เทคนิคการลดเสียงสะท้อนและเสียงรบกวนสำหรับระบบสื่อสารแบบแฮนด์ฟรี

นายรัฐพล ทูลแสงงาม

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

ACOUSTIC ECHO AND NOISE REDUCTION TECHNIQUES FOR HANDS-FREE COMMUNICATION SYSTEMS

Mr. Rattapol Thoonsaengngam

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Electrical Engineering

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

Academic Year 2007

หัวข้อวิทยานิพนธ์	เทคนิคการลดเสียงสะท้อนและเสียงรบกวนสำหรับระบบสื่อสารแบบแฮนด์ฟรี
โดย	นายรัฐพล ทูลแสงงาม
สาขาวิชา	วิศวกรรมไฟฟ้า
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.นิศาชล ตั้งเสงี่ยมวิสัย

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.สมชาย จิตะพันธ์กุล)

..... อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร.นิศาชล ตั้งเสงี่ยมวิสัย)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.เจษฎา ชินรุ่งเรือง)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.วันเฉลิม โปรา)

รัฐพล ทูลแสงงาม : เทคนิคการลดเสียงสะท้อนและเสียงรบกวนสำหรับระบบการสื่อสารแบบแฮนด์ฟรี (ACOUSTIC ECHO AND NOISE REDUCTION TECHNIQUES FOR HANDS-FREE COMMUNICATION SYSTEMS) อ. ที่ปรึกษา : ผศ.ดร.นิศาชล ตั้งเสงี่ยมวิสัย, 118 หน้า.

ระบบการสื่อสารทางเสียงแบบแฮนด์ฟรีถูกพัฒนาขึ้นเพื่อความสะดวกสบายของผู้ใช้งานเป็นหลัก โดยอุปกรณ์สื่อสารปลายทาง ซึ่งได้แก่ ลำโพงและไมโครโฟนจะถูกติดตั้งไว้ ณ บริเวณที่ต้องการทำการสนทนาแทนหูโทรศัพท์ในระบบการสื่อสารแบบเดิม โครงสร้างเช่นนี้นำมาซึ่งปัญหาเสียงก่อกวนที่ถูกพิจารณาในวิทยานิพนธ์ฉบับนี้สองชนิด ได้แก่ เสียงสะท้อนและเสียงรบกวน ดังนั้นวิธีการที่สามารถลดปริมาณเสียงก่อกวนทั้งสองลงได้จึงมีความจำเป็นอย่างยิ่งในระบบการสื่อสารทางเสียงแบบแฮนด์ฟรี

วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการเพิ่มประสิทธิภาพให้กับการลดเสียงสะท้อนและเสียงรบกวนที่ขึ้นกับเทคนิคการกดทางสเปกตรัม ซึ่งมีความซับซ้อนในการคำนวณต่ำเหมาะสำหรับงานประยุกต์เช่นระบบการสื่อสารแบบแฮนด์ฟรี โดยเสนอขั้นตอนการปรับปรุงการประมาณค่าความหนาแน่นสเปกตรัมกำลังเสียงสะท้อนที่อาศัยหลักการของการหักล้างเสียงสะท้อน นอกจากนี้ยังได้ทำการพัฒนาวิธีการประมาณค่าอัตราส่วนสัญญาณต่อสัญญาณก่อกวนก่อนประสบ (a priori SDR) ในส่วนของเทคนิคการกดทางสเปกตรัม โดยทั้งนี้สมการช่วงเปลี่ยนที่นิยามขึ้นถูกใช้เป็นเครื่องมือในการวิเคราะห์และออกแบบการประมาณค่า a priori SDR ดังกล่าว ผลการทดลองเชิงปริวิสัยจากการจำลองด้วยระบบคอมพิวเตอร์ ชี้ให้เห็นว่าวิธีการที่นำเสนอสามารถลดเสียงสะท้อนลงในปริมาณที่มากกว่าวิธีการเดิม นอกจากนี้ผลการทดลองทั้งเชิงปริวิสัย และเชิงอัตวิสัย ยังชี้ให้เห็นอีกว่าวิธีการที่นำเสนอสามารถลดผลของความผิดเพี้ยนของเสียงพูดลงจากวิธีการลดเสียงสะท้อนและเสียงรบกวนที่ขึ้นกับเทคนิคการกดทางสเปกตรัมแบบเดิมได้

ภาควิชา วิศวกรรมไฟฟ้า ลายมือชื่อนิสิต .....

สาขาวิชา วิศวกรรมไฟฟ้า ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2550

4870437021 : MAJOR ELECTRICAL ENGINEERING
KEYWORD : NOISE SUPPRESSION / SPEECH ENHANCEMENT / ACOUSTIC ECHO SUPRESSION / COMBINED ACOUSTIC ECHO SUPPRESSION AND NOISE SUPPRESSION / DISTURBANCE REDUCTION BASED ON SPECTRAL SUPPRESSION TECHNIQUE

RATTAPOL THOONSAENGNGAM : ACOUSTIC ECHO AND NOISE REDUCTION TECHNIQUES FOR HANDS-FREE COMMUNICATION SYSTEMS. THESIS ADVISOR : ASST.PROF.NISACHON TANGSANGIUMVISAI, Ph.D., 118 pp.

Hands-free communication systems have been designed by considering mainly on "comfortable" use. A hands-free terminal comprising of a loudspeaker and a microphone set up around the conversation area is used instead of a conventional telephone terminal or a handset. This structure, however, brings about two kinds of disturbances namely acoustic echo and noise, which will be considered in this thesis. Hence, an approach to alleviate these problems is necessary for hands-free communication systems.

In this thesis, the proposed acoustic echo and noise reduction technique is based upon the spectral suppression method, which requires low computational complexity. Thus, it is suitable for hands-free communication systems. An echo power spectral density (EPSD) estimation using a principle of acoustic echo cancellation (AEC) has been introduced. In addition, a priori signal-to-disturbance ratio (SDR) estimation has been developed using transition equation analysis. Experiments carried out using computer simulations show that the proposed technique not only gives higher echo attenuation but also reduces speech distortion as compared to the conventional technique.

Department Electrical Engineering Student's signature .....
Field of study Electrical Engineering Advisor's signature .....
Academic year 2007

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี ข้าพเจ้าใคร่ขอกราบขอบพระคุณอย่างสูงสำหรับความช่วยเหลืออย่างดียิ่งของ ผศ.ดร.นิศาชล ตั้งเสงี่ยมวิสัย อาจารย์ที่ปรึกษา วิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่างๆ พร้อมทั้งแรงกระตุ้นและแรงบันดาลใจในการทำวิจัยมาด้วยดีตลอดมา

ขอกราบขอบพระคุณครูอาจารย์ทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้ พร้อมทั้งคุณธรรมจริยธรรม ตลอดชั่วชีวิตของขาพเจ้า

ขอขอบคุณเพื่อน รุ่นพี่ รุ่นน้อง และผู้ที่อยู่รอบตัวข้าพเจ้าทุกท่าน สำหรับความช่วยเหลือ และกำลังใจในการทำวิจัยเสมอมา ขอขอบคุณที่ทำให้วันเวลาของข้าพเจ้าผ่านไปอย่างมีความหมาย

ขอขอบคุณทุนวิจัยกองทุนรัชดาภิเษกสมโภชน์จุฬาลงกรณ์มหาวิทยาลัยที่ช่วยสนับสนุนในการทำวิจัยเป็นอย่างดี

สุดท้ายนี้ขอกราบขอบพระคุณบิดามารดา และครอบครัว ตลอดจนญาติพี่น้องทุกท่านที่เป็นกำลังใจและให้การสนับสนุนแก่ข้าพเจ้าตลอดมา

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

หน้า

ดัชนีคำศัพท์

A priori	ก่อนประสบ
Acoustic Echo Reduction (AER)	วิธีการ/การลดเสียงสะท้อน
Acoustic Echo Cancellation (AEC)	วิธีการ/การหักล้างเสียงสะท้อน
Acoustic Echo Suppression (AES)	วิธีการ/การกดเสียงสะท้อน
(AES คือ AER ที่อาศัย Spectral Suppression Technique ในการลดเสียงสะท้อน)	
Acoustic Echo and Noise Reduction (AENR)	วิธีการ/การลดเสียงรบกวนและเสียงสะท้อน
Acoustic Echo and Noise Suppression (AENS)	วิธีการ/การกดเสียงรบกวนและเสียงสะท้อน
(AENS คือ AENR ที่อาศัย Spectral Suppression Technique ในการลดเสียงก่อกวนทั้งสอง)	
Adaptive Filter	วงจรกรองปรับตัว
Amplitude of Spectral Speech Estimate	ค่าประมาณขนาดสเปกตรัมเสียงพูด
Clean Speech	เสียงพูดสะอาด
Combined System	ระบบร่วม
Detector	ตัวตรวจหา
Disturbance Signal	สัญญาณเสียงก่อกวน
Echo Power Spectral Density (EPSD)	ความหนาแน่นสเปกตรัมกำลังเสียงสะท้อน
Echo Signal	สัญญาณเสียงสะท้อน
Enhanced Speech	เสียงพูดที่ถูกปรับปรุง
Estimate	ค่าประมาณ
Far-end Signal	สัญญาณทางห้องไกล
Far-end Speech	เสียงพูดทางห้องไกล
Finite Impulse Response	ผลตอบสนองอิมพัลส์จำกัด
Frame step	ช่วงก้าวระหว่างเฟรม
Frequency-bin	องค์ประกอบทางความถี่
Hypothesis	สมมติฐาน
Infinite Impulse Response	ผลตอบสนองอิมพัลส์อนันต์
Input	สัญญาณขาเข้า
Linear Time-Invariant System (LTI system)	ระบบเชิงเส้นไม่แปรผันตามเวลา
Magnitude	ขนาด
Microphone Signal	สัญญาณไมโครโฟน

Multiple-microphone noise reduction	วิธีการ/การลดเสียงรบกวนที่อาศัยไมโครโฟนหลายตัว
Near-end Signal	สัญญาณทางห้องใกล้
Near-end Speech	เสียงพูดทางห้องใกล้
Noise Reduction (NR)	วิธีการ/การลดเสียงรบกวน
Noise Suppression (NS) (NS คือ NR ที่อาศัย Spectral Suppression Technique ในการลเสียงรบกวน)	วิธีการ/การกดเสียงรบกวน
Noise Signal	สัญญาณเสียงรบกวน
Noisy Speech Signal	สัญญาณเสียงพูดที่ถูกก่อกวน
Noise Power Spectral Density (NPSD)	ความหนาแน่นสเปกตรัมกำลังเสียงรบกวน
Objective Test	การทดสอบเชิงปริวิสัย
Phase	เฟส
Posterior	หลังประสบ
Power Spectrum	สเปกตรัมกำลัง
Probability Density Function	ฟังก์ชันความหนาแน่นความน่าจะเป็น
Probability Distribution	การแจกแจงความน่าจะเป็น
Pulse	พัลส์สัญญาณ
Single-microphone noise reduction	วิธีการ/การลดเสียงรบกวนที่อาศัยไมโครโฟนหนึ่งตัว
Signal	สัญญาณ
Spectral Amplitude	ขนาดสเปกตรัม
Spectral Component of Echo	สเปกตรัมเสียงสะท้อน
Spectral Component of Noise	สเปกตรัมเสียงรบกวน
Spectral Component of Noisy Speech	สเปกตรัมเสียงพูดที่ถูกก่อกวน หรือสเปกตรัมสัญญาณไมโครโฟน
Spectral Component of Speech	สเปกตรัมเสียงพูด
Spectral Power	สเปกตรัมกำลัง
Spectral Speech Estimate	ค่าประมาณสเปกตรัมเสียงพูด
Spectral Suppression	วิธีการ/การกดทางสเปกตรัม
Spectral Suppression Technique	เทคนิคการกดทางสเปกตรัม
Speech Signal	สัญญาณเสียงพูด
Speech Estimate	ค่าประมาณเสียงพูด
Speech Distortion	ความผิดเพี้ยนของเสียงพูด
Speech Enhancement	การเพิ่มสมรรถนะเสียงพูด
Stationary Noise	เสียงรบกวนจุดนิ่ง

Subjective Test	การทดสอบเชิงอัตวิสัย
Time frame index	ดัชนีบ่งบอกเฟรมเวลา
Time-frequency index	ดัชนีบ่งบอกเวลาความถี่

หมายเหตุ

1. คำว่า “สัญญาณ” ถูกใช้เมื่อต้องการเน้นว่าเป็นสัญญาณไฟฟ้าของสิ่งที่ถูกขยาย เช่น สัญญาณเสียงพูด หมายความถึงสัญญาณไฟฟ้าของเสียงพูดที่อยู่ในสายไฟหรือถูกรับมาโดย Sensor เรียบร้อยแล้ว เป็นต้น
2. หากใช้เฉพาะคำว่า “เสียงพูด” จะหมายถึงเสียงพูดในอากาศทั่วไป
3. ชื่อของวิธีการต่างๆ จะไม่ใช้คำว่า “สัญญาณ” แม้ว่าในความหมายแล้วคือสัญญาณทางไฟฟ้าก็ตาม ทั้งนี้เพื่อความกระชับในการเรียบเรียง และเนื่องจากวิธีการที่เกี่ยวข้องต่างๆ ในวิทยานิพนธ์ฉบับนี้เกี่ยวข้องกับเฉพาะสัญญาณเท่านั้นดังนั้นจึงเป็นการสะดวกกว่าในการละไว้ในฐานที่เข้าใจ อย่างไรก็ตามคำขยายความของวิธีการจะเป็นสิ่งซึ่งบ่งบอกถึงความหมายดังกล่าวอีกครั้งหนึ่ง

บทที่ 1

บทนำ

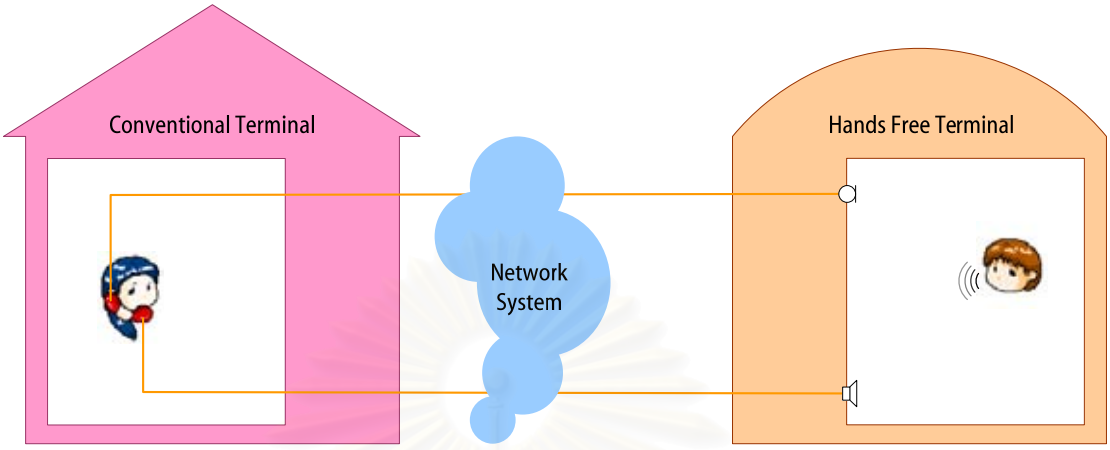
1.1. ความสำคัญของปัญหา

มนุษย์ใช้เสียงในการสื่อสารมาเป็นเวลายาวนาน ตั้งแต่ยังไม่รู้จักภาษา จนกระทั่งทุกวันนี้ภาษาได้กลายเป็นสิ่งที่สำคัญเป็นอย่างยิ่งแก่อารยธรรมของมนุษยชาติ เสียงที่ถูกเปล่งเป็นภาษาออกมาเรียกว่า “เสียงพูด” (Speech) เสียงพูดทำให้มนุษย์สามารถสื่อสารโต้ตอบกันได้อย่างที่สิ่งมีชีวิตชนิดอื่นไม่สามารถทำได้ และยังช่วยให้มนุษย์สามารถจดจำกันและกันได้ จากผลงานของ Alexander Graham Bell ทำให้มนุษย์สามารถแปรเปลี่ยนไปมาระหว่างเสียงพูดกับสัญญาณทางไฟฟ้า ปรากฏการณ์ดังกล่าวนี้ไม่เพียงทำให้เสียงพูดสามารถเดินทางไปได้ไกลยิ่งขึ้น ผู้ป่วยทางโสตสามารถได้ยินเสียงพูดที่ชัดเจนมากยิ่งขึ้น แต่ยังนำมาซึ่งความพยายามจำนวนมากที่จะทำให้ผู้ที่สามารถเข้าใจในเสียงพูดนี้ไม่จำกัดอยู่แค่เพียงมนุษย์ งานประยุกต์ที่ตรงกับความสามารถที่เพิ่มขึ้นข้างต้นได้แก่ ระบบโทรศัพท์ (Telephone System) เครื่องช่วยฟัง (Hearing Aids) ระบบรู้จำเสียงพูด (Auto Speech Recognition System) และระบบรู้จำผู้พูด (Speaker Recognition System) เป็นต้น จะเห็นว่าเทคโนโลยียิ่งก้าวหน้ามากเพียงใดเสียงพูดก็ยิ่งมีบทบาทกับมนุษย์มากขึ้นเท่านั้น ตราบเท่าที่มนุษย์ยังคงพูด

ระหว่างการสื่อสารทางเสียงหรือการสนทนา สัญญาณเสียงที่ไม่พึงประสงค์มักเข้ามามีส่วนร่วมเสมอ วิทยานิพนธ์ฉบับนี้จะเรียกเสียงที่ไม่พึงประสงค์นี้ว่า เสียงก่อกวน (Acoustic Disturbance) เสียงก่อกวนทำให้เกิดความคลาดเคลื่อนในการสื่อสารขึ้นได้กับทั้งผู้รับสารที่เป็นมนุษย์และมีผลอย่างมากกับผู้รับสารเช่นระบบรู้จำเสียงพูด หรือกล่าวคือคอมพิวเตอร์ อย่างไรก็ตามด้วยความซับซ้อนของกลไกสมองและโสตประสาทของมนุษย์ มนุษย์จึงสามารถทนต่อระดับเสียงก่อกวนที่สูงกว่าคอมพิวเตอร์มาก เช่น ที่ระดับเสียงก่อกวนที่เท่ากัน มนุษย์สามารถตีความหมายของเสียงพูดที่ถูกก่อกวนได้อย่างถูกต้องกว่าเครื่องคอมพิวเตอร์มาก เป็นต้น อย่างไรก็ตามความรู้สึกเป็นสิ่งที่เกิดขึ้นในมนุษย์เท่านั้น และเสียงก่อกวนนำมาซึ่งความรู้สึกที่ไม่ประทับใจนัก ดังนั้นระบบซึ่งสามารถที่จะลดระดับของเสียงก่อกวนลงได้ หรืออาจเรียกว่าระบบการเพิ่มสมรรถนะเสียงพูดที่ถูกก่อกวน จึงเป็นระบบที่จำเป็นต่อทั้งผู้รับสารที่เป็นมนุษย์และผู้รับสารที่เป็นคอมพิวเตอร์ ในจุดประสงค์ที่แตกต่างกัน

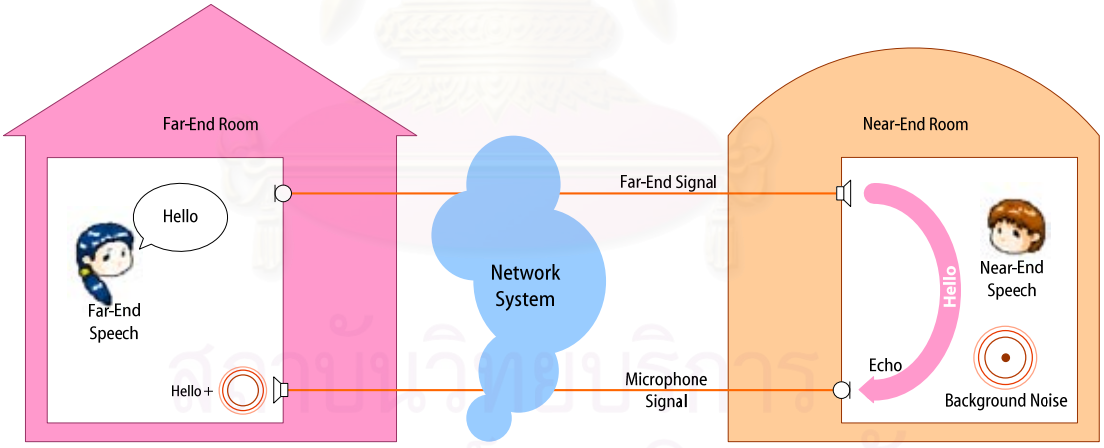
ระบบสื่อสารทางเสียง อันได้แก่ ระบบโทรศัพท์ ได้รับการพัฒนาอย่างรวดเร็วในช่วงหลายทศวรรษที่ผ่านมา ทั้งนี้เนื่องจากระบบโทรศัพท์สามารถเข้าถึงความต้องการของมนุษย์ได้อย่างกว้างขวาง โดย ณ ปลายทางของระบบโทรศัพท์ (Telephone Terminal) ทั่วไป อุปกรณ์สื่อสารปลายทาง อันได้แก่ หูโทรศัพท์ (Handset) ซึ่งประกอบด้วยลำโพงและไมโครโฟน จะถูกใช้เป็นตัวแปลงสัญญาณเสียงทางไฟฟ้าให้กลายเป็นเสียง และจากเสียงให้กลายเป็นสัญญาณไฟฟ้า ตามลำดับ โดยทั้งนี้ผู้สนทนาจะต้องถือหรือ พกพาอุปกรณ์สื่อสารปลายทางดังกล่าวเอาไว้กับตัว ทำให้เกิดความไม่อิสระในการสนทนาขึ้นดังแสดงไว้ในรูปที่ 1.1 (ซ้ายมือ) ความไม่อิสระดังกล่าวทำให้ผู้ออกแบบเกิดความพยายามที่จะพัฒนารูปแบบการสนทนาแบบใหม่ โดยมีจุดประสงค์ให้ผู้สนทนามีความอิสระในการสนทนามากยิ่งขึ้น รูปแบบใหม่สำหรับการสนทนาที่ถูกพัฒนาขึ้นนี้ได้แก่ ระบบสื่อสารทางเสียงที่มีปลายทางแบบแฮนด์ฟรี (Hands Free Terminal) กล่าวคือหูโทรศัพท์จะถูกแทนที่ด้วยลำโพงและไมโครโฟนซึ่งถูก

ติดตั้งไว้ภายในบริเวณที่ทำการสนทนา (ห้องใกล้) ดังรูปที่ 1.1 (ขวามือ) ผู้สนทนาจึงไม่ต้องถือ หรือพกพาอุปกรณ์สื่อสารปลายทางไว้กับตัวดังเดิม ทำให้ผู้สนทนามีความอิสระในการสนทนามากยิ่งขึ้น



รูปที่ 1.1 การเปรียบเทียบการสื่อสารทางเสียงที่มีปลายทางแบบเดิมและปลายทางแบบแฮนด์ฟรี

แม้ว่ารูปแบบการวางตัวของอุปกรณ์สื่อสารปลายทางแบบแฮนด์ฟรีดังกล่าวจะทำให้เกิดความอิสระในการสนทนามากยิ่งขึ้นก็ตาม แต่ก็นำมาซึ่งปัญหาเสียงก่อกวนที่ไม่อาจหลีกเลี่ยงได้ 2 ประเภท ได้แก่ เสียงสะท้อน (Acoustic Echo) และ เสียงรบกวนพื้นหลัง (Background Noise)



รูปที่ 1.2 เสียงสะท้อนและเสียงรบกวนในระบบการสื่อสารแบบแฮนด์ฟรี

เมื่อพิจารณาทางห้องใกล้[1] ในการสื่อสารทางเสียงแบบแฮนด์ฟรี เสียงที่ออกจากลำโพงในห้องใกล้จะถูกขยายขนาดจากสัญญาณเสียงที่ส่งมาจากห้องไกล ดังนั้นจึงสามารถย้อนเข้าสู่ไมโครโฟนทางห้องใกล้ และส่งกลับ

[1] วิทยานิพนธ์ฉบับนี้เลือกพิจารณาปัญหาที่เกิดขึ้นทางห้องใกล้เป็นหลัก อย่างไรก็ตามการวิเคราะห์และแก้ปัญหาในลักษณะเดียวกันนี้สามารถนำไปพิจารณาใช้กับห้องไกลได้เช่นเดียวกัน

ไปยังผู้พูดทางห้องไกลได้ดังรูปที่ 1.2 ทำให้ผู้พูดทางห้องไกลได้ยินเสียงพูดของตนเอง เรียกว่า “เสียงสะท้อน” ซึ่งสร้างความไม่เป็นธรรมชาติในการสนทนาแก่ผู้พูดทางห้องไกล ดังนั้นจึงจัดว่าเสียงสะท้อนเป็นเสียงก่อกวนชนิดหนึ่ง สังเกตได้ว่าในปลายทางโทรศัพท์แบบเดิมนั้นเสียงพูดทางห้องไกลจะไม่ถูกส่งกลับไปยังห้องไกล เนื่องจากเสียงที่ออกจากลำโพงในหูโทรศัพท์ของผู้พูดทางห้องใกล้นั้นมีพลังงานค่อนข้างน้อยเมื่อเทียบกับเสียงพูดที่ออกจากลำโพงในอุปกรณ์สื่อสารปลายทางแบบแฮนด์ฟรี รวมทั้งการที่เสียงที่ออกลำโพงในหูโทรศัพท์ถูกกีดขวางด้วยศีรษะของผู้สนทนาจนทำให้ไม่สามารถเดินทางผ่านไปถึงไมโครโฟนในหูฟังโทรศัพท์เองได้

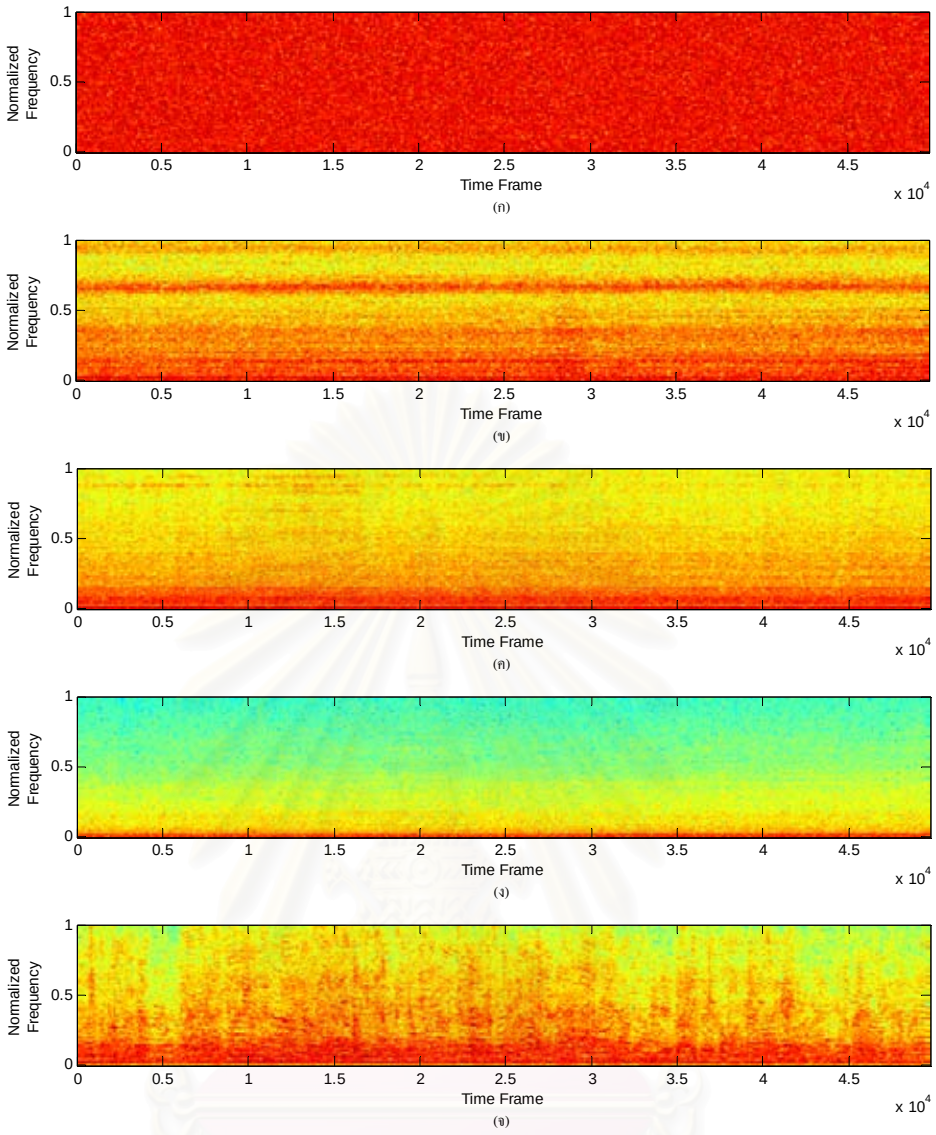
นอกจากเสียงสะท้อนที่กล่าวถึงไปแล้ว การวางตัวของไมโครโฟนทางห้องใกล้ที่อยู่ห่างจากผู้พูดทางห้องใกล้ (แหล่งกำเนิดเสียงพูด) ยังทำให้เสียงจากแหล่งกำเนิดเสียงอื่นๆ ที่ไม่ใช่ผู้พูดทางห้องใกล้สามารถเข้าสู่ไมโครโฟนทางห้องใกล้ได้ดังรูปที่ 1.2 เสียงจากแหล่งกำเนิดเสียงอื่นๆ เป็นเสียงซึ่งไม่พึงประสงค์ และถูกจัดเป็นเสียงก่อกวนอีกชนิดหนึ่ง ซึ่งเรียกว่า “เสียงรบกวนพื้นหลัง” ตัวอย่างของเสียงรบกวนพื้นหลังได้แก่ เสียงเครื่องปรับอากาศในห้องประชุม เสียงเครื่องยนต์ของยานพาหนะที่เล็ดรอดเข้ามาภายในห้องโดยสาร เสียงฝนตก หรือแม้แต่กระทั่งเสียงผู้คนจอแจในงานเลี้ยงสังสรรค์ เป็นต้น วิทยานิพนธ์ฉบับนี้พิจารณาเฉพาะเสียงรบกวนพื้นหลังที่มีความเป็นจุดนิ่ง หรือเรียกว่า “เสียงรบกวนจุดนิ่ง” (Stationary Noise) ซึ่งเป็นเสียงก่อกวนที่มีคุณลักษณะทางสถิติ เช่น ค่าเฉลี่ย และค่าความแปรปรวน เป็นต้น คงที่ติดต่อกันยาวนานเท่านั้น ตัวอย่างของเสียงรบกวนจุดนิ่งได้แก่ เสียงเครื่องยนต์ เสียงฝนตก และ เสียงเครื่องปรับอากาศ เป็นต้น ตัวแทนของเสียงรบกวนจุดนิ่งที่นิยมใช้ในการจำลองสถานการณ์จริงได้แก่ เสียงรบกวนแบบเกาส์เซียนสีขาว (White Gaussian Noise, WGN) เนื่องจากเป็นเสียงรบกวนที่มีรูปแบบการแจกแจงทางความถี่แบบคงที่เท่ากันทุกความถี่ ตัวอย่างของเสียงรบกวนแบบเกาส์เซียนสีขาวที่อาจได้ยินในชีวิตประจำวันได้แก่ เสียงซ่าของสัญญาณวิทยุที่ถูกส่งเมื่อสถานีส่งเกิดขัดข้อง เป็นต้น สเปกโทรแกรมซึ่งถูกใช้บรรยาย สเปกตรัมกำลังของสัญญาณที่เวลาต่างๆ ถูกใช้แสดงตัวอย่างของเสียงรบกวนพื้นหลังชนิดต่างๆ ดังรูปที่ 1.3 (ข้อมูลสัญญาณเสียงรบกวนชนิดต่างๆ นี้ นำมาจากฐานข้อมูลเสียงรบกวน NOISEX-92 [60]) โดยจะเห็นว่าเสียงผู้คนจอแจในงานเลี้ยงมีคุณลักษณะทางสถิติ[2]ที่ไม่คงที่ดังเช่นเสียงรบกวนอื่นๆ ดังนั้นเสียงผู้คนในงานเลี้ยงจึงจัดเป็นเสียงรบกวนพื้นหลังที่ไม่เป็นจุดนิ่ง (Non-Stationary Noise)

ในอดีต วิธีการลดเสียงก่อกวนทั้งสองประเภทนี้ถูกนำเสนออย่างอิสระจากกัน โดยวิทยานิพนธ์ฉบับนี้จะเรียกวิธีการลดเสียงก่อกวนที่เป็นเสียงรบกวนพื้นหลังแบบจุดนิ่งว่า วิธีการลดเสียงรบกวน (Noise Reduction, NR) และเรียกวิธีการลดเสียงก่อกวนที่เป็นเสียงสะท้อนว่า วิธีการลดเสียงสะท้อน (Acoustic Echo Reduction, AER)

NR ถูกพัฒนาขึ้นเพื่อใช้กับสัญญาณดิจิทัลเป็นครั้งแรกเมื่อประมาณ 40 ปีก่อน ในราว ค.ศ. 1970 [1] โดยเทคนิคแรกที่ถูกนำเสนอ ได้แก่ เทคนิคการลบทางสเปกตรัม (Spectral Subtraction, SpS) [5] ซึ่งเป็นหลักการของเทคนิคการกดทางสเปกตรัม (Spectral Suppression, SS) [6]-[25] ที่เป็นที่รู้จักกันอย่างกว้างขวางและได้รับความสนใจจากนักวิจัยจำนวนมากในปัจจุบัน และโดยทั่วไปแล้ว NR ที่อาศัยเทคนิคการกดทางสเปกตรัมจะถูกเรียกว่าการกดเสียงรบกวน (Noise Suppression, NS)

---

[2] คุณลักญณะทางสถิติมีความสัมพันธ์อย่างลึกซึ้งกับค่าสเปกตรัมกำลังในแต่ละช่วงเวลาของสัญญาณนั้นๆ โดยอาจกล่าวได้อย่างสังเขปว่า สัญญาณที่ให้รูปแบบสเปกตรัมกำลังที่ต่อเรื่องยาวนานสามารถตีความได้ว่าเป็นสัญญาณที่เป็นจุดนิ่งยาวนานได้เช่นเดียวกัน

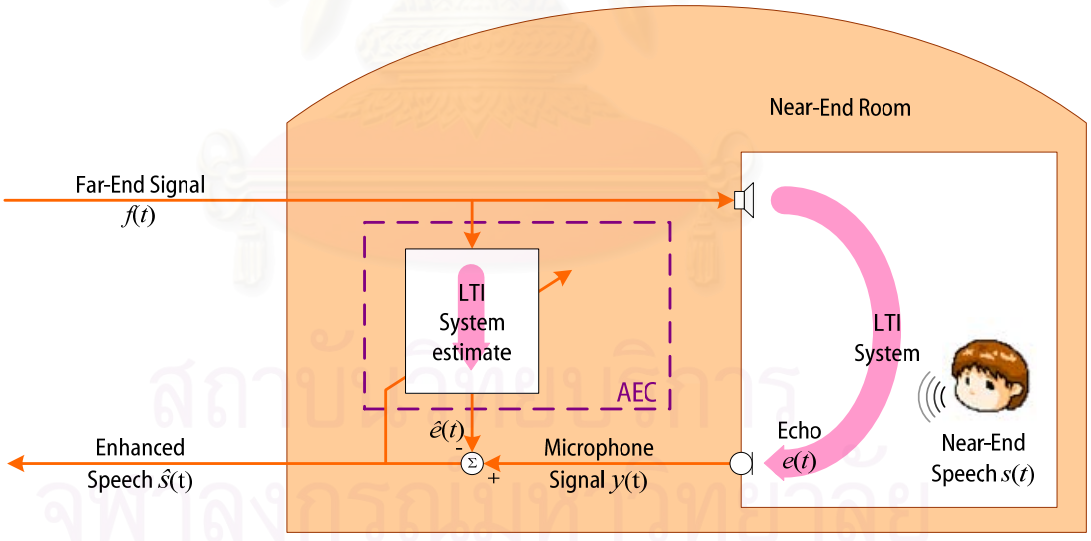


รูปที่ 1.3 สเปกโตรแกรมของเสียงรบกวนพื้นหลังชนิดต่างๆ
(ก) เสียงรบกวนสีขาว (ข) เสียงจากเครื่องยนต์ของเครื่องบิน f 16 (ค) เสียงเครื่องตัดหญ้า
(ง) เสียงเครื่องยนต์ที่เล็ดรอดเข้ามายังห้องโดยสารภายในรถยนต์ และ (จ) เสียงผู้คนจอแจในงานเลี้ยง

จากการที่ปัญหาเสียงรบกวนเป็นปัญหาที่ได้รับความสนใจจากนักวิจัยจำนวนมาก ดังนั้นจึงทำให้วิธีการลดเสียงรบกวนถูกคิดค้นและพัฒนาขึ้นเป็นจำนวนมากตามไปด้วย ซึ่งสามารถแบ่งออกได้เป็น 2 จำพวก ตามจำนวนไมโครโฟนที่ใช้ คือ พวกที่อาศัยไมโครโฟนหนึ่งตัว (Single-Microphone NR) และ พวกที่อาศัยไมโครโฟนหลายตัว (Multiple-Microphone NR) โดยในแต่ละจำพวกยังสามารถแบ่งออกได้เป็นหลายเทคนิคในการแก้ไขอีกเช่นกัน NR ที่อาศัยไมโครโฟนหนึ่งตัวใช้เพียงข้อมูลจากสัญญาณเสียงพูดที่ถูกรบกวน (Noisy Speech) ที่รับมาจากไมโครโฟนหนึ่งตัวนั้นในการปรับปรุงคุณภาพของเสียงพูด ส่วน NR ที่อาศัยไมโครโฟนหลายตัวใช้ข้อมูลสัญญาณเสียงพูดที่ถูกรบกวนจากไมโครโฟนจำนวนมากกว่าหนึ่งตัวขึ้นไป ทำให้ได้มาซึ่งข่าวสารเชิงปริภูมิ (Spatial Information) ที่มากกว่ากรณีการใช้ไมโครโฟนหนึ่งตัว และอาจให้ประสิทธิภาพที่เพิ่มขึ้นด้วย อย่างไรก็ตามวิทยานิพนธ์ฉบับนี้มุ่งเน้นไปที่ NR ที่อาศัยไมโครโฟนหนึ่งตัว เท่านั้น

AER ถูกนำเสนอขึ้นครั้งแรกในราว ค.ศ. 1979 โดยเทคนิคแรกที่ได้รับการเสนอขึ้น ได้แก่ การหักล้างเสียงสะท้อน (Acoustic Echo Cancellation, AEC) [33] ซึ่งเป็นแนวทางที่ได้รับความสนใจและพัฒนาจากนักวิจัยจำนวนมากจนกระทั่งปัจจุบัน [3], [34]-[44] นอกจาก AEC แล้ว การลดเสียงสะท้อนโดยอาศัยเทคนิคการกดทางสเปกตรัม หรือที่เรียกว่า การกดเสียงสะท้อน (Acoustic Echo Suppression, AES) [46]-[47] ก็เป็นอีกเทคนิคหนึ่งซึ่งกำลังได้รับความสนใจจากนักวิจัยเช่นกัน

การลดเสียงก่อกวนอย่างเสียงสะท้อนอาจเป็นสิ่งซึ่งกระทำได้ยาก เนื่องจากสัญญาณเสียงสะท้อนมีลักษณะทางสถิติที่คล้ายกับสัญญาณเสียงพูดที่ต้องการรักษาไว้ กล่าวคือ เสียงสะท้อนเป็นสัญญาณที่มีความเป็นจุดนิ่งเพียงช่วงสั้นๆ เช่นเดียวกับเสียงพูด แต่อย่างไรก็ตาม ในระบบการสื่อสารแบบแฮนด์ฟรีที่พิจารณาสามารถรับสัญญาณต้นกำเนิดของสัญญาณเสียงสะท้อนหรือที่เรียกว่าสัญญาณทางห้องไกล $f(t)$ ก่อนที่จะถูกขยายออกสู่ลำโพงทางฝั่งห้องใกล้ได้ (ดูรูปที่ 1.4 ประกอบ) สัญญาณ $f(t)$ นี้เป็นสัญญาณที่มีสหสัมพันธ์ (Correlation) กับสัญญาณเสียงสะท้อน $e(t)$ สูงมาก ดังนั้นหากสามารถหาความสัมพันธ์ระหว่างทั้งสองสัญญาณดังกล่าวได้ ก็จะสามารถอาศัยสัญญาณเสียงทางห้องไกล $f(t)$ ในการทำนายหรือประมาณค่าสัญญาณเสียงสะท้อน $e(t)$ ได้อย่างแม่นยำ และเมื่อนำค่าประมาณสัญญาณเสียงสะท้อน $\hat{e}(t)$ ไปหักลบออกจาสัญญาณไมโครโฟน $y(t)$ ซึ่งประกอบไปด้วยสัญญาณเสียงพูด $s(t)$ และสัญญาณเสียงสะท้อน $e(t)$ สัญญาณเสียงผลลัพธ์จึงสามารถปลอดจากสัญญาณเสียงสะท้อนได้ และทั้งนี้ยังคงความเป็นการสื่อสารสองทางเต็มอัตรา (Full Duplex Communication) เอาไว้ได้อีกด้วย วิธีการลดเสียงสะท้อนที่อาศัยหลักการข้างต้น ได้แก่ AEC ซึ่งเป็นวิธีการที่เป็นที่นิยมใช้ในการลดเสียงสะท้อนเป็นอย่างมากนั่นเอง



รูปที่ 1.4 ระบบ AEC

ประสิทธิภาพของวิธีการหักล้างเสียงสะท้อนขึ้นอยู่กับว่าสามารถจำลองความสัมพันธ์ระหว่างสัญญาณ $f(t)$ และสัญญาณเสียงสะท้อน $e(t)$ ได้ดีเพียงใด โดยการตั้งสมมติฐานว่าห้องใกล้ประพฤติตัวเป็นดังระบบเชิงเส้นที่ไม่แปรเปลี่ยนตามเวลา (Linear Time-Invariant System, LTI) วงจรกรองปรับตัว (Adaptive Filter) แบบผลตอบสนองอิมพัลส์จำกัด (Finite impulse response, FIR) ซึ่งสามารถจำลองระบบ LTI ได้อย่างมีประสิทธิภาพ

จึงถูกนำมาใช้ในการจำลองระบบห้องใกล้ดังกล่าวนี้ อย่างไรก็ตามในการจำลองหรือเลียนแบบระบบห้องใกล้วงจรกรองแบบปรับตัวที่ใช้จะต้องมีจำนวนสัมประสิทธิ์ซึ่งเหมาะสมกับระยะเวลาสะท้อนกลับ (Reverberation time) ของห้องใกล้ซึ่งขึ้นกับขนาดของห้องใกล้นั้นด้วย เช่น เวลาสะท้อนกลับสำหรับห้องโดยสารภายในรถยนต์อยู่ที่ 50-100 ms ซึ่งจำนวนสัมประสิทธิ์ของวงจรกรองปรับตัวที่เหมาะสมได้แก่ 400-800 ตัวอย่าง ณ ความถี่ชักตัวอย่าง 8 kHz และอยู่ที่ 300-500 ms สำหรับห้องสัมมนาในสำนักงาน โดยจำนวนสัมประสิทธิ์ของวงจรกรองปรับตัวที่เหมาะสมได้แก่ 2400-4000 ตัวอย่าง ณ ความถี่ชักตัวอย่าง 8 kHz เป็นต้น จะเห็นได้ว่าแม้เป็นห้องใกล้ที่มีขนาดเล็ก (ห้องโดยสารรถยนต์) จำนวนสัมประสิทธิ์ของวงจรกรองแบบปรับตัวที่เหมาะสมยังคงมีค่าสูงอยู่ ดังนั้นขั้นตอนวิธีที่ใช้ในการปรับตัวของวงจรกรองปรับตัวจึงต้องมีการคำนึงถึงความซับซ้อนในการคำนวณซึ่งแปรผันตามจำนวนสัมประสิทธิ์ดังกล่าวอีกด้วย ขั้นตอนวิธีการปรับตัวของวงจรกรองแบบปรับตัวที่ใช้ในการหักล้างเสียงสะท้อนถูกพัฒนาขึ้นจำนวนมาก เช่น NLMS [3], FRLS [34], APA [35] และ PCP [37] เป็นต้น

AES ถูกนำเสนอโดยอาศัยหลักการของเทคนิค SS ซึ่งถูกใช้ใน NS โดยความซับซ้อนในการคำนวณของ AES อยู่ในระดับที่ต่ำมากเมื่อเทียบกับ AEC อย่างไรก็ตามความผิดเพี้ยนของเสียงพูดของคู่สนทนาทางห้องใกล้ซึ่งเป็นสัญญาณขาออกของ AES จะมีค่าสูงมากกว่าเสียงพูดทางห้องใกล้ที่ได้รับการปรับปรุงจาก AEC โดยเฉพาะอย่างยิ่งในสถานการณ์ที่คู่สนทนาทั้งคู่พูดพร้อมๆ กัน ซึ่งเรียกสถานการณ์ดังกล่าวนี้ว่า สถานการณ์ดับเบิลทอล์ก (Double talk situation, DTS)

ในสถานการณ์จริงของการสื่อสารแบบแฮนด์ฟรี ห้องใกล้ที่ทำการพิจารณามักประสบกับปัญหาเสียงก่อกวนทั้งสองชนิดพร้อมๆ กัน เช่น นาย ก. อาศัยระบบสื่อสารแบบแฮนด์ฟรีเพื่อสนทนากับ นาย ข. ในระหว่างขับรถ (เพื่อความปลอดภัยในการขับขี่) สถานการณ์ดังกล่าวนี้ ทั้งเสียงสะท้อนที่เกิดขึ้นจากการวางตัวของอุปกรณ์สื่อสารปลายทางแบบแฮนด์ฟรี และเสียงเครื่องยนต์ที่เล็ดรอดเข้ามายังห้องโดยสารล้วนแล้วแต่เป็นเสียงก่อกวนที่ต้องการลดระดับลงทั้งสิ้น เป็นต้น ดังนั้น NR และ AER จึงสมควรได้รับการออกแบบให้ทำงานได้พร้อมๆ กัน หรืออย่างน้อยทำงานได้อย่างสอดคล้องกันเพื่อลดเสียงก่อกวนทั้งสองในระบบการสื่อสารแบบแฮนด์ฟรี ความพยายามครั้งแรกในการออกแบบการทำงานร่วมกันระหว่าง NR และ AER ดังกล่าวเกิดขึ้นเมื่อประมาณ ค.ศ. 1994 โดยเทคนิค NR ที่ถูกเลือกมาพิจารณาได้แก่ NS ส่วนเทคนิค AER ที่ถูกเลือกมาพิจารณาได้แก่ AEC โดยเรียกว่าเป็นการพัฒนาระบบร่วม (Combined System) โดยช่วงเริ่มแรกของการพัฒนาเป็นการพยายามหาลำดับการทำงานก่อนหลังที่เหมาะสมระหว่าง NS และ AEC [48] หลังจากนั้นการวิเคราะห์นำมาซึ่งสมการที่แสดงให้เห็นว่าวิธีการ AEC ควรเป็นวิธีการที่เกิดขึ้นก่อนวิธีการลดเสียงรบกวนและเสียงสะท้อนที่ยังคงหลงเหลืออยู่พร้อมๆ กัน [49] นอกจากนี้ยังมีผู้เสนอให้ใช้เทคนิคการกดสเปกตรัมในการแก้ปัญหาทั้งเสียงสะท้อนและเสียงรบกวน โดยเรียกวิธีการดังกล่าวว่า การกดเสียงสะท้อนและเสียงรบกวน (Acoustic Echo and Noise Suppression, AENS) [53] ซึ่งวิธี AENS ช่วยลดความซับซ้อนในการคำนวณของระบบร่วมลงได้อย่างมาก แต่อย่างไรก็ตามผลการทดลองของผู้พัฒนา AENS ชี้ให้เห็นว่า AENS นำมาซึ่งความผิดเพี้ยนของเสียงพูดของคู่สนทนาทางห้องใกล้ที่สูงขึ้นกว่าระบบร่วมระหว่าง AEC และ NS โดยเฉพาะอย่างยิ่งในกรณี DTS

วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการเพิ่มประสิทธิภาพให้กับวิธีการลดเสียงก่อกวนทั้งสองประเภทอันได้แก่ เสียงสะท้อน และเสียงรบกวนพื้นหลัง (ซึ่งถูกสมมติให้เป็นเสียงรบกวนจุดนิ่ง) โดยพัฒนาจากเทคนิคการกดเสียงสะท้อนและเสียงรบวน AENS ใน [53] เพื่อให้สามารถลดทอนเสียงสะท้อนและเสียงรบกวนลงได้ดียิ่งขึ้น รวมทั้ง

ลดผลของความผิดเพี้ยนของเสียงพูดทางห้องใกล้ลง ในขณะที่ยังคงรักษาความซับซ้อนในการคำนวณที่ต่ำกว่าระบบร่วมระหว่าง AEC และ NS เอาไว้

1.2. วัตถุประสงค์

เพื่อพัฒนาการเพิ่มสมรรถนะเสียงพูดในระบบการสื่อสารทางเสียงแบบแฮนด์ฟรี โดยทำการลดเสียงก่อกวนสองชนิดได้แก่ เสียงรบกวนพื้นหลังและเสียงสะท้อน

1.3. ขอบเขตของวิทยานิพนธ์

1. พิจารณาระบบ AENS ทางฝั่งห้องใกล้ของระบบการสื่อสารทางเสียงแบบแฮนด์ฟรีที่มีไมโครโฟนและลำโพงอย่างละหนึ่งตัว ในสถานการณ์ซิงเกิลทอล์ก และดับเบิลทอล์ก
2. พัฒนาประสิทธิภาพระบบ AENS ในด้าน การลดลงของเสียงสะท้อนและเสียงรบกวน รวมทั้งความผิดเพี้ยนของสัญญาณเสียงขาออกของระบบ
3. พิจารณากรณีของสัญญาณเสียงรบกวนแบบบวกเพิ่ม (Additive Noise) และมีความเป็นจุดนิ่ง (Stationary) เท่านั้น
4. พัฒนาระบบ AENS ที่ให้ความผิดเพี้ยนของสัญญาณเสียงพูดที่ต่ำ และยังคงมีความซับซ้อนในการคำนวณที่ต่ำเช่นกัน

1.4. ประโยชน์ที่คาดว่าจะได้รับ

แนวทางในการแก้ไขปัญหาเสียงสะท้อนและเสียงรบกวนในระบบการสื่อสารทางเสียงแบบแฮนด์ฟรี ที่มีประสิทธิภาพสูงโดยรักษาไว้ซึ่งคุณภาพเสียงพูดของสัญญาณเสียงขาออกของระบบ

1.5. ขั้นตอนและวิธีดำเนินการ

1. ศึกษาปัญหาเสียงสะท้อนและเสียงรบกวนที่เกิดขึ้นในระบบการสื่อสารแบบแฮนด์ฟรี
2. ศึกษาวิธีการลดสัญญาณเสียงรบกวนที่อาศัยเทคนิคการกดทางสเปกตรัม
3. ศึกษาวิธีการลดสัญญาณเสียงสะท้อน
4. ศึกษาระบบการลดเสียงสะท้อนและเสียงรบกวนในระบบการสื่อสารทางเสียงแบบแฮนด์ฟรี
5. พัฒนาวิธีการลดเสียงสะท้อนและเสียงรบกวนในระบบการสื่อสารแบบแฮนด์ฟรี
6. จำลองวิธีการที่พัฒนาขึ้น เพื่อทดสอบผลการลดเสียงสะท้อนและเสียงรบกวน
7. วิเคราะห์ และสรุปผลการวิจัย
8. เขียนวิทยานิพนธ์

บทที่ 2

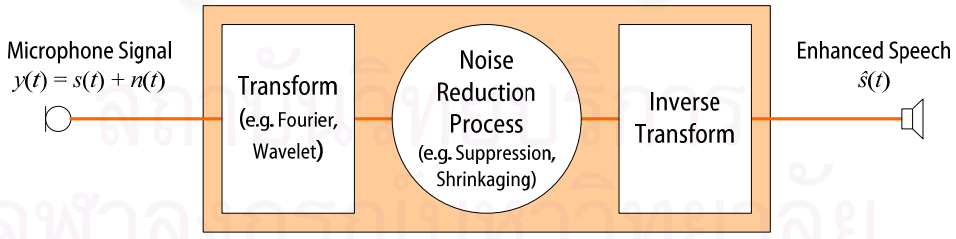
หลักการและขั้นตอนวิธีสำคัญ

การเพิ่มประสิทธิภาพให้กับการสื่อสารแบบแฮนด์ฟรีด้วยการเพิ่มสมรรถนะเสียงพูด (Speech Enhancement) สามารถทำได้โดยการแก้ปัญหาเสียงก่อกวน 2 ชนิด ได้แก่ เสียงรบกวนพื้นหลัง ซึ่งวิทยานิพนธ์ฉบับนี้สมมติให้เป็นเสียงรบกวนจุดนิ่ง (โดยต่อไปจะใช้เฉพาะคำว่า “เสียงรบกวน” โดยละคำว่าพื้นหลังไว้เพื่อความกระชับในการเขียนวิทยานิพนธ์) และเสียงสะท้อน ถึงแม้ว่าวิทยานิพนธ์ฉบับนี้จะอาศัยเทคนิค SS ในการแก้ปัญหาเสียงก่อกวนทั้งสอง แต่ NR และ AER ที่อาศัยเทคนิคอื่นๆ ก็จะถูกกล่าวถึงอย่างสังเขปด้วย เนื้อหาในบทที่ 2 ประกอบไปด้วย การลดเสียงรบกวน การลดเสียงสะท้อน และระบบร่วมการลดเสียงสะท้อนและเสียงรบกวน

2.1. การลดเสียงรบกวน

การลดเสียงรบกวนที่อาศัยไมโครโฟนหนึ่งตัว สามารถแบ่งออกได้เป็น 3 กลุ่ม ได้แก่ การลดเสียงรบกวนที่อาศัยการแปลง การลดเสียงรบกวนที่อาศัยวงจรกรองแบบปรับตัว และการลดเสียงรบกวนที่อาศัยแบบจำลองของเสียงพูด

การลดเสียงรบกวนที่อาศัยการแปลง มีหลักการร่วมกันคือ ทำการแปลงสัญญาณเสียงที่ถูกรบกวนที่รับมาได้ไปสู่โดเมนอื่นๆ ผ่านทางการแปลง (Transform) ต่างๆ และดำเนินการจัดการกับสัญญาณเสียงรบกวนบนโดเมนที่ทำการแปลงไปนั้นๆ ก่อนที่จะทำการแปลงกลับ (Inverse transform) สัญญาณที่ได้รับการปรับปรุงคืนสู่โดเมนเวลาต่อไปดังรูปที่ 2.1 ทั้งนี้การแปลงที่เลือกใช้ในการลดเสียงรบกวนจะเป็นการแปลงที่ทำให้แต่ละองค์ประกอบในโดเมนที่ทำการแปลงไปนั้นๆ มีความอิสระจากกันค่อนข้างมาก จึงสามารถพิจารณาในแต่ละองค์ประกอบได้อย่างอิสระจากกัน ส่งผลให้กระบวนการลดเสียงรบกวนอยู่ในรูปแบบที่ง่ายลงกว่าการลดเสียงรบกวนดังกล่าวในโดเมนเวลามาก

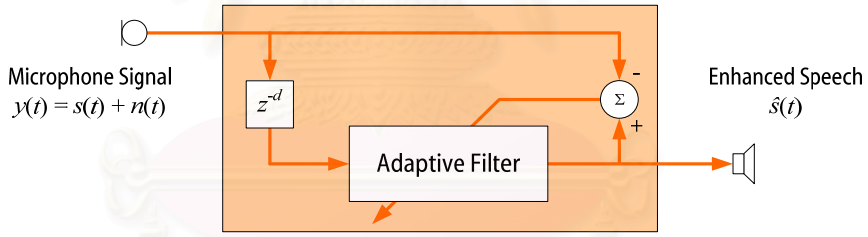


รูปที่ 2.1 การลดเสียงรบกวนที่อาศัยการแปลง

ตัวอย่างการแปลงที่มีคุณสมบัติดังกล่าวได้แก่ การแปลงฟูริเยร์ (Fourier transform) การแปลงเวฟเล็ต (Wavelet transform) ซึ่งเป็นการแปลงที่ไม่ขึ้นกับสัญญาณ (Signal independent transformation) และการแปลงคาฮูเนน-เลิฟ (Karhunen-Loève transform) ซึ่งเป็นการแปลงที่ขึ้นกับสัญญาณ (Signal dependent transformation) เป็นต้น กระบวนการจัดการกับเสียงรบกวนถูกพิจารณาโดยขึ้นกับการแปลงแบบต่างๆ ที่เลือกใช้ เช่น การกด (Suppression) ถูกใช้เป็นกระบวนการลดเสียงรบกวนในโดเมนความถี่ ซึ่งได้มาจากการแปลงฟูริเยร์ในช่วงเวลา

สั้นๆ (Short-Time Fourier Transform, STFT) หรือก็คือ เทคนิค SS นั่นเอง การหด (Shrinkage) และการตัด (Thresholding) ถูกใช้กับการลดเสียงรบกวนในโดเมนของการแปลงเวฟเล็ต ซึ่งโดยทั่วไปถูกเรียกว่า การลดเสียงรบกวนในโดเมนเวฟเล็ต (Wavelets Denoising) [26]-[27] และการกดที่ถูกใช้เป็นกระบวนการลดเสียงรบกวนในโดเมนของการแปลงคาฮูเนน-เลิฟ ซึ่งถูกเรียกว่า การลดเสียงรบกวนในปริภูมิย่อยสัญญาณ (Signal Subspace Method) [28] เป็นต้น จะเห็นว่า NS ที่ถูกวิคราะห์และพัฒนาในวิทยานิพนธ์ฉบับนี้ก็อยู่ในกลุ่มของการลดเสียงรบกวนที่อาศัยการแปลงนี้เอง

การลดเสียงรบกวนที่อาศัยวงจรกรองปรับตัว ใช้หลักการทำงานของวงจรกรองปรับตัว ที่สามารถประมาณสัญญาณที่มีลักษณะเป็นรายคาบได้จากสัญญาณรายคาบนั้นๆ ในอดีต ทั้งนี้เนื่องจากเสียงพูดเป็นสัญญาณรายคาบหลายความถี่รวมเข้าด้วยกัน ดังนั้นวงจรกรองปรับตัวจึงสามารถทำนายสัญญาณเสียงพูดได้จากสัญญาณเสียงพูดในอดีตด้วยเช่นกัน ในขณะเดียวกันเสียงรบกวนแถบความถี่กว้าง (Wideband noise) มีความเป็นรายคาบต่ำ ดังนั้นวงจรกรองปรับตัวจึงไม่สามารถประมาณค่าสัญญาณเสียงรบกวนดังกล่าวได้จากสัญญาณเสียงรบกวนในอดีต คุณสมบัติที่กล่าวมานี้มีบทบาทสำคัญอย่างมากในการช่วยลดเสียงรบกวน กล่าวคือ หากอาศัยค่าในอดีตของสัญญาณเสียงพูดที่ถูกรบกวนเป็นสัญญาณขาเข้าของวงจรกรองปรับตัวแล้ว วงจรกรองปรับตัวก็จะสามารถประมาณค่าได้เฉพาะสัญญาณเสียงพูดที่ปราศจากเสียงรบกวนเท่านั้น นำมาซึ่งวิธีการลดเสียงรบกวนดังกล่าวได้ อย่างไรก็ตามสัญญาณเสียงพูดที่ประมาณได้จากวงจรกรองปรับตัวมีความผิดเพี้ยน (Distortion) ไปจากเสียงพูดสะอาด (Clean speech) มาก จึงมีการพัฒนาการประมาณสัญญาณเสียงรบกวนขึ้น เพื่อนำไปหักลบออกจากสัญญาณเสียงพูดที่ถูกรบกวนแทน ตัวอย่างงานวิจัยในกลุ่มนี้ได้แก่ [29]-[30]



รูปที่ 2.2 ลดเสียงรบกวนที่อาศัยวงจรกรองปรับตัว

การลดเสียงรบกวนที่อาศัยแบบจำลองของเสียงพูด เนื่องจากเสียงพูดสามารถถูกจำลองเป็นระบบพลวัตสถานะ (State dynamic system) ได้ ดังนั้นค่าประมาณสัญญาณเสียงพูดที่ทำให้ค่าผิดพลาดเฉลี่ยกำลังสองระหว่างสัญญาณเสียงพูดและค่าประมาณสัญญาณเสียงพูดมีค่าน้อยที่สุดเมื่อให้มาเฉพาะสัญญาณเสียงพูดที่ถูกรบกวนและระบบสถานะพลวัตของเสียงพูดนั้นๆ สามารถหาได้โดยอาศัยการทำงานของ Kalman filter [58] อย่างไรก็ตาม เนื่องจากระบบสถานะพลวัตของสัญญาณเสียงพูดดังกล่าวเป็นแบบจำลองซึ่งไม่ทราบค่า การประมาณระบบสถานะพลวัตดังกล่าวจึงต้องดำเนินการควบคู่ไปกับการประมาณสัญญาณเสียงพูด โดยวิธีการส่วนใหญ่อาศัยการประมาณแบบกลับไปกลับมาที่เรียกว่า Expectation-Maximization หรือ EM algorithm [57] ผลของเสียงพูดที่ได้จากการปรับปรุงโดยวิธีการในกลุ่มนี้มีความเป็นธรรมชาติมากกว่าวิธีการในกลุ่มอื่นๆ [32] แต่อย่างไรก็ตามความซับซ้อนในการทำงานของวิธีการประเภทนี้มากกว่าทั้งสองกลุ่มข้างต้นมาก และยังไม่สามารถลดผลของเสียงรบกวนลงได้มากนัก ตัวอย่างของงานวิจัยทางด้านนี้ได้แก่ [31]-[32]

วิทยานิพนธ์ฉบับนี้มุ่งเน้นลงไปที่การลดเสียงรบกวนที่อาศัยเทคนิคการกดทางสเปกตรัม (Noise Reduction based on Spectral Suppression Technique) หรือ NS หลักการ และวิธีการทำงานโดยละเอียดของ NS จะถูกบรรยายอย่างละเอียดในหัวข้อย่อยที่ต่อไปนี้

2.1.1. การกดเสียงรบกวน (Noise Suppression)

กำหนดให้สัญญาณเสียงพูดที่ถูกรบกวน หรือสัญญาณไมโครโฟน $y(t)$ เกิดจากผลรวมของสัญญาณเสียงพูด $s(t)$ และสัญญาณเสียงรบกวนแบบบวก (Additive noise) $n(t)$ ที่ไม่มีสหสัมพันธ์ต่อกัน เมื่อ t คือดัชนีที่บ่งบอกเวลาแบบไม่ต่อเนื่อง (Discrete time index) ดังสมการต่อไปนี้

$$y(t) = s(t) + n(t) \tag{2.1}$$

สัญญาณไมโครโฟนจะถูกแปลงไปสู่โดเมนความถี่ผ่านทาง การแปลงฟูริเยร์แบบไม่ต่อเนื่องในช่วงเวลาสั้นๆ (Short-Time Discrete Fourier Transform, STFT) ดังนี้

$$Y(k,\ell) = \sum_{t=0}^{T-1} y(t+\ell M)\, h(t)\, e^{-j\frac{2\pi}{T}kt} \tag{2.2}$$

เมื่อ $k = 0,1,\ldots,T-1$ คือ ดัชนีบ่งบอกองค์ประกอบทางความถี่ (Frequency-bin Index) $\ell = 1,2,\ldots$ คือ ดัชนีบ่งบอกเฟรมเวลา (Time frame Index) M คือ ค่าช่วงก้าวระหว่างเฟรม (Frame step) กล่าวคือจำนวนตัวอย่างที่ถูกก้าวข้ามไปก่อนทำการแปลงในแต่ละเฟรม T คือ จำนวนตัวอย่างที่ใช้ในการแปลง STFT แต่ละเฟรมซึ่งบ่งบอกถึงความละเอียดของแต่ละองค์ประกอบทางความถี่ (Frequency Resolution) และ $h(t)$ คือ หน้าต่างการวิเคราะห์ (Analysis window) ซึ่งใช้ลดผลของการรั่วไหลระหว่างองค์ประกอบทางความถี่ (Frequency leakage) ในการวิเคราะห์เชิงความถี่ (Frequency analysis) การแปลงฟูริเยร์แบบไม่ต่อเนื่องในช่วงเวลาสั้นๆ สำหรับการวิเคราะห์เชิงความถี่ถูกแสดงดังรูปที่ 2.3

หลังจากทำการแปลงไปสู่โดเมนความถี่แล้ว สมการที่ (2.1) สามารถเขียนได้ใหม่เป็น

$$Y(k,\ell) = S(k,\ell) + N(k,\ell) \tag{2.3}$$

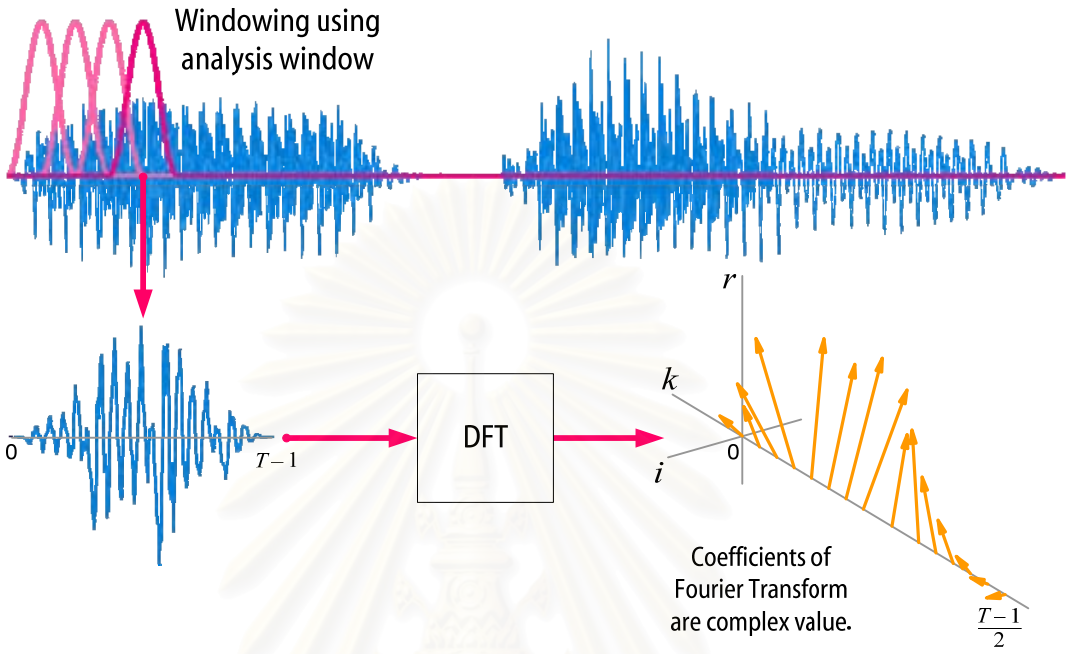
โดยสัมประสิทธิ์ $Y(k,\ell)$ ถูกเรียกว่า สปกตรัมเสียงพูดที่ถูกรบกวน หรือ สเปกตรัมสัญญาณไมโครโฟน $S(k,\ell)$ คือ สเปกตรัมเสียงพูด และ $N(k,\ell)$ คือ สเปกตรัมเสียงรบกวน ณ องค์ประกอบทางความถี่ที่ k และเฟรมเวลาที่ ℓ โดยสัมประสิทธิ์ทั้งสามนี้ เป็นจำนวนเชิงซ้อนซึ่งสามารถเขียนในรูปของขนาดและเฟสได้ดังนี้

$$Y(k,\ell) = R(k,\ell)\, e^{j\theta_Y(k,\ell)} \tag{2.4}$$

$$S(k,\ell) = A(k,\ell)\, e^{j\theta_S(k,\ell)} \tag{2.5}$$

$$N(k,\ell) = B(k,\ell)\, e^{j\theta_N(k,\ell)} \tag{2.6}$$

เมื่อ R , A , B , $\theta_Y(k,\ell)$, $\theta_S(k,\ell)$ และ $\theta_N(k,\ell)$ คือขนาดสเปกตรัมเสียงพูดที่ถูกรบกวน ขนาดสเปกตรัมเสียงพูด ขนาดสเปกตรัมเสียงรบกวน เฟสสเปกตรัมเสียงพูดที่ถูกรบกวน เฟสสเปกตรัมเสียงพูด และ เฟสสเปกตรัมเสียงรบกวน ตามลำดับ ณ องค์ประกอบทางความถี่ที่ k และเฟรมเวลาที่ ℓ



รูปที่ 2.3 การแปลงฟูริเยร์แบบไม่ต่อเนื่องในช่วงเวลาสั้นๆ สำหรับการวิเคราะห์เชิงความถี่

จุดประสงค์ของการกดเสียงรบกวนคือต้องการหาค่าประมาณสเปกตรัมเสียงพูด (Spectral speech estimate) $\hat{S}(k,\ell)$ ซึ่งปราศจากส่วนของเสียงรบกวน จากสเปกตรัมสัญญาณไมโครโฟน $Y(k,\ell)$ โดยวิธีการประมาณค่าดังกล่าวจะถูกบรรยายอย่างละเอียดในหัวข้อย่อยที่ 2.1.1.1-2.1.1.4 จากนั้น ค่าประมาณสเปกตรัมเสียงพูดจะถูกแปลงกลับสู่โดเมนเวลาโดยผ่านทางการแปลงฟูริเยร์แบบไม่ต่อเนื่องผกผันในช่วงเวลาสั้นๆ และวิธีการ Overlap add เพื่อให้ได้มาซึ่งสัญญาณเสียงพูดที่ถูกปรับปรุง $\hat{s}(t)$ ในทางเวลาต่อไป ดังนี้

$$\hat{s}(t) = \sum_{\ell}\sum_{k=0}^{T-1} \hat{S}(k,\ell)\,\tilde{h}(t-\ell M)\,e^{j\frac{2\pi}{T}k(t-\ell M)} \qquad (2.7)$$

เมื่อ $\tilde{h}(t)$ คือ หน้าต่างสังเคราะห์ (Synthesis window) ซึ่งเป็น Biorthogonal กับ $h(t)$

2.1.1.1. การหา ค่าประมาณสเปกตรัมเสียงพูด

เนื่องจากข้อมูลที่รับมาได้มีเพียงสเปกตรัมเสียงพูดที่ถูกรบกวน $Y(\cdot,\ell)$ [3] เท่านั้น การประมาณหาค่าสเปกตรัมเสียงพูด $S(k,\ell)$ จากข้อมูลที่มีอยู่นี้จึงสามารถทำได้โดย การพยายามหาค่า $\hat{S}(k,\ell)$ ที่สอดคล้องกับเงื่อนไขต่อไปนี้

$$\hat{S}(k,\ell) = \arg\min_{\hat{S}(k,\ell)} E\{\, d[S(k,\ell),\hat{S}(k,\ell)] \,|\, Y(\cdot,\ell)\} \tag{2.8}$$

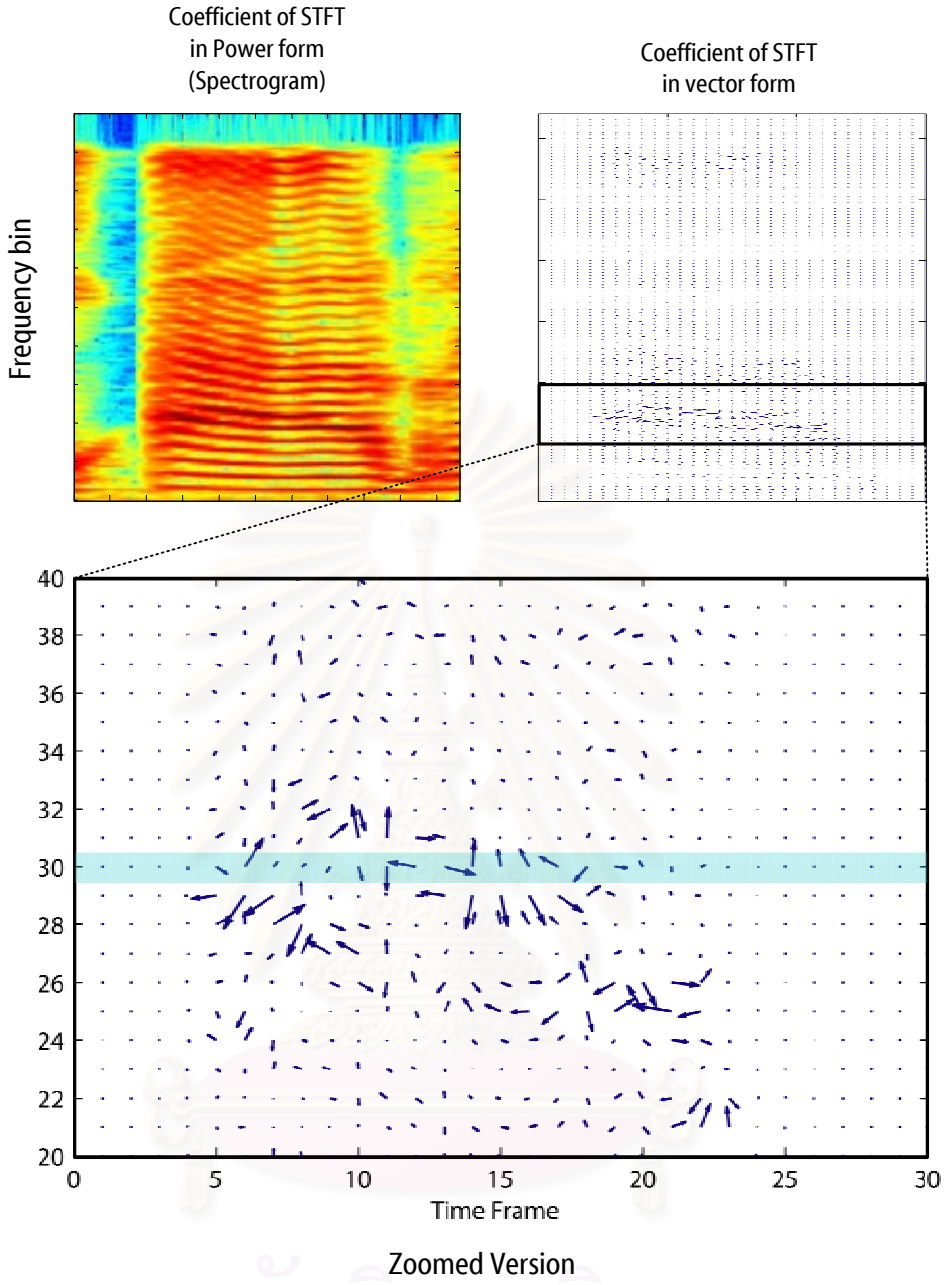
เมื่อ $E\{\cdot\}$ คือค่าคาดหวัง (Expectation) และ $d[S(k,\ell),\hat{S}(k,\ell)]$ คือค่าผิดเพี้ยน (Distortion measure) ระหว่างสเปกตรัมเสียงพูด และค่าประมาณสเปกตรัมเสียงพูด [12] และเนื่องจากสมมติฐานที่ว่าแต่ละองค์ประกอบทางความถี่มีความอิสระจากกัน [7] ทำให้สมการที่ (2.8) สามารถเขียนได้เป็น

$$\hat{S}(k,\ell) = \arg\min_{\hat{S}(k,\ell)} E\{\, d[S(k,\ell),\hat{S}(k,\ell)] \,|\, Y(k,\ell)\} \tag{2.9}$$

จึงทำให้สามารถทำการประมาณค่าสเปกตรัมเสียงพูดในแต่ละองค์ประกอบทางความถี่อย่างแยกจากกันได้

จากสมการที่ (2.9) จะเห็นว่านอกจาก ค่าผิดเพี้ยน $d[S(k,\ell),\hat{S}(k,\ell)]$ ที่สามารถถูกเลือกอย่างเหมาะสมแล้ว การแจกแจงความน่าจะเป็นของสเปกตรัมเสียงพูด และของสเปกตรัมเสียงรบกวน ก็เป็นสิ่งซึ่งสามารถถูกเลือกให้เหมาะสมได้เช่นเดียวกัน ปัจจุบันยังไม่เป็นที่แน่ชัดว่า ค่าผิดเพี้ยน และการแจกแจงความน่าจะเป็นชนิดใด ที่มีความเหมาะสมต่อสเปกตรัมเสียงพูดมากที่สุด [13]

[3] $Y(\cdot,\ell)$ คือสเปกตรัมสัญญาณไมโครโฟนทุกๆ องค์ประกอบทางความถี่ ณ เฟรมเวลาที่ ℓ



รูปที่ 2.4 ภาพลักษณะของสเปกตรัมเสียงพูด (ซึ่งอาศัยเวกเตอร์แทนจำนวนเชิงซ้อน) ในโดเมนความถี่ จะเห็นว่าหากดูตามแกนเฟรมเวลา (แถบที่แรเงา) แล้วจะเห็นถึงความเป็นกระบวนการสโตแคสติกเชิงซ้อนของสเปกตรัมเสียงพูด (Complex Stochastic Process)

นักวิจัยหลายกลุ่มพยายามหาและเสนอการแจกแจงความน่าจะเป็นของสเปกตรัมเสียงพูดที่เหมาะสม เช่น การกระจายตัวแบบลาปลาส [14] การกระจายตัวแบบเอกซ์โพเนนเชียล [14] และการกระจายตัวที่ได้มาจากการเก็บข้อมูล [16] เป็นต้น แต่อย่างไรก็ตาม [13] แสดงให้เห็นว่า การแจกแจงความน่าจะเป็นแบบเกาส์เซียนสามารถพัฒนาไปสู่การแจกแจงแบบอื่นๆ ได้เช่นกัน หากคิดคำนึงถึงความน่าจะเป็นของค่าความแปรปรวนของสเปกตรัมเสียงพูดด้วย

<u>ตัวอย่างการหาค่าประมาณสเปกตรัมเสียงพูดโดยเลือกค่าผิดเพี้ยน เป็นแบบค่าผิดพลาดกำลังสอง และเลือกการแจกแจงความน่าจะเป็นของสเปกตรัมเสียงพูดแบบเกาส์เซียน</u>

ให้สเปกตรัมเสียงพูดมีการแจกแจงความน่าจะเป็นแบบเกาส์เซียน (Gaussian distribution) ทั้งส่วนจริงและส่วนจินตภาพดังนี้

$$p_{S_R(k,\ell)}(X) = \frac{1}{\sqrt{2\pi\lambda_{S_R}(k,\ell)}} \exp\left(-\frac{|X|^2}{2\lambda_{S_R}(k,\ell)} \right) \tag{2.10}$$

$$p_{S_I(k,\ell)}(X) = \frac{1}{\sqrt{2\pi\lambda_{S_I}(k,\ell)}} \exp\left(-\frac{|X|^2}{2\lambda_{S_I}(k,\ell)} \right) \tag{2.11}$$

เมื่อ ตัวห้อย R แทนส่วนจริง ตัวห้อย I แทนส่วนจินตภาพ (โดย $S = S_R + jS_I$) $\lambda_{S_R}(k,\ell)$ คือความแปรปรวนของส่วนจริงสเปกตรัมเสียงพูด และ $\lambda_{S_I}(k,\ell)$ คือ ความแปรปรวนของส่วนจินตภาพสเปกตรัมเสียงพูด จากสมการที่ (2.10) และ (2.11) สามารถหาการแจกแจงความน่าจะเป็นของสเปกตรัมเสียงพูดได้เป็นดังนี้

$$p_{S(k,\ell)}(X) = \frac{1}{\pi\lambda_S(k,\ell)} \exp\left(-\frac{|X|^2}{\lambda_S(k,\ell)} \right) \tag{2.12}$$

โดยที่

$$\lambda_S(k,\ell) = 2\lambda_{S_R}(k,\ell) = 2\lambda_{S_I}(k,\ell) = E\{|S(k,\ell)|^2\} \tag{2.13}$$

คือ ค่าความแปรปรวนของสเปกตรัมเสียงพูด หรือ ค่าความหนาแน่นสเปกตรัมกำลังเสียงพูด (Speech Power Spectral Density, SPSD)

เช่นเดียวกันหากทำการสมมติให้สเปกตรัมเสียงรบกวนมีการกระแจกแจงความน่าจะเป็นแบบเกาส์เซียนแล้วจะได้ว่า

$$p_{N(k,\ell)}(X) = \frac{1}{\pi\lambda_N(k,\ell)} \exp\left(-\frac{|X|^2}{\lambda_N(k,\ell)} \right) \tag{2.14}$$

เมื่อ

$$\lambda_N(k,\ell) = E\{|N(k,\ell)|^2\} \tag{2.15}$$

คือ ค่าความแปรปรวนของสเปกตรัมเสียงรบกวน หรือ ค่าความหนาแน่นสเปกตรัมกำลังเสียงรบกวน (Noise Power Spectral Density, NPSD)

โดยเลือกค่าผิดเพี้ยน $d[S(k,\ell), \hat{S}(k,\ell)]$ เป็นแบบค่าผิดพลาดกำลังสอง (Square error distortion) ดังนี้

$$d_{\mathrm{SE}}[S(k,\ell), \hat{S}(k,\ell)] = |S(k,\ell) - \hat{S}(k,\ell)|^2 \tag{2.16}$$

ผลเฉลยของสมการที่ (2.9) จะตรงกับตัวประมาณเบย์เซียน (Bayesian estimator) ซึ่งได้แก่ ค่าเฉลี่ยแบบมีเงื่อนไข (Conditional mean) ของสเปกตรัมเสียงพูดดังนี้

$$\hat{S}(k,\ell) = E\{S(k,\ell) \mid Y(k,\ell)\} \tag{2.17}$$

ซึ่งสามารถหาได้โดยอาศัย Bayes' rule ดังนี้

$$E\{S(k,\ell) \mid Y(k,\ell)\} = \frac{\int_0^{\infty}\int_0^{2\pi} S(k,\ell)\, p_{Y(k,\ell)|S(k,\ell)}(Y(k,\ell),S(k,\ell))\, p_{S(k,\ell)}(S(k,\ell))\, dA\, d\theta_S}{\int_0^{\infty}\int_0^{2\pi} p_{Y(k,\ell)|S(k,\ell)}(Y(k,\ell),S(k,\ell))\, p_{S(k,\ell)}(S(k,\ell))\, dA\, d\theta_S} \tag{2.18}$$

โดยจากสมการที่ (2.3) และ (2.14) สามารถหาการแจกแจงความน่าจะเป็น $p_{Y(k,\ell)|S(k,\ell)}$ ได้ดังนี้

$$p_{Y(k,\ell)|S(k,\ell)}(Y(k,\ell),S(k,\ell)) = \frac{1}{\pi\lambda_N(k,\ell)} \exp\left(-\frac{|Y(k,\ell)-S(k,\ell)|^2}{\lambda_N(k,\ell)}\right) \tag{2.19}$$

และโดยอาศัยสมการที่ (2.12), (2.17), (2.18) และ (2.19) จะได้ว่า

$$\hat{S}(k,\ell) = G_{\mathrm{SE}}(k,\ell) Y(k,\ell) \tag{2.20}$$

เมื่อ

$$G_{\mathrm{SE}}(k,\ell) = \frac{\xi(k,\ell)}{\xi(k,\ell)+1} \tag{2.21}$$

ถูกเรียกว่า Spectral Gain ซึ่งมีชื่อเรียกพิเศษสำหรับกรณี ที่ใช้ค่าผิดเพี้ยนเป็นค่าผิดพลาดกำลังสอง นี้ว่า Wiener Gain และ

$$\xi(k,\ell) = \frac{\lambda_S(k,\ell)}{\lambda_N(k,\ell)} \tag{2.22}$$

คือ a priori SNR [6] ซึ่งถือเป็นตัวแปรที่สำคัญเป็นอย่างมากในเทคนิคการกดทางสเปกตรัมดังจะกล่าวถึงต่อไป

นอกจากค่าผิดเพี้ยน $d[S(k,\ell),\hat{S}(k,\ell)]$ แบบค่าผิดพลาดกำลังสอง ดังแสดงในตัวอย่างด้านบนแล้ว ยังมีการนิยามค่าผิดเพี้ยนชนิดอื่นๆ อีกจำนวนมาก เช่น

1. ค่าผิดเพี้ยนแบบ สเปกตรัมแอมปลิจูด (Spectral amplitude distortion, SA) [7]

$$d_{\mathrm{SA}}[S(k,\ell),\hat{S}(k,\ell)] = |A(k,\ell) - \hat{A}(k,\ell)|^2 \tag{2.23}$$

2. ค่าผิดเพี้ยนแบบ ลอการิทึม-สเปกตรัมแอมปลิจูด (Log-spectral amplitude distortion, LSA) [8]

$$d_{\mathrm{LSA}}[S(k,\ell),\hat{S}(k,\ell)] = |\log A(k,\ell) - \log \hat{A}(k,\ell)|^2 \tag{2.24}$$

3. ค่าผิดเพี้ยนแบบ สเปกตรัมกำลัง (Spectral power distortion, SP) [10]

$$d_{\mathrm{SP}}[S(k,\ell),\hat{S}(k,\ell)] = |\, A^2(k,\ell) - \hat{A}^2(k,\ell)\,|^2 \tag{2.25}$$

เป็นต้น

ค่าผิดเพี้ยนที่แตกต่างกันนำมาซึ่งรูปแบบของ Spectral gain ที่แตกต่างกันตามไปด้วย โดย Spectral gain ที่ตรงกับ ค่าผิดเพี้ยนที่ถูกยกตัวอย่างไว้ด้านบน ได้แก่ G_{SA} , G_{LSA} และ G_{SP} ตามลำดับ โดยมีรูปแบบฟังก์ชันเป็น

$$G_{\mathrm{SA}}(k,\ell) = \frac{\sqrt{\pi v(k,\ell)}}{2\gamma(k,\ell)}\left[(1+v(k,\ell))I_0\left(\frac{v(k,\ell)}{2}\right)+v(k,\ell)I_1\left(\frac{v(k,\ell)}{2}\right)\right]\exp(-\frac{v(k,\ell)}{2}) \tag{2.26}$$

$$G_{\mathrm{LSA}}(k,\ell) = \frac{\xi(k,\ell)}{\xi(k,\ell)+1}\exp\left(\frac{1}{2}\int_{v(k,\ell)}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{2.27}$$

$$G_{\mathrm{SP}}(k,\ell) = \sqrt{\frac{\xi(k,\ell)}{\xi(k,\ell)+1}\left(\frac{1}{\gamma(k,\ell)}+\frac{\xi(k,\ell)}{\xi(k,\ell)+1}\right)} \tag{2.28}$$

ตามลำดับ เมื่อ $I_0(\cdot)$ และ $I_1(\cdot)$ คือ Modified Bessel functions อันดับที่ 0 และอันดับที่ 1 ตามลำดับ และ

$$v(k,\ell) = \gamma(k,\ell)\frac{\xi(k,\ell)}{\xi(k,\ell)+1} \tag{2.29}$$

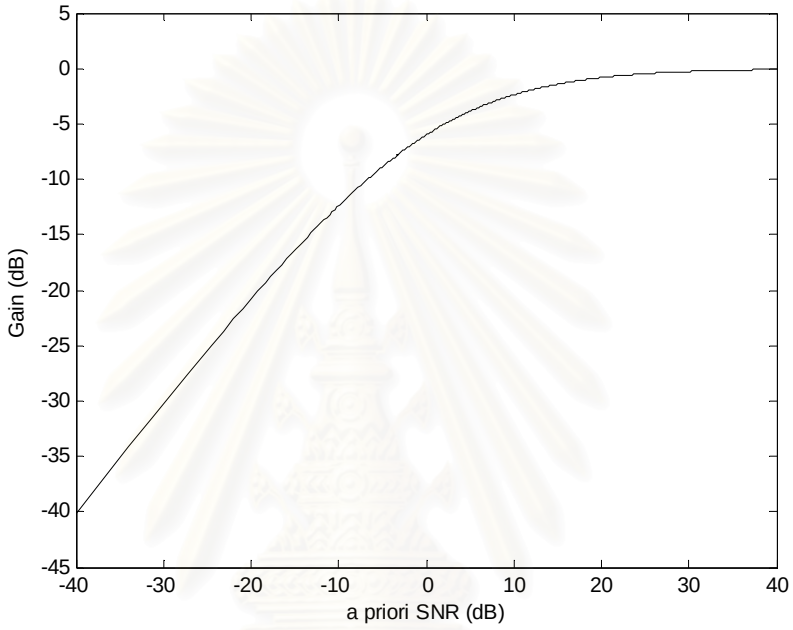
เมื่อ

$$\gamma(k,\ell) = \frac{|\,Y(k,\ell)\,|^2}{\lambda_N(k,\ell)} \tag{2.30}$$

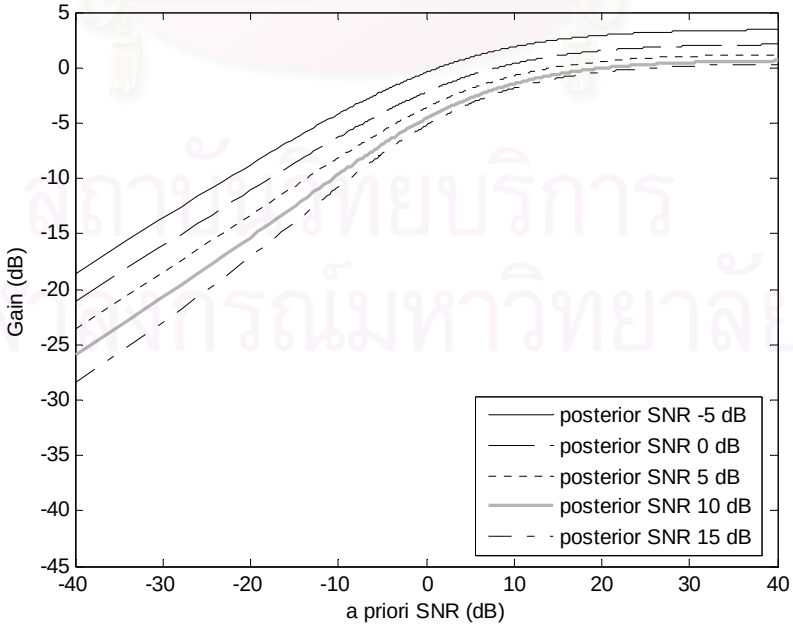
คือ posterior SNR [6]

สังเกตได้ว่า Spectral Gain ที่ได้จากค่าผิดเพี้ยนแบบ SA, LSA และ SP นั้น ล้วนแล้วแต่เป็นฟังก์ชันของทั้ง a priori SNR และ posterior SNR ในขณะที่ Spectral Gain ที่ได้จากค่าผิดเพี้ยน SE เป็นฟังก์ชันของตัวแปร a priori SNR เท่านั้น ทั้งนี้เนื่องจากการเลือกใช้ค่าผิดเพี้ยนที่เกิดจากความแตกต่างของขนาดสเปกตรัมในรูปแบบต่างๆ ดังเช่น สมการที่ (2.23)-(2.25) จะให้ค่าประมาณสเปกตรัมเสียงพูดที่มีการคำนึงถึงความเป็นเวกเตอร์ของปริมาณสเปกตรัมด้วย ในขณะที่การเลือกใช้ค่าผิดเพี้ยนที่เกิดจากความแตกต่างของสเปกตรัม ดังเช่น SE ให้ค่าประมาณสเปกตรัมเสียงพูดที่ไม่มีการคำนึงถึงความเป็นเวกเตอร์ของปริมาณสเปกตรัม การคำนึงถึงความเป็นเวกเตอร์ของปริมาณสเปกตรัมนี้ ทำให้ต้องอาศัยทั้งปริมาณที่สามารถบ่งบอกหรือส่อความ (Imply) ถึงทั้งขนาดและเฟสของปริมาณสเปกตรัมดังกล่าวได้ จากนิยามของ a priori SNR ในสมการที่ (2.22) สามารถกล่าวได้ว่า a priori SNR เป็นปริมาณที่ส่อความถึงเฉพาะขนาดของปริมาณสเปกตรัม ในขณะที่ปริมาณที่เป็นตัวตั้งหารของค่า posterior SNR ในสมการที่ (2.30) คือ ขนาดสเปกตรัมสัญญาณไมโครโฟนนั้น มีการรวมผลของการรวมกันแบบเวกเตอร์ของสเปกตรัมเสียงพูดและสเปกตรัมเสียงรบกวนเอาไว้ ทำให้ posterior SNR ส่อความได้ถึงเฟสที่แตกต่างกัน

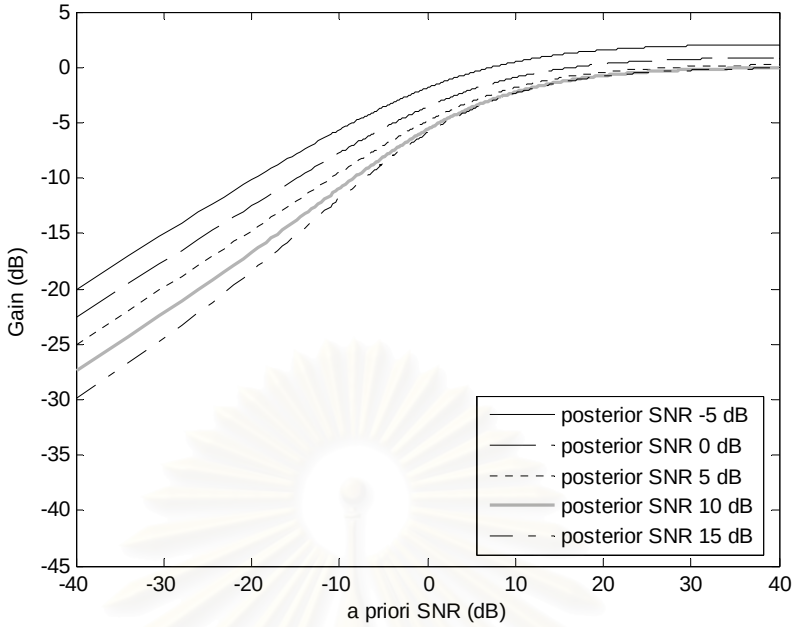
ระหว่างสเปกตรัมเสียงพูดและสเปกตรัมเสียงรบกวนได้ ดังนั้น Spectral Gain ที่ได้จากการเลือกใช้ผิดเพี้ยนที่ให้ค่าประมาณสเปกตรัมเสียงพูดที่มีการคิดคำนึงถึงความเป็นเวกเตอร์ของปริมาณสเปกตรัม จึงประกอบด้วยตัวแปร 2 ชนิดที่ส่อความถึงขนาดและเฟสของปริมาณสเปกตรัมอันได้แก่ a priori SNR และ posterior SNR ตามลำดับนั่นเอง ในขณะที่ การเลือกใช้ผิดเพี้ยนที่ให้ค่าประมาณสเปกตรัมเสียงพูดที่ไม่มีการคำนึงถึงความเป็นเวกเตอร์ของปริมาณสเปกตรัม นำมาซึ่ง Spectral Gain ที่ประกอบไปด้วยปริมาณที่ส่อความถึงเฉพาะขนาดของปริมาณสเปกตรัม ซึ่งได้แก่ a priori SNR เท่านั้น



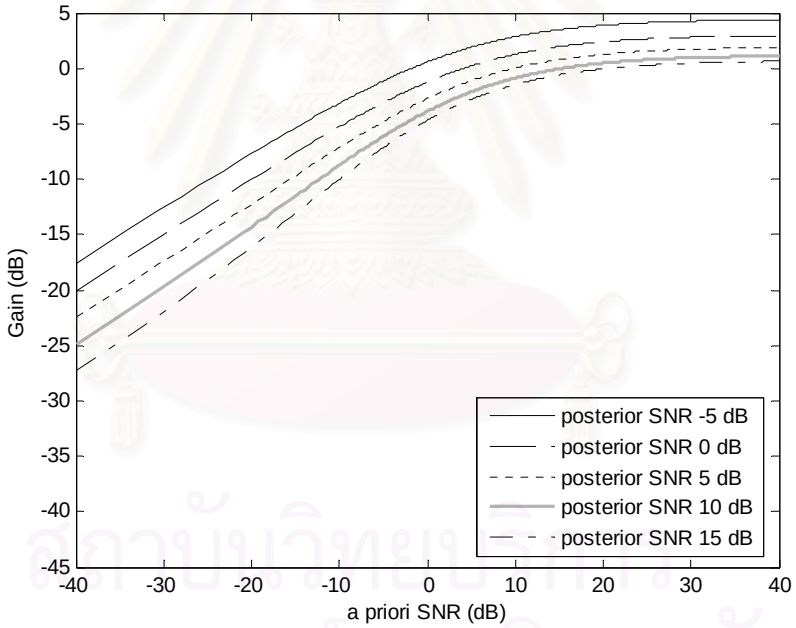
(ก)



(ข)



(ค)



(ง)

รูปที่ 2.5 Spectral Gain ชนิดต่างๆ (ก) G_{SE} (ข) G_{SA} (ค) G_{LSA} และ (ง) G_{SP}

รูปที่ 2.5 ถูกใช้แสดง Spectral Gain ชนิดต่างๆ เทียบกับค่า a priori SNR ในแนวแกนนอน (เนื่องจากเป็นตัวแปรหลัก) และค่า posterior SNR ต่างๆ ตามที่ระบุ จะเห็นได้ว่าสำหรับ Spectral Gain ที่เป็นฟังก์ชันของทั้ง a priori SNR และ posterior SNR แล้ว ณ ค่า a priori SNR ที่เท่ากัน Spectral Gain จะมีมากขึ้นเมื่อค่า posterior SNR มีค่าน้อยลง ซึ่งสอดรับกับที่ได้กล่าวไว้ข้างต้น เนื่องจาก ณ ค่า a priori SNR หนึ่งๆ การที่ค่า posterior SNR มีค่า

น้อยกว่าสามารถส่อความได้ว่า ในขณะนั้นสเปกตรัมเสียงพูดและสเปกตรัมเสียงรบกวนเกิดการบวกแบบไม่เสริมกัน (Destructive addition) ทำให้ไม่สมควรไปกดขนาดสเปกตรัมสัญญาณไมโครโฟนซึ่งมีค่าน้อยอยู่แล้วลงไปอีกนั่นเอง

หลังจากทำการหาค่า Spectral gain แล้ว จะสามารถหาค่าประมาณขนาดสเปกตรัมเสียงพูด ได้จากผลคูณระหว่าง ขนาดสเปกตรัมเสียงพูดที่ถูกรบกวน และ Spectral gain ดังนี้

$$\hat{A}(k,\ell) = G_\eta(k,\ell)R(k,\ell) \tag{2.31}$$

เมื่อ $\hat{A}(k,\ell)$ คือค่าประมาณขนาดสเปกตรัมเสียงพูด และ $G_\eta(k,\ell)$ คือ Spectral gain ตามค่าผิดเพี้ยนใดๆ ตามแต่ตัวห้อย η และจากนั้นเฟสสเปกตรัมเสียงพูดที่ถูกรบกวน $\theta_Y(k,\ell)$ จะถูกรวมเข้ากับ ค่าประมาณขนาดสเปกตรัมเสียงพูด เพื่อให้ได้เป็น ค่าประมาณสเปกตรัมเสียงพูด ดังนี้

$$\hat{S}(k,\ell) = \hat{A}(k,\ell)e^{\,j\theta_Y(k,\ell)} \tag{2.32}$$

ทั้งนี้เนื่องจากหูมนุษย์ไม่ไวต่อการเปลี่ยนแปลงเฟสเหมือนดังเช่นขนาด จึงสามารถนำเฟสสเปกตรัมสัญญาณไมโครโฟน $\theta_Y(k,\ell)$ ซึ่งมีเฟสสเปกตรัมที่เกิดจากสัญญาณรบกวนรวมอยู่ มาใช้ได้ [25] นอกจากนี้จากการคำนวณเชิงสถิติยังพบว่าค่าประมาณเฟสสเปกตรัมเสียงพูดที่ดีที่สุดเมื่อให้มาเพียงสเปกตรัมเสียงพูดที่ถูกรบกวนก็ได้แก่ เฟสสเปกตรัมเสียงพูดที่ถูกรบกวนนั่นเองด้วย [7] โดยอาจเห็นตัวอย่างได้จาก การใช้เฟสจากเฟสสเปกตรัมเสียงพูดที่ถูกรบกวนในสมการที่ (2.20) ซึ่งเป็นผลตอบที่ได้จากการคำนวณเชิงสถิตินั่นเอง จึงอาจรวมสมการที่ (2.31) และ (2.32) ได้เป็น

$$\hat{S}(k,\ell) = G_\eta(k,\ell)Y(k,\ell) \tag{2.33}$$

จากตัวอย่าง Spectral gain ในสมการที่ (2.20), (2.26), (2.27) และ (2.28) จะเห็นได้ว่า Spectral Gain นั้นไม่ว่าจะได้มาจากค่าผิดเพี้ยนแบบใด จะอยู่ในรูปของฟังก์ชันค่าจริงของตัวแปรสองตัวได้แก่ a priori SNR $\xi(k,\ell)$ และ posterior SNR $\gamma(k,\ell)$ เสมอ ซึ่งในทางปฏิบัติแล้วค่าตัวแปรทั้งสองนี้ไม่สามารถได้มาจากการวัดโดยตรง ดังนั้นจึงจำเป็นที่จะต้องทำการประมาณค่าตัวแปรดังกล่าว หัวข้อต่อไปเป็นการกล่าวถึงวิธีการประมาณค่าตัวแปรทั้งสอง

2.1.1.2. การประมาณค่า a priori SNR และ posterior SNR

การประมาณค่า NPSD

จากสมการที่ (2.22) และ (2.30) a priori SNR และ posterior SNR ต่างเป็นฟังก์ชันของ NPSD $\lambda_N(k,\ell)$ โดยจากการที่เสียงรบกวนมีความเป็นจุดนิ่งยาวนาน $\lambda_N(k,\ell)$ จึงอาจถูกประมาณในช่วงที่ไม่มีเสียงพูดได้ดังนี้

$$\hat{\lambda}_N(k,\ell) = \rho\,\hat{\lambda}_N(k,\ell-1) + (1-\rho)|\,Y(k,\ell)\,|^2 \tag{2.34}$$

เมื่อ $\rho \in (0,1)$ คือ ค่าความจำ (Forgetting factor) การประมาณค่า NPSD เช่นนี้ต้องอาศัยตัวตรวจหาเสียงพูด (Voice activity detector, VAD) [6], [24] เพื่อระบุว่าช่วงใดมีหรือไม่มีเสียงพูด อย่างไรก็ตามในปัจจุบันวิธีการประมาณค่า NPSD จำนวนมากถูกพัฒนาขึ้นให้สามารถประมาณค่า NPSD ได้โดยไม่ต้องอาศัยการทำงานของ VAD อีกต่อไป [11], [15], [23], [21]

การประมาณค่า Posterior SNR

หลังจากทำการประมาณค่า NPSD เรียบร้อยแล้ว ค่า posterior SNR $\gamma(k,\ell)$ สามารถถูกประมาณได้จาก

$$\hat{\gamma}(k,\ell) = \frac{|Y(k,\ell)|^2}{\hat{\lambda}_N(k,\ell)} \tag{2.35}$$

การประมาณค่า A priori SNR

ในทางตรงกันข้าม ค่า a priori SNR ไม่อาจถูกประมาณได้โดยตรงดังเช่น posterior SNR เนื่องจากไม่ทราบค่า SPSD $\lambda_S(k,\ell)$ ดังนั้นวิธีการประมาณ a priori SNR จึงถูกนำเสนอขึ้น [7], [12], [18], [20] วิทยานิพนธ์ฉบับนี้จะกล่าวสรุปถึงวิธีการประมาณค่า a priori SNR 3 วิธี คือ Decision Direct (DD) [7], Two-Step Noise Reduction (TSNR) [18] และ Self Adaptive Averaging Factor (SAAF) [20]

I. Decision Direct (DD)

DD เป็นวิธีการประมาณ a priori SNR ที่มีชื่อเสียงอย่างมากตั้งแต่อดีตมาจนกระทั่งถึงปัจจุบัน โดยมีสมการการประมาณดังนี้

$$\hat{\xi}_{\mathrm{DD}}(k,\ell) = \alpha_{\mathrm{DD}}\,\tilde{\xi}(k,\ell-1) + (1-\alpha_{\mathrm{DD}})\,\delta(k,\ell) \tag{2.36}$$

เมื่อ

$$\tilde{\xi}(k,\ell-1) = \frac{\hat{A}^2(k,\ell-1)}{\lambda_N(k,\ell-1)} \tag{2.37}$$

คือ ค่าประมาณ a priori SNR ที่ได้จากค่าประมาณสเปกตรัมเสียงพูดในเฟรมที่แล้ว $\alpha_{\mathrm{DD}} \in (0,1)$ คือ ค่าถ่วงน้ำหนัก (Weighting factor) ซึ่ง [7] เสนอให้ใช้ $\alpha_{\mathrm{DD}} = 0.98$ และ

$$\delta(k,\ell) = \max[\gamma(k,\ell)-1, 0] \tag{2.38}$$

ถูกเรียกว่า Instantaneous SNR ณ องค์ประกอบทางความถี่ที่ k และเฟรมเวลาที่ ℓ

เนื่องด้วย DD มีประสิทธิภาพเหนือการประมาณ a priori SNR อื่นๆ ในช่วง ค.ศ. 1984 ถึง ค.ศ. 2000 เป็นอย่างมากในด้านความสามารถในการลดเสียงรบกวนตกค้างแบบเสียงดนตรี (Musical Noise) (คุณลักษณะของ Musical Noise จะถูกกล่าวถึงในหัวข้อตอนท้ายของหัวข้อย่อยนี้) ซึ่งเป็นข้อเสียหลักของ NS คุณสมบัติการทำงาน

และพฤติกรรมที่ค่อนข้างซับซ้อนของ DD ถูกบรรยายโดยละเอียดใน [9] โดยสามารถสรุปอย่างสังเขปได้เป็นช่วงๆ ของการทำงานได้ดังนี้

1. ช่วงที่ไม่มีเสียงพูด (Non-speech Activity Period)
   - เนื่องจากในช่วงที่ไม่มีเสียงพูด เสียงพูดที่ถูกปรับปรุงจะมีค่าน้อยจนสามารถประมาณพจน์ $\tilde{\xi}(k,\ell-1)$ เป็น 0 ได้ ทำให้สมการประมาณของวิธี DD ลดรูปลงเหลือเพียง

$$\hat{\xi}_{\mathrm{DD}}(k,\ell) \approx (1-\alpha_{\mathrm{DD}})\delta(k,\ell) \tag{2.39}$$

   วิทยานิพนธ์ฉบับนี้จะเรียกสมการที่ (2.39) นี้ว่า สมการช่วงเปลี่ยน (Transition Equation, TE) (การวิเคราะห์และตีความหมายของ TE เพื่อใช้เป็นเครื่องมือในการพัฒนาการประมาณค่า a priori SNR ที่นำเสนอจะถูกบรรยายอย่างละเอียดในหัวข้อย่อยที่ 3.2.1) และจาก TE จะเห็นว่าในช่วงที่ไม่มีเสียงพูดนี้ $\hat{\xi}_{\mathrm{DD}}(k,\ell)$ จะเป็นค่าเฉลี่ยแบบต่ำๆ ของ $\delta(k,\ell)$

   คุณสมบัติข้อนี้เองที่ทำให้ DD สามารถลดปัญหา Musical Noise ลงได้เมื่อใช้กับ Spectral gain ที่มีให้ค่าน้อยมากพอเมื่อค่าประมาณ a priori SNR ที่ได้มีค่าน้อย

2. ในช่วงที่มีเสียงพูด (Speech Activity Period)
   - Instantaneous SNR $\delta(k,\ell)$ มีค่าค่อนข้างสูง
   - $\hat{\xi}_{\mathrm{DD}}(k,\ell)$ จะตามหลัง $\delta(k,\ell)$ อยู่ 1 เฟรมดังนี้

$$\hat{\xi}_{\mathrm{DD}}(k,\ell) \approx \delta(k,\ell-1) \tag{2.40}$$

3. ช่วงเปลี่ยนสถานะ (Transition State)
   - คือ ช่วงที่เกิดการเปลี่ยนแปลงระหว่างช่วงไม่มีเสียงพูดและช่วงที่มีเสียงพูด [22]
   - Instantaneous SNR $\delta(k,\ell)$ จะเกิดการเปลี่ยนแปลงจากที่มีค่าน้อยๆ ไปเป็นมีค่ามาก
   - สมการที่ใช้อธิบายลักษณะการทำงานในช่วงเปลี่ยนสถานะนี้ยังคงได้แก่ TE ทั้งนี้เนื่องจากเฟรมก่อนหน้าก่อนเกิดการเปลี่ยนสถานะนั้นยังคงเป็นช่วงที่ไม่มีเสียงพูดและสามารถประมาณว่า $\tilde{\xi}(k,\ell-1) \approx 0$ ได้เช่นกัน
   - ทั้งนี้ค่าประมาณ a priori SNR ที่ได้จากวิธี DD จะสามารถติดตามการเปลี่ยนแปลง (Tracking) ค่า Instantaneous SNR [4] ได้ทันในช่วงเปลี่ยนสถานะนี้ก็ต่อเมื่อ $\delta(k,\ell) \gg 1/(1-\alpha_{\mathrm{DD}})$

คุณสมบัติของ DD ในช่วงที่ 2 ทำให้มองได้ว่า Spectral gain ที่หาได้มาจาก DD อาจไม่เหมาะสมกับสเปกตรัมเสียงพูดที่ถูกรบกวนในเฟรมนั้น ซึ่งเป็นสาเหตุให้เกิดผลข้างเคียงคือ การก้องกังวาน ในสัญญาณเสียงพูดที่ถูกปรับปรุงได้ [18] ดังนั้น [18] จึงนำเสนอ Two-Step Noise Reduction, TSNR ขึ้นเพื่อช่วยในการแก้ปัญหาดังกล่าว

---

[4] สาเหตุที่ต้องติดตามค่า Instantaneous SNR เป็นเพราะค่าดังกล่าวเป็นตัวแทนที่ดีที่สุดสำหรับค่า a priori SNR ซึ่งจะถูกอธิบายอย่างละเอียดในหัวข้อย่อยที่ 3.2

II. Two-Step Noise Reduction (TSNR)

TSNR มีขั้นตอนการทำงานแบ่งออกเป็น 2 ขั้นดังนี้

1. ทำการประมาณค่า a priori SNR ด้วย DD $\hat{\xi}_{\mathrm{DD}}(k,\ell)$
2. ทำการประมาณค่า a priori SNR $\hat{\xi}_{\mathrm{TSNR}}(k,\ell)$ โดยใช้ค่า Spectral Gain ซึ่งได้จากคำนวณโดยใช้ $\hat{\xi}_{\mathrm{DD}}(k,\ell)$ ดังนี้

$$\hat{\xi}_{\mathrm{TSNR}}(k,\ell)=\left(G_{\psi}(k,\ell)\Big|_{\hat{\xi}_{\mathrm{DD}}(k,\ell),\gamma(k,\ell)}\right)^2\gamma(k,\ell) \tag{2.41}$$

เมื่อ G_{ψ} คือ Spectral gain ใดๆ โดยไม่จำเป็นต้องเหมือนกับที่เลือกใช้ในสมการที่ (2.33) การวิเคราะห์และทดลอง TSNR โดยเลือก $G_{\psi}=G_{\mathrm{SE}}$ หรือ Wiener gain ซึ่งจะเรียกการประมาณ TSNR ที่ใช้ $G_{\psi}=G_{\mathrm{SE}}$ นี้ว่า Two-Step Wiener (TSW) ผลการทดลองชี้ให้เห็นว่า TSW สามารถแก้ปัญหาเรื่องการตามหลังอยู่หนึ่งเฟรมใน DD ได้เป็นอย่างดี [18]

เนื่องจาก TE จะถูกใช้วิเคราะห์และพัฒนาการประมาณค่า a priori SNR ที่นำเสนอในบทที่ 3 ดังนั้นจึงขอทำการหา TE ของ TSW ไว้ ณ ที่นี้ด้วย โดยการประมาณให้พจน์ $\tilde{\xi}(k,\ell-1)$ ของสมการการประมาณค่า a priori SNR ของวิธี TSW มีค่าเป็น 0 จะได้ว่า TE ของวิธี TSW คือ

$$\hat{\xi}_{\mathrm{TSW}}(k,\ell)=\left(\frac{(1-\alpha_{DD})\delta(k,\ell)}{1+(1-\alpha_{DD})\delta(k,\ell)}\right)^2\gamma(k,\ell) \tag{2.42}$$

III. Self Adaptive Averaging Factor (SAAF)

ในปีเดียวกันกับที่ TSNR ถูกนำเสนอ [20] นำเสนอวิธีการประมาณ a priori SNR ซึ่งมีจุดประสงค์ที่จะเพิ่มความสามารถในการติดตามการเปลี่ยนแปลงค่า Instantaneous SNR ได้ดียิ่งขึ้น โดยเฉพาะอย่างยิ่งค่า Instantaneous SNR ที่มีค่าไม่สูงมาก ซึ่ง DD ไม่สามารถทำการติดตามได้ อาศัยข้อสังเกตที่ว่าวิธี DD สามารถทำการติดตามการเปลี่ยนแปลงค่า Instantaneous SNR ได้ดียิ่งขึ้นหากเปลี่ยนค่าถ่วงน้ำหนัก α_{DD} ให้มีค่ามากขึ้น แต่ทั้งนี้ก็แลกมาด้วยเสียงรบกวนตกค้างแบบ Musical Noise ที่เพิ่มมากขึ้น ดังนั้นการใช้ค่าถ่วงน้ำหนักที่เหมาะสมและเปลี่ยนแปลงตามเวลาอาจนำมาซึ่งวิธีการประมาณค่า a priori SNR ที่มีประสิทธิภาพมากยิ่งขึ้นได้

เริ่มจากการทำการหาค่าถ่วงน้ำหนักที่เหมาะสมดังนี้

$$\alpha_{\mathrm{opt}}=\arg\min_{\alpha_{\mathrm{DD}}}E\{(\xi(k,\ell)-\hat{\xi}_{\mathrm{DD}}(k,\ell))^2\mid\tilde{\xi}(k,\ell-1)\} \tag{2.43}$$

ทำให้ได้มาซึ่งค่า α_{opt} ที่สามารถปรับตัวได้ดังต่อไปนี้

$$\alpha_{\mathrm{opt}}(k,\ell)=\frac{1}{1+\left(\dfrac{\delta(k,\ell)-\tilde{\xi}(k,\ell-1)}{\delta(k,\ell)+1}\right)^2} \tag{2.44}$$

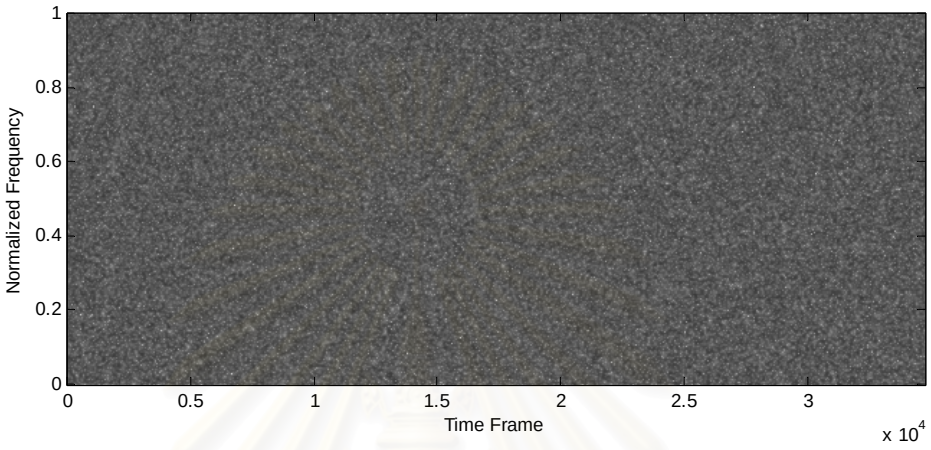
โดยเรียกวิธีการประมาณค่า a priori SNR นี้ว่า Self adaptive averaging factor (SAAF) ดังนั้นต่อไปจะทำการใช้ $\alpha_{\text{SAAF}}(k,\ell)$ แทน $\alpha_{\text{opt}}(k,\ell)$ การทดลอง [20] แสดงให้เห็นว่า SAAF สามารถรักษาไว้ซึ่งองค์ประกอบของเสียงพูดได้มากกว่า DD โดย TE ของ SAAF สามารถเขียนได้เป็น

$$\alpha_{\text{SAAF}}(k,\ell) = \frac{1}{1+\left(\frac{\delta(k,\ell)}{\delta(k,\ell)+1}\right)^2} \tag{2.45}$$

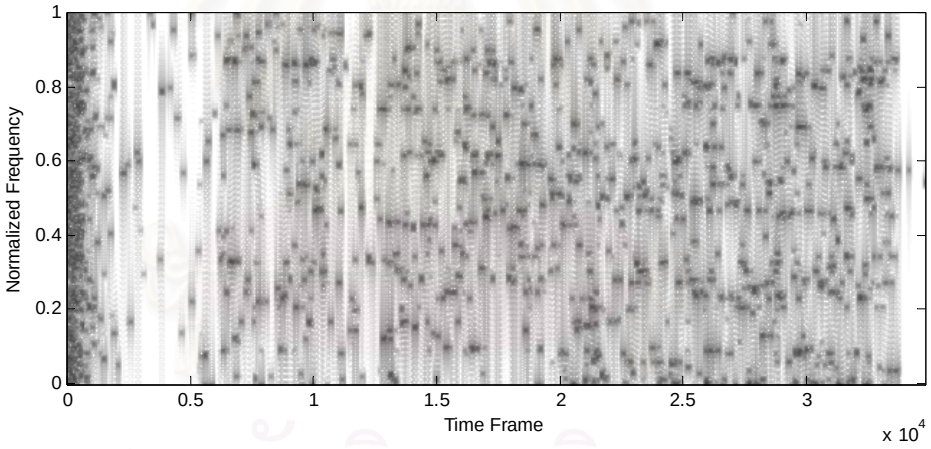
เสียงรบกวนตกค้างแบบ Musical Noise

เสียงรบวนตกค้างแบบ Musical Noise เป็นเสียงรบกวนที่หลงเหลืออยู่ในสัญญาณเสียงขาออกของ NS โดยที่ถูกเรียกว่า Musical Noise ก็เนื่องมาจากเสียงรบกวนตกค้างดังกล่าวมีลักษณะคล้ายเสียงดนตรีซึ่งผู้เล่นทำการเล่นโน้ตแบบสุ่ม และเล่นที่เวลาแบบสุ่มเช่นกัน กล่าวคือหากพิจารณาในเชิงความถี่แล้ว Musical Noise ณ ขณะเวลาหนึ่งๆ จะมีพลังงานอยู่ในองค์ประกอบทางความถี่ใดๆ (ไม่จำเป็นต้องประจำอยู่เพียงหนึ่งความถี่ หรือทุกๆ ความถี่) อย่างสุ่ม และเมื่อพิจารณาในเวลาต่อๆ ไป องค์ประกอบทางความถี่ที่มีพลังงานอยู่ของ Musical Noise ก็จะเปลี่ยนไปอย่างสุ่ม เหตุนี้เองทำให้เสียงที่ได้ยินมีลักษณะคล้ายเสียงโน้ตดนตรี ที่ถูกเล่นอย่างสุ่ม

สาเหตุของการเกิดเสียงรบกวนตกค้างแบบ Musical Noise ใน NS ได้แก่ การประมาณค่า a priori SNR ที่ไม่ราบเรียบพอ เช่น สำหรับในช่วงที่ไม่มีเสียงพูดค่าประมาณ a priori SNR ที่ควรจะเป็นคือ ค่าน้อยๆ และราบเรียบ ทั้งนี้เพื่อจะได้ค่า Spectral Gain (เนื่องจากค่า a priori SNR เป็นตัวแปรหลักของ Spectral Gan จึงมีผลโดยตรงต่อค่า Spectral Gain) ที่มีค่าน้อย และไม่แกว่งมาก อย่างไรก็ตาม การประมาณค่า a priori SNR ในอดีต ไม่สามารถให้ค่าประมาณ a priori SNR ที่ราบเรียบพอ เป็นเหตุให้ค่า Spectral Gain ในแต่ละองค์ประกอบทางความถี่ เกิดการแกว่งตัวอย่างสุ่มในทางเวลา และนำมาซึ่งเสียงรบกวนตกค้างที่หลงเหลือเฉพาะในบางความถี่อย่างสุ่ม และในบางเวลาอย่างสุ่มด้วยเช่นเดียวกัน ตัวอย่างสเปกโทรแกรมของสัญญาณเสียงรบกวนตกค้างแบบ Musical Noise ถูกแสดงดังรูปที่ 2.6



(ก)

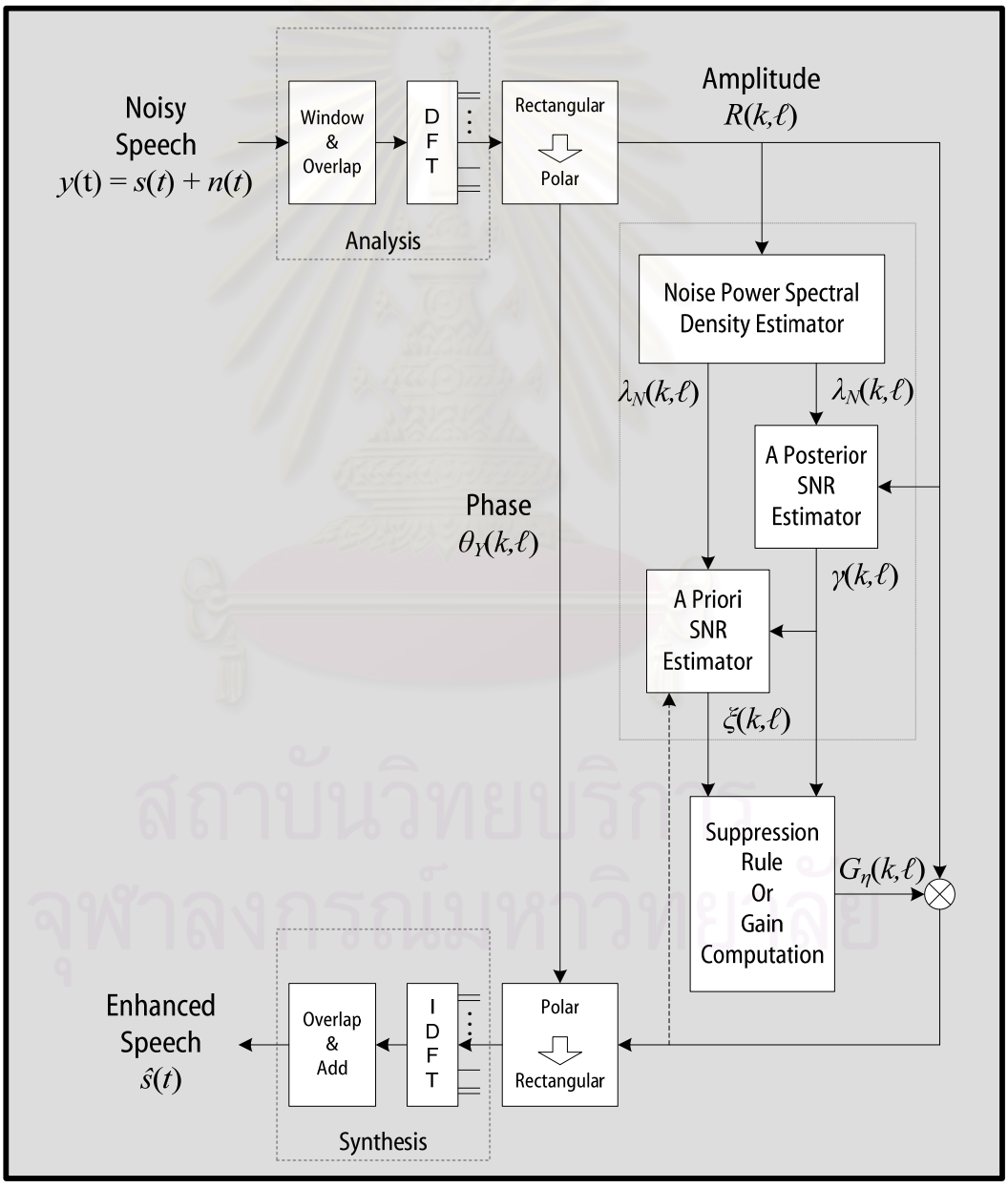


(ข)

รูปที่ 2.6 ตัวอย่างเสียงรบกวนตกค้างแบบ Musical Noise (ก) สเปกโทรแกรมของเสียงรบกวนสีขาว และ (ข) สเปกโทรแกรมของสัญญาณที่ถูกปรับปรุงด้วย NS ที่อาศัยการประมาณค่า a priori SNR แบบ Maximum Likelyhood (ML) [7] ซึ่งเป็นการประมาณค่า a priori SNR ในอดีตที่ไม่สามารถลดผลของเสียงรบกวนตกค้างแบบ Musical Noise ลงได้

2.1.1.3. ระบบ NS

จากหัวข้อย่อยที่ 2.1.1.1 และ 2.1.1.2 ระบบ NS หนึ่งๆ ถูกแสดงได้ดังรูปที่ 2.7 โดยเริ่มจากการแปลงสัญญาณไมโครโฟนที่รับมาได้ไปสู่โดเมน STFT และแยกข้อมูลทางขนาด $R(k,\ell)$ และเฟสสเปกตรัมของสัญญาณไมโครโฟน $\theta_Y(k,\ell)$ ออกจากกัน ค่า NPSD $\lambda_N(k,\ell)$ ถูกประมาณค่าขึ้น จากนั้นค่าประมาณ NPSD $\hat{\lambda}_N(k,\ell)$ จึงถูกใช้ในการประมาณหาค่า posterior SNR $\gamma(k,\ell)$ และ ค่า a priori SNR $\xi(k,\ell)$ เพื่อนำไปใช้ในการคำนวณหาค่า Spectral Gain $G_\eta(k,\ell)$ ตามแต่ที่เลือกไว้ ค่าประมาณขนาดสเปกตรัมเสียงพูด $\hat{A}(k,\ell)$ สามารถหามาได้จากผลคูณระหว่าง Spectral Gain $G_\eta(k,\ell)$ และ ขนาดสเปกตรัมสัญญาณไมโครโฟน $R(k,\ell)$ ก่อนจะถูกนำมารวมกับเฟสสเปกตรัมของสัญญาณไมโครโฟน $\theta_Y(k,\ell)$ และแปลงกลับสู่โดเมนเวลาต่อไป



รูปที่ 2.7 ระบบ NS

2.1.1.4. การกดเสียงรบกวนที่อาศัยความน่าจะเป็นในการมีอยู่ของสเปกตรัมเสียงพูด

เทคนิคการกดเสียงรบกวนในหัวข้อย่อยนี้ไม่ได้ถูกใช้ในการพัฒนาระบบในวิทยานิพนธ์ฉบับนี้ แต่เนื่องจากเป็นเทคนิคที่ค่อนข้างน่าสนใจดังนั้นจึงขอกล่าวถึงอย่างสังเขปไว้ ณ ที่นี้เพื่อความครอบคลุมของเนื้อหาในวิทยานิพนธ์

เนื่องจากเสียงพูดมีคุณสมบัติจุดนิ่งเป็นช่วงๆ ดังนั้นสเปกตรัมเสียงพูดจึงสามารถแบ่งออกได้เป็น 2 Hypothesis ได้แก่ มีเสียงพูด (Speech presence, $H_1(k,\ell)$) และไม่มีเสียงพูด (Speech absence, $H_0(k,\ell)$) กