינת על ידי המטריצה $Y(Z) = B^{-1}(Z)$. הפתרונות השונים של (26) נבדלים ביניהם רק בחלוקת הקטבים של $Y(Z)$ בין החלק הסיבתי $(k \ge 0)$ והחלק הלא-סיבתי $(k < 0)$ של הסדרה Y_k. ידוע שקיימת רק חלוקת קטבים אחת המבטיחה יציבות ביחס לכל כניסה $\underline{v}_n$ וזו החלוקה שמתאימה את כל הקטבים שבתוך מעגל היחידה לחלק הסיבתי ואת השאר לחלק הלא סיבתי. הפתרון המתקבל הוא בדיוק $\underline{\tilde{u}}_n$ שהוצג לעיל כפי שקל לודא.

<u>מסקנה</u>: ניתן לפרש את הפתרון $\tilde{\underline{u}}_n$ שהוצג כאן, כפתרון היחיד של בעית הסינתזה שהיא ill-posed, (הנובעת מאי-יכולת לקבוע את תנאי ההתחלה באופן חד-ערכי), שאינו רגיש ביחס לתנאים אלו. לפיכך הפתרון הנ"ל מכונה על ידינו פתרון במצב-יציב.

<u>טענה 10</u>: לאחר שקבענו $\varepsilon > 0$ קטן כרצוננו וקבענו T_ε כך שנקבל קירוב $\varepsilon/2$ של $\tilde{\underline{u}}_n$, ניתן לקבל קירוב ε של $\tilde{\underline{u}}_n$ על ידי החלפת הסדרה $\{Y_k\}_{k=-(T_\varepsilon-1)}^{(T_\varepsilon-1)}$ בקירוב הבא שלה:

(א) חשב את $Y(f)$ בתדרים $f_m = \frac{m}{Q}$ עבור $0 \le m \le (Q-1)$, על ידי DFT ממימד Q של הסדרה הסופית B_k $(Q \ge (2K_h-1))$, והיפוך מטריצות pointwise.

(ב) חשב את $\{\tilde{Y}_k\}_{|k|\le(T_\varepsilon-1)}$ על ידי קירוב האינטגרל ב-(30) על ידי סכום סופי, דהיינו:

$$\tilde{Y}_k = \frac{1}{Q}\sum_{m=o}^{(Q-1)} Y\left(\frac{m}{Q}\right)e^{j\frac{2\pi}{Q}mk} \tag{31}$$

דהיינו על ידי IDFT ממימד Q (כאשר $(2T_\varepsilon-1) \ge Q$).

וזאת כמובן על ידי בחירת $Q_\varepsilon \ge Q$ גדול מספיק.

<u>הוכחה</u>: מאחר וכל איבר של $Y(f)$ הוא פולינום טריגונומטרי רציונלי עם מכנה שונה מאפס, הרי $Y(f)$ אינטגרבילית רימאן, וכך גם $Y(f)e^{j2\pi fk}$. לכן לכל k ולכל איבר של Y_k קיים מספר Q_ε כך שההבדל בין האינטגרל והטור-הסופי שב-(32) קטן מ-$\varepsilon/2$, עבור איבר זה. לכן על ידי בחירת מספר Q_ε גדול יותר ניתן לקבל ש-

$$\sum_{n=o}^{\infty} ||\underline{v}_n|| \max_{|k|\le(T_\varepsilon-1)} (||\tilde{Y}_k - Y_k||_2) < \varepsilon_k$$

(וזאת כי T_ε סופי, ו-$\underline{v}_n$ ב-ℓ_1). בדומה להוכחת סעיף (ג) של משפט 7, ניתן להראות ש-

$$||\tilde{\underline{u}}_n - \sum_{k=-(T_\varepsilon-1)}^{(T_\varepsilon-1)} \tilde{Y}_k \underline{v}_{n-k}||_2 \le \varepsilon/2 + \max_{|k|\le(T_\varepsilon-1)} (||\tilde{Y}_k - Y_k||_2)\cdot \sum_{n=o}^{\infty} ||\underline{v}_n||_2 \le \varepsilon$$

כשהאי-שיוויון השמאלי, מתקבל מאי-שיוויון המשולש ומחסמים על $||A\underline{v}||_2$. מאחר ו- T_ε סופי ניתן גם להשתמש במקום בנורמה של $\underline{v}_n$ ב- ℓ_1 (שעלולה להשתנות מ-MDSTT אחד למשנהו) בנורמה של $\underline{v}_n$ ב-ℓ_∞, שהיא בדרך כלל חסומה גלובלית.

<u>סיכום:</u> ניתן על ידי DFT גדול מספיק, בעזרת האלגוריתם שתואר בטענה 10, לייצר סידרה סופית של $(2T_\varepsilon-1)$ מטריצות $\{\tilde{Y}_k\}$ ממימד $J \times J$ כל אחת, כך שהסינתזה הרקורסיבית על ידן של $\underline{x}_p$, מהווה קירוב ε (במידה שווה), של הפתרון היציב היחיד של בעית הסינתזה האופטימלית עבור $L \to \infty$.

VI. מערכות יחידה המכילות מודיפיקציה לינארית

נדון כעת במערכת שבה ה- DSTT $\{\underline{X}_{sR}\}_{s=1}^{\infty}$ מוסב ל-MDSTT על ידי מודיפיקציה לינארית קבועה שנעשית על כל וקטור בנפרד קרי ה-MDSTT $\hat{\underline{X}}_{sR}$ מתקבל על ידי:

$$\hat{\underline{X}}_{sR} = P\underline{X}_{sR} \qquad s \geq 1 \tag{33}$$

כש-P זו מטריצה רגולרית ממימד M×M , הידועה בעת השחזור.

זהו המודל המתאים למשל לערבול של דיבור (ראה [12] ליתר פירוט), או לסינון לינארי במישור הטרנספורם. על ידי סינתזה מתוך ה-MDSTT (למשל בשיטת WOLA) יוצרים סידרת דגימות שהיא שמשודרת, ובמקלט משחזרים מתוכה את הסידרה המקורית על ידי מערכת אנליזה וסינתזה שניה המכילה בתוכה את המודיפיקציה P^{-1} (ראה למשל בציור 6.1 מערכת כזו).

נגדיר להלן מודיפיקציה "חוקית" (למטרות ערבול), ונראה מיד שעבורה ניתן לשחזר את האות המקורי ללא שגיאה במערכת הצפנה / פענוח המבוקרת על שתי מערכות A/S.

<u>הגדרה:</u> מודיפיקציה P תיקרא "חוקית", (LM), עבור אנליזה על ידי המסנן $\{h(n)\}_{n=0}^{\infty}$ עם הפרמטרים R,M שעבורה קיימת ftUS, אם קיימת מערכת אנליזה שניה כלשהי בעלת אותם פרמטרים R,M ומסנן אנליזה $\{\hat{h}(n)\}_{n=0}^{\infty}$ אחר כלשהו, כך שלכל סידרה חצי-אינסופית $x(\cdot)$, קיימת סידרה אחרת חצי-אינסופית $y(\cdot)$ כך שה-DSTT של $y(\cdot)$ במערכת השניה מזדהה עם ה-MDSTT של $x(\cdot)$ במערכת המקורית.

<u>מסקנה:</u> לכל מודיפיקציה "חוקית", ניתן לשחזר את הסידרה המקורית במקלט ללא שגיאה, וזאת על ידי שימוש במקלט בסינתזה המכילה את P^{-1} ואחריה את ה-ftUS של מערכת האנליזה המקורית. לשם כך פלט המשדר יהא הסידרה $y(\cdot)$ המתאימה לסידרה המקורית על פי ההגדרה דלעיל, והאנליזה במקלט תיעשה על ידי מערכת האנליזה השניה המיוצגת על ידי $\{\hat{h}(n)\}_{n=o}^{\infty}$.

<u>הערה:</u> כפי שנראה בהמשך ה"חוקיות" של מודיפיקציה P תהא בדרך כלל תלויה במסנן האנליזה $\{h(n)\}_{n=o}^{\infty}$. לפיכך נגדיר גם מודיפיקציה "חוקית" אוניברסלית (ULM), כדלקמן:

<u>הגדרה:</u> P היא ULM אם"ם היא LM ביחס לכל מסנן אנליזה $\{h\ (n)\}_{n=o}^{\infty}$.

נסמן על ידי $\hat{P}$ את המטריצה שאבריה נתונים על ידי:

$$(34)\qquad \hat{P}(\ell,m) = (T^{-1}PT)\ (M-\ell\ ,M-m) \qquad 1 \le \ell,m \le M$$

נגדיר בהמשך את מחלקת המטריצות $\hat{P}$ המייצגות LM ו-ULM (זו הדרך היותר טבעית לאפיין מחלקות אלו), ומאחר והמעבר מ- P ל- $\hat{P}$ על ידי (34) הוא הפיך, ניתן מאפיון זה לקבל אפיון מקביל של מחלקת המטריצות P המייצגת LM ו-ULM.

לאור האמור בסעיף II של נספח זה, הרי מאחר וההעתקה H מהסדרה הזמנית ל-DSTV פריקה לסכום ישר של M העתקות (שסומנו שם $(H_1,\ldots,H_M$) ניתן להגדיר LM כמטריצה $\hat{P}$ שעבורה מתקיים:

$$(\exists\, \hat{h})\ (\forall\ x)\ (\forall\ M \ge \ell \ge 1)\ (\exists\, y_\ell(x,h,\hat{h},\hat{P}))\ (\forall\ s \ge 1)$$

$$(35)\qquad \sum_{t=1}^{\infty} \hat{h}_\ell\ (sJ-tI)y_\ell(t) = \sum_{m=1}^{M} \hat{P}(\ell,m) \sum_{r=1}^{\infty} h_m(sJ-rI)\ x_m(r)$$

כש-$x_m(\cdot)$, $h_m(\cdot)$, $y_\ell(\cdot)$, $\hat{h}_\ell(\cdot)$ הם בעקבות ההגדרות של תחילת סעיף II.

<u>טענה 11:</u> תנאי מספיק והכרחי לכך ש-$\hat{P}$ היא LM הוא ש-(35) מתקיימת עבור MJ הסדרות הבאות: $x_m(r) = \delta(m-m_o)\delta(r-r_o)$, $1 \le r_o \le J$, $1 \le m_o \le M$.

<u>הוכחה</u>: ברור שזהו תנאי הכרחי. נוכיח שהוא גם תנאי מספיק. נניח שהתנאי הנ"ל מתקיים ונסמן על ידי $y(\ell,m_o,t,r_o)$ את הערך של $y_\ell(t)$ המתקבל ב-(35) עבור $.x_m(r) = \delta(m-m_o)\delta(r-r_o)$

עתה עבור $x_m(r) = \delta(m-m_o)\ \delta(r-r_o-\theta J)$, $\theta \geq 0$, (35) יתקיים **עבור**:

$$y_\ell(t) = \begin{cases} y(\ell,m_o,t-\theta J,r_o) & t > \theta J \\ 0 & t \leq \theta J \end{cases}$$

כפי שקל לודא.

לכן לסדרה כלשהי $x(\cdot)$, הסדרה המתאימה $y(\cdot)$ תתקבל על ידי:

$$(36)\qquad y_\ell(t) = \sum_{m=1}^{M}\sum_{r=1}^{J}\sum_{\theta=o}^{\lfloor\frac{t-1}{J}\rfloor} y(\ell,m,t-\theta J,r)\ x_m(r+\theta J)$$

מ.ש.ל.

<u>מסקנה</u>: תנאי הכרחי ומספיק לכך ש- $\hat{P}$ היא LM,הוא שקיים $\{\hat{h}(n)\}_{n=o}^{\infty}$ וכן $\{z(\ell,m,t,r)\}_{t=1}^{\infty}$ המוגדר עבור $1 \leq \ell \leq M$, $1 \leq r \leq J$ ו-$m \in I_\ell$ (כש- $I_\ell \in \{1,\dots,M\}$ הוא אוסף ה-m-ים כך ש-$\hat{P}(\ell,m) \neq 0$), כך ש:

$$(37)\qquad \sum_{t=1}^{\infty} \hat{h}_\ell(sJ-tI)z(\ell,m,t,r) = h_m(sJ-rI) \qquad\qquad \infty > s \geq 1$$

<u>הוכחה</u>: המסקנה נובעת מטענה 11, כי אם $\hat{P}(\ell,m_o) = 0$ הרי ברור ש-$y_\ell(t) = 0$ מאפשר פתרון של (35) עבור $x_m(r) = \delta(m-m_o)\ \delta(r-r_o)$ לכל r_o.

עתה עבור $\hat{P}(\ell,m) \neq 0$ הרי על פי סימוני טענה 11:

$$z(\ell,m,t,r) = y(\ell,m,t,r)\ /\ \hat{P}(\ell,m)$$

<u>הערה</u>: מהמסקנה דלעיל אנו רואים שהעובדה ש-$\hat{P}$ היא LM נקבעת רק על פי <u>מיקום האיברים השונים מאפס</u> של $\hat{P}$ ואינה תלויה בערכם הספציפי.

<u>טענה 12</u>:

(א) לכל $1 \leq \ell \leq M$, $I_\ell \neq \phi$.

(ב) אם $\hat{P}$ היא LM הרי ל- $\{\hat{h}(n)\}_{n=0}^{\infty}$ קיימת ftUS, קרי $\hat{h}(n) \neq 0$ עבור $1 \leq n \leq R$.

<u>הוכחה</u>:

(א) זה נובע מיידית מכך ש-$\hat{P}$ רגולרית ולכן בכל שורה שלה ישנו איבר אחד לפחות השונה מאפס.

(ב) נניח ש-$\hat{P}$ היא LM, אזי $\hat{h}(\cdot)$ חייב לאפשר פתרון של (37) לפחות עבור $1 \leq s \leq I$.

מאחר ונתון של-$h(\cdot)$ קיימת ftUS, הרי בפרט ניתן לשחזר את $\{x_m(r)\}_{r=1}^{J}$ מתוך I וקטורי ה-DSTV הראשונים ולכן לכל $1 \leq m \leq M$, J הוקטורים $\{h_m(sJ-rI)\}_{r=1}^{J}$ הם וקטורים <u>בת"ל</u> ב-$\mathbb{C}^I$.

את אגף שמאל של (37) עבור $1 \leq s \leq I$ נתאר על ידי כפל מטריצה $\hat{H}_\ell$ ממימד $J \times J$ (שאיבריה $\hat{h}_\ell(sJ-tI) \triangleq (\hat{H}_\ell(s,t)$, בוקטור $\underline{z}^{(\ell,m,r)} \in \mathbb{C}^J$.

לאור חלק (א) של הטענה הרי (37) מתקיים לכל $1 \leq r \leq J$ לפחות עבור m_o מסוים. לכן, על פי האמור לעיל, הטווח של $\hat{H}_\ell$ חייב להיות תת-מרחב לינארי ממימד $J \leq \dim(\text{Image}(\hat{H}_\ell))$. אך מכיון שב-$\hat{H}_\ell$ יש J עמודות, נובע מכך שלכל $1 \leq \ell \leq M$, $J = \text{rank}(\hat{H}_\ell)$, ועל פי טענה 5 בסעיף II של נספח זה, זהו בדיוק התנאי לקיום ftUS עבור $\hat{h}(\cdot)$.

מ.ש.ל.

<u>מסקנה (מתוך הוכחת טענה 12)</u>:

אם $\hat{P}$ היא LM אזי לכל $m \in I_\ell$, $\text{Image}(\hat{H}_\ell) = \text{Image}(H_m)$.

<u>טענה 13</u>: תנאי הכרחי לכך ש-$\hat{P}$ תייצג LM הוא ש- $\hat{P}$ היא מטריצה Block-Diagonal עם בלוקים ממימד $g \times g$.

<u>הוכחה</u>: לאור המסקנה דלעיל, מספיק להוכיח שלכל $m = u + gv$, $\ell = \omega + gq$, כש- $1 \leq \omega,u \leq g$ ו-$0 \leq v,q \leq I-1-$, המקיימים $v \neq q$, מתקיים ש:

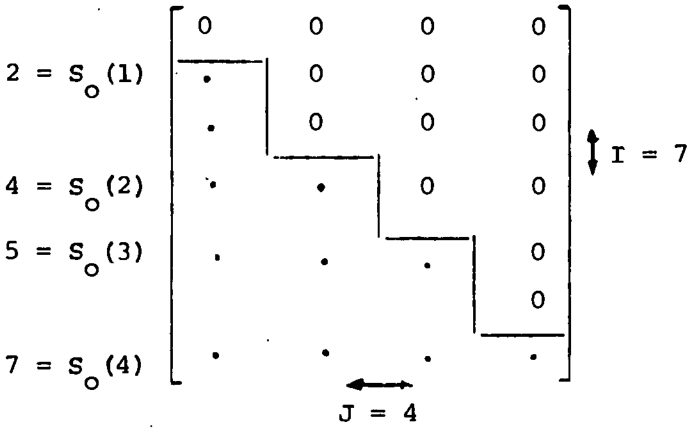
$\text{Image}(\hat{H}_\ell) \neq \text{Image}(H_m)$.

עתה:

$$1 \le s \le I \qquad H_m(s,t) = h(u+g(v+sJ-tI))$$
$$1 \le t \le J \qquad \hat{H}_\ell(s,t) = h(\omega+g(q+sJ-tI))$$

ומכיון ש-$h(\cdot)$ ו-$\hat{h}(\cdot)$ הם בעלי ftUS (על פי הנתון וטענה 12), הרי לאור משפט 3 (בסעיף II), ישנו בכל עמודה של H_m ושל $\hat{H}_\ell$ איבר השונה מאפס, כשהאיבר הראשון השונה מאפס בכל עמודה מתאים לארגומנט של $h(\cdot)$ ו-$\hat{h}(\cdot)$ שהוא בתחום $1,2,\dots,gJ$, וזה מתאים לשורה $s_o(t) = \lceil (tI-v)/J \rceil$ עבור H_m ול-$\hat{s}_o(t) = \lceil (tI-q)/J \rceil$ עבור $\hat{H}_\ell$.

מכיון ש- $I \ge J$ הרי $s_o(t+1) \ge s_o(t) + 1$ (ובאופן דומה ל-$\hat{s}_o(t)$), לכן המטריצות H_m ו-$\hat{H}_\ell$ יראו למשל:



ברור לכן שאם קיים t עבורו $s_o(t) \ne \hat{s}_o(t)$ אזי $\text{Image}(H_m) \ne \text{Image}(\hat{H}_\ell)$, ללא תלות בערכי האברים השונים מאפס.

נתון ש-$q \ne v$, ונניח ללא הגבלת הכלליות ש-$v < q$. מאחר ו-$\gcd(I,J) = 1$, הרי $(tI)\bmod J$ עבור $1 \le t \le J$ עובר על כל המספרים בתחום $0,\dots,J-1$ בדיוק פעם אחת, ולכן קיים t_o כך ש-$(t_oI-v) = 0 \bmod J$ ולכן:

$$\hat{s}_o(t_o) = \left\lceil \frac{t_oI-q}{J} \right\rceil \ge \left\lceil \frac{t_oI-v}{J} \right\rceil + 1 = s_o(t_o) + 1$$

וזה בדיוק מה שדרוש לסיום ההוכחה.

מ.ש.ל.

<u>מסקנה (מתוך הוכחת טענה 13):</u>

לכל תת-מטריצה שמאלית-עליונה של $\hat{H}_\ell$ (או H_m) ממימד $\Lambda \times \varepsilon$ כש- $\Delta \ge \hat{s}_o(\varepsilon)$ ישנה דרגה ε.

<u>טענה 14</u>: אם משוואה (37) מתקיימת עבור צמד $1 \le m,\ell \le M$, כש-$\lceil \ell/g \rceil = \lceil m/g \rceil$, וזאת לכל $1 \le r \le J$, הרי משוואה (37) תתקיים גם עבור הצמדים $\ell+gv$, $m+gv$, לכל $1 \le r \le J$, וזאת על ידי אותו מסנן אנליזה $\hat{h}(\cdot)$.

<u>הוכחה</u>: הנתון (על פי (37)) הוא ש-

$$\sum_{t=1}^{\infty} \hat{h}(\ell+g(sJ-tI))z(\ell,m,t,r) = h(m+g(sJ-rI))$$

וזאת לכל $1 \ge r \ge J$, ולכל $-\infty < s < \infty$ (כי עבור $s \le 0$, שני האגפים הם אפס זהותית, כפי שקל לודא).

לכל v ולכל r, קיים $1 \ge r' \ge J$ כך ש-$(r'I-(rI-v)) \equiv 0 \bmod J$ (וזאת כי I ו-J זרים).

עתה נסמן $\varepsilon = (r'-r)$, אזי $r'I-(rI-v) = \varepsilon I+v = \Delta J$ כש-Δ שלם.

לכן: $gv = g(\Delta J-\varepsilon I)$.

לפיכך עבור $\varepsilon \le 0$:

$$\sum_{t=1-\varepsilon}^{\infty} \hat{h}(\ell+gv+g(sJ-tI))z(\ell,m,t+\varepsilon,r') = \sum_{t=1-\varepsilon}^{\infty} \hat{h}(\ell+g((s+\Delta)J-(t+\varepsilon)I))z(\ell,m,t+\varepsilon,r') =$$

$$\underset{\substack{\uparrow \\ \text{ע"פ הנתון עבור } r'}}{=} h(m+g((s+\Delta)J-r'I)) \underset{\substack{\uparrow \\ \text{ע"פ הגדרת } \Delta}}{=} h(m+gv+g(sJ-rI))$$

מכאן שעל ידי שימוש ב-

$$z(\ell+gv,m+gv,t,r) \triangleq \begin{cases} z(\ell,m,t+\varepsilon,r') & t \ge 1-\varepsilon \\ 0 & t < (1-\varepsilon) \end{cases}$$

הוכחנו את הטענה למקרה של $\varepsilon \le 0$.

כאשר $\varepsilon > 0$ הטענה מתקבלת באותו אופן ובלבד שנוכיח ש-$z(\ell,m,t,r') = 0$ עבור $1 \le t \le \varepsilon$.

נבחין עתה בכך ש-$0 < \varepsilon < J$ (ברור כי $J \geq r,r' \geq 1$), וכן $\hat{s}_o(\varepsilon) = s_o(\varepsilon) \geq \Delta < I$,

וזאת כי $\lceil m/g \rceil = \lceil \ell/g \rceil \Leftarrow \hat{s}_o(\varepsilon) = s_o(\varepsilon)$ ו-$J > \varepsilon \Leftarrow I > \Delta$ כי

$(I-1) \leq v$, וכן $\hat{s}_o(\varepsilon) \geq \Delta$ כי $v \geq -(\lceil \ell/g \rceil - 1)$.

בנוסף

$$m + g(\Delta J - r'I) = (m+gv) - rM \leq 0$$

ולכן $h(m+g(sJ-r'I)) = 0$ עבור $1 \leq s \leq \Delta$.

באופן דומה

$$\ell + g(\Delta J-(\varepsilon+1)I) = (\ell+gv) - M \leq 0$$

ולכן $\hat{h}(\ell+g(sJ-tI)) = 0$ עבור $1 \leq s \leq \Delta$ ו-$t > \varepsilon$.

לכן מערכת המשוואות הנתונה, מצטמצמת עבור $1 \leq s \leq \Delta$, לתת-מטריצה שמאלית עליונה של $\hat{H}_\ell$ ממימד $\varepsilon \times \Delta$ מוכפלת ב-$\{z(\ell,m,t,r')\}_{t=1}^{\varepsilon}$ ושערכה הוא הוקטור $\underline{0} \in \mathbb{C}^\Delta$.

מהמסקנה מהטענה הקודמת, לתת-מטריצה זו יש דרגה ε ולכן מתקיים בהכרח

$z(\ell,m,t,r') = 0$, $1 \leq t \leq \varepsilon$.

מ.ש.ל.

<u>משפט 8</u>: עבור $R = M$, <u>כל</u> מטריצה $\hat{P}$ רגולרית היא ULM, כאשר לכל מסנן $\{h(n)\}_{n=o}^{\infty}$, המסנן המתאים יהא $\hat{h}=h$.

<u>הוכחה</u>: זו למעשה מסקנה ישירה של משפט 1 בסעיף II, והיא מובאת כאן רק למטרות שלמות ההצגה. כאשר $R = M$ ולמסנן $h(\cdot)$ יש ftUS, הרי קיימת עבורו גם DUS. לפיכך עבור $\hat{h} = h$, לכל MDSTT קיימת סדרה זמנית $y(\cdot)$ שה-DSTT שלה (על ידי $\hat{h}=h$) מזדהה עם ה-MDSTT הנתון (זו למשל הסדרה הנוצרת על ידי ה-DUS), ולכן לכל מטריצה $\hat{P}$ משוואה (35) מתקיימת לכל $h(\cdot)$, קרי כל מטריצה $\hat{P}$ רגולרית היא ULM.

<u>הערה</u>: במקרה זה מערכת הצפנה / פענוח תבוסס על אנליזה עם המסנן $h(\cdot)$ הן במשדר והן במקלט וסינתזה על ידי ה-DUS שקיים עבור מסנן זה (שהוא גם ftUS עבור מסנן זה).

<u>משפט 9</u>: עבור $R < M$ קיים מסנן אנליזה $h(\cdot)$ שעבורו כל מטריצה $\hat{P}$ שהיא Block-Diagonal עם בלוקים ממימד $g \times g$ מייצגת LM. בפרט אם מסנן זה הוא FIR, אזי אורכו L_h הוא כפולה שלמה של g, וכן ה-LM היא ביחס למסנן אנליזה שני $\hat{h} = h$.

<u>הוכחה</u>: מסנן האנליזה $h(\cdot)$ המופיע במשפט יהא בעל g-יות דגמים שערכן קבוע, קרי $\{h_m(x)\}_{x=0}^{\infty}$ ב"ת ב-m, עבור $1 \leq m \leq g$. לכן אם זהו מסנן FIR ברור שאורכו הוא כפולה שלמה של g. מאחר ו-$\hat{h} = h$ הרי ש-$\hat{h}_\ell(x) = h_m(x)$ לכל $x \geq 0$ ולכל צמד ℓ,m כך ש-$\lceil m/g \rceil = \lceil \ell/g \rceil$. לפיכך, קל להראות שלכל צמד כזה יש פתרון ל-(37) על ידי בחירת $z(\ell,m,t,r) = \delta(t-r)$. לאור המסקנה מטענה 11, נובע שכל מטריצה $\hat{P}$ שהיא Block-Diagonal עם בלוקים ממימד $g \times g$, מייצגת LM.

<u>משפט 10</u>: עבור $R < M$, מחלקת המודיפיקציות $\hat{P}$ שהן ULM, ניתנת על ידי

$$\hat{P} = \Lambda \begin{bmatrix} \pi & 0 & 0 \\ 0 & \ddots\ \pi & 0 \\ 0 & 0 & \ddots\ \pi \end{bmatrix}$$

כאשר Λ היא מטריצה אלכסונית (רגולרית) ואילו π זו מטריצת פרמוטציה ממימד $g \times g$.

המסנן $\hat{h}(\cdot)$ מוגדר במקרה זה על ידי $\hat{h}(u+gv) = h(\pi(u)+gv)$, $1 \leq u \leq g$, $v \geq 0$.

יתר על כן, לכל מודיפיקציה $\hat{P}$ שאינה ULM ישנו מסנן אנליזה $h(\cdot)$ שהוא FIR באורך $2R$, שהיא אינה LM עבורו.

<u>הוכחה</u>:

(א) נראה שמטריצות $\hat{P}$ מהצורה שניתנה במשפט הן אכן ULM.

במטריצות כאלה עבור $\ell = u+gv$ הרי $I_\ell = \{\pi(u) + gv\}$. קל לבדוק שעבור $\hat{h}(\cdot)$ המוגדר במשפט לעיל, מתקיימת משוואה (37) עבור כל $m \in I_\ell$, על ידי בחירת $z(\ell,m,t,r) = \delta(t-r)$.

לכן לאור המסקנה מטענה 11 המטריצות $\hat{P}$ הללו הן ULM (כי הן LM ביחס לכל מסנן $h(\cdot)$).

(ב) על מנת להשלים הוכחת המשפט תנתאר מסנן $h(\cdot)$ שהוא FIR מאורך 2R, בעל ftUS, כך שכל מודיפיקציה $\hat{P}$ שבה ישנו ℓ כך ש-I_ℓ אינו איבר בודד, היא לא LM עבורו. מזה כבר נובע שעבור מסנן זה כל LM היא מהצורה :

$$\hat{P} = \Lambda \left[\begin{array}{c|c|c} \pi_0 & 0 & 0 \\ \hline 0 & \pi_1 & 0 \\ \hline 0 & 0 & \ddots\ \pi_{I-1} \end{array}\right]$$

כשהמטריצות $\pi_{I-1},\ldots,\pi_0$ הן מטריצות פרמוטציה ממימד $g \times g$.

נותר להוכיח שכל ה-π_i חייבים להיות שווים זה לזה על מנת ש-$\hat{P}$ תהא LM .

נניח בשלילה שקיימים $\pi_i(u) \neq \pi_j(u)$, כך ש-$\hat{P}$ שכזו היא עדיין LM, ביחס ל-$h(\cdot)$. קרי, לאור המסקנה מטענה 11 ישנם $\hat{h}(\cdot)$ ו-$z(\cdot)$ שפותרים את (37) לכל $m \in I_\ell$, ובפרט ל-$\ell = u+gj$, $m = \pi_j(u)+gj$. לאור טענה 14, הרי (37) מתקיימת במקרה זה גם עבור הצמד $\ell = u+gi$, $m = \pi_j(u)+gi$, וזאת על ידי אותו מסנן אנליזה $\hat{h}(\cdot)$. לפיכך עבור $\ell = u+gi$ ישנו מסנן $\hat{h}(\cdot)$ כך שעבורו (37) מתקיימת הן ביחס ל-$\pi_j(u)+gi$ והן ביחס ל-$\pi_i(u)+gi$. לאור המסקנה מטענה 11, הרי שמודיפיקציה $\hat{P}$ בעלת $I_{u+gi} = \{\pi_i(u)+gi, \pi_j(u)+gi\}$ תהא LM ביחס ל-$h(\cdot)$, על ידי שימוש במסנן $\hat{h}(\cdot)$ הנ"ל. אך זו סתירה לטענתנו המקורית שעבור $h(\cdot)$ שבנינו, אם I_ℓ אינו איבר בודד הרי ש-$\hat{P}$ אינה LM.

(ג) לפני שנבנה את המסנן $h(\cdot)$ המבוקש, נציין שמספיק להוכיח ש-I_ℓ חייב להיות איבר בודד עבור $g(I-J)+1 \le \ell \le g(I-J)+g$, וזאת כמובן לאור טענה 14.

(ד) לאור טענה 13 והמסקנה מטענה 12, מספיק לתאר מסנן $h(\cdot)$ בעל התכונה ש-$Image(H_m)$ עבור $g(I-J)+1 \le m \le g(I-J)+g$, מתאר g תת-מרחבים לינאריים <u>שונים זה מזה ב-$\mathbb{C}^I$</u>.

(ה) עתה נבחין בכך ש-H_m זו מטריצה ממימד $J \times I$ כשמאחר ו-$R < M$ הרי $I > J \ge 1$. יתר על כן, עבור $h(\cdot)$ שהוא FIR מאורך 2R כשכל 2R דגמיו אינם אפסים, הרי כל עמודה של H_m (פרט אולי לעמודה האחרונה) מכילה בדיוק שני אברים עוקבים שונים מאפס, בשורות $s_o(t)$ ו-$s_o(t)+1$ (עבור העמודה ה-t-ית). לשם הבהרת טענה זו ניתן לעיין בהוכחת טענה 13. מאחר וכאמור שם $s_o(t+1) \ge s_o(t)+1$, וכן $s_o(1) = 1$ (כי $m \ge g(I-J)$!), וכן $s_o(J) > J$ עבור $J > 1$ (כי $m \le g(I-J)+g$), הרי קיים $1 \le t \le J$ כך ש-$s_o(t)+2 \le s_o(t+1)$. לכן אם נבחר את $h(m+g(s_o(t)J-tI))$ להיות קבוע עבור $g(I-J)+1 \le m \le g(I-J)+g$,

ואת ה-g-יה הבאה $h(m+g((s_o(t)+1)J-tI))$ להיות בעלת g ערכים שונים, הרי מובטח לנו (עבור $J > 1$) שהטווח של כל אחת מ-g המטריצות H_m הנ"ל שונה מהטווח של יתר $(g-1)$ המטריצות, וזאת ללא תלות בערך של יתר דגמי **המסנן** $h(\cdot)$. כאשר $J = 1$, הרי מאחר ו-$I > 1$, ניתן להשיג אותה תוצאה על ידי הבניה דלעיל, עם $s_o(t) = t = 1$.

מ.ש.ל.

DESIGN OF DIGITAL FIR FILTER ARRAYS

Research Thesis

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Science

AMIR DEMBO

Submitted to the Senate of the Technion-Israel Institute of Technology

Nisan 5746 Haifa May, 1986

This final paper was carried out in the
Faculty of Electrical Engineering under
the supervision of Prof. DAVID MALAH.

The generous financial help of the Guttwirth Fund
is gratefully acknowledged.

I thank Prof. DAVID MALAH on his dedicated
supervision, and Mr. Yoram Or-Chen on his
encouragement during the research.

CONTENTS

			page
SUMMARY			1
LIST OF SYMBOLS AND DEFINITIONS			4
Chapter 1:	INTRODUCTION		5
	1.1	Subject Description and Importance	5
	1.2	Objectives and Structure of Thesis	12
Chapter 2:	REVIEW OF KNOWN RESULTS		14
	2.1	Design Methods of FIR Digital Filters	14
	2.2	Design Methods of Filter Banks	17
	2.3	Design Methods of Analysis/Synthesis Systems	20
Chapter 3:	DESIGN OF FILTER BANKS WITH SPECIFIED COMPOSITE RESPONSE		28
	3.1	WMMSE Design of Optimal Digital Filter Banks	28
	3.2	Existence, Uniqueness and Properties of Optimal Filter Banks for Various Error Criteria	37
Chapter 4:	DESIGN METHODS FOR UNIFORM FILTER BANKS		42
	4.1	Properties of Optimal Uniform Filter Banks	42
	4.2	WMMSE Design of Optimal Uniform Filter Banks	43
	4.3	Min-Max Design of Optimal Uniform Filter Banks	46
	4.4	The Design of Filter Banks Using an "Optimal" Window	47

		page
Chapter 5:	DESIGN OF OPTIMAL ANALYSIS/SYNTHESIS SYSTEMS WITH QUANTIZATION	53
	5.1 The Statistical Model and the Error Measures	53
	5.2 Optimal Synthesis Filters for Fine Quantization (Additive Noise) and Matrix Quantization	58
	5.3 Optimal Analysis/Synthesis Systems for Fine Quantization	61
Chapter 6:	FILTER BANKS FOR OPTIMAL SYNTHESIS OF MODIFIED SIGNALS	64
	6.1 Conditions for Existence of Unity Systems when no Modification is Applied	64
	6.2 Optimal WMMSE Synthesis of Finite Duration Signals	65
	6.3 Steady-State Optimal Synthesis for Infinite Signals	67
	6.4 Conditions for the Existence of Unity Systems when Linear Modifications are Applied	70
Chapter 7:	CONCLUSIONS AND OPEN PROBLEMS	73
REFERENCES		76
APPENDICES		
Appendix A:	WMMSE Design of Digital Filter Banks with Specified Composite Response	83
Appendix B:	Existence, Uniqueness and General Properties of a Class of Vector Approximation Problems	111
Appendix C:	The Design of Optimal Uniform Digital Filter Banks with Specified Composite Response	150
Appendix D:	Statistical Design of Analysis/Synthesis Systems with Quantization	173
Appendix E:	Optimal Synthesis from Modified Short-Time Transform	217

S U M M A R Y

Digital filter banks are widely used in digital signal processing, particularly in speech processing. Conventional filter banks are composed of finite impulse response (FIR) digital filters, since these filters can guarantee a linear phase (i.e., the outpus of all the filters correspond to the same time instance in the input signal). The design of these filters is an important issue in the area of digital signal processing and many works were reported on this subject in the last decade. The earlier works were on the "classical" design problem, which is the design of a single FIR filter whose frequency response approximates a desired response. At first, simple sub-optimal methods were proposed, and afterwords more suphisticated approaches which enable an optimal solution (under appropriate norm in the space of the frequency responses) were developed.

More recently, the interest in Analysis/Synthesis (A/S) systems increased, especially for speech processing. Thus, many of the systems for speeach enhancement, coding and recognition contain digital filter banks (which are related to the short time Fourier transform STFT). The appearence of these systems brought about the need for the design of filter banks which except for the optimality of each of their individual filters, also obey some global constraint. When the filter bank is used for analysis of the signal which is not followed by its reconstruction (synthesis) e.g., in speech recognition, speeker verification and analyzing various parameters of the speech signal, the usual constraint is a unity composite response, (i.e., the sum of the frequency responses of the filters in the filter bank is identically one). On the otherhand, in systems in which the analysis is followed by a synthesis (and therefore the system contains two filter banks), usually a unity

system (analysis + synthesis) constriant is imposed. This constraint guarantees that when no modification is applied, the output signal of the synthesis stage exactly equalls the input signal to the analysis stage (up to a constant delay). That means that the separation of the signal into its frequency components has not distorted the signal. There is much interest in filter banks which answer these specifications,or at least perform the optiaml reconstruction given a modified signal. However, due to the analytic (and numerical) difficulties in solving the design problems described above, there exists only partial (usually non-optimal) solutions to most of them.

In this work we describe original optimal solutions to the problem of design of digital FIR filter banks, (used for analysis and synthesis of signals) under various composite response specifications.

The main results of this work are:

1. For analysis filter banks, we describe the design of optimal (in the minimum weighted L_2 norm sense) filter banks having a specified composite response. The composite response is specified by an allowed tollerance (in the L_2 norm) with respect to the desired (ideal) frequency response.The use of L_2 norm has a mathematical advantage which enables an analytic solution of the design problem. Furthermore, a statistical interpretation of this design method connects it to the minimum variance criterion. The proposed design method is quite general and enables any composite response specification (not necessarily a unity response), and furthermore the filter bank need not be composed of FIR filters, but can be constructed from linear combinations of given arbitrary components. These more general structures are sometimes required due to implementation considerations (e.g., using VLSI).

2. As a first step towards the design of general filter banks which are optimal under other norms (e.g. the weighted L_∞ norm), and have a specified composite response, we present sufficient conditions for existence and uniqueness of the solution. Furthermore, we prove some general properties of this solution, such as: realness of the coefficients of the filters, sufficient conditions for phase linearity of the filters, and the relation between the allowed tollerance of the composite response error and the overall weighted error in the frequency responses of the individual filters.

3. The particular case of uniform filter banks (in which the length of all the filters is identical as well as their passband bandwidth) is highly important in signal processing, since it can be efficiently implemented using the fast Fournier transform (FFT). For this case, we prove that the optimal solution can be derived by frequency translations of an optimal real prototype filter. The general design problem is then further simplified and restated in terms of this prototype filter coefficients. As a consequence of these results we present simplified design methods for optimal uniform filter banks either in the minimum weighted L_∞ norm sense, or in the weighted L_2 norm. A sub-optimal solution, which is very close to the optimal (L_∞) solution is also derived, based on our generalizations of the "window" method for FIR filter design.

4. In A/S system that contains quantization, there is an inherent distortion in the reconstructed signal (since the quantization is a non-invertible operation). Therefore, the optimal system need not necessarily be a unity system. Since the A/S system is a time varying system, its output signal is not a wide-sense stationary process. In this part of the work we developed a statistical error criterion for non-stationary error signals, which is used to the design of optimal filter banks.

   Two different quantization models are considered - fine quantization (FQ) modelled by an additive noise, and matrix quantization (MQ) using a codebook. For both of them we present a method for the design of the optimal synthesis prototype filter. For FQ, in which case the output signal is approximately a linear function of the analysis filters, we also present an iterative algorithm for the design of optimal A/S systems, and investigate its convergence properties. We also show that as the additive noise (which models the distortion due to quantization) level approaches zero, the optimal A/S system converges to a unity system, and for some important applications the synthesis filters can be interpreted as Wiener filters.

5. For a general A/S system (not necessarily used for coding), we extend the discussion to systems based on any short time regular linear transform (not necessarily the STFT), and present necessary and sufficient conditions for the existence of unity systems. The optimal synthesis (in the min. L_2 norm sense) of a time domain sequence from a given modified (by an unknown modification) output signal of the analysis stage is solved, both for finite duration and infinite duration signals (the latter corresponds to a steady-state solution). These results extend some known results which were applied in speech enhancement and time/frequency scaling applications.

```
% DIGITAL FILTER
%
w = 50  ; % friquency of a signal [1/sec]
S = 1   ;  % Amplitude of a signal
N = S/10; % Amplitude of noise
Ts = (1/w)/10 ; % Sampling time [ sec ]
Nts = 50 ; % number of samples
w_n = 50 ; % friquency of a harmonic noise [1/sec]
M = 5 ; % filter window weidth
%%%%%%%%%%%%%%%%%%%%  INPUT DATA  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%
output = [];

 t=(0:Ts:Ts*Nts)';l=length(t);
 signal = (-1).^(floor(w*t) ) ;
 noise_rand = N*randn(l,1);
 noise_harm = N*sin(w_n*t);
  input = signal + noise_rand ;
 for i=1:l,
  upp = i+(M-1)/2 ;low = i-(M-1)/2 ;
  if (i-(M-1)/2 <=0 ) , low=1; elseif (i+(M-1)/2 >=l ) upp=l; e
nd;
  for j=low:upp,
   if (j == low) , output(i)=0; end;
   output(i) = output(i)+input(j);
  end;
  output(i) = output(i)/M ;
 end;

plot(output);
figure(2);
plot(input);
%size()
```

DESIGN OF DIGITAL FIR FILTER ARRAYS

Research Thesis

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Science

AMIR DEMBO

Submitted to the Senate of the Technion-Israel Institute of Technology
Nisan 5746 Haifa May, 1986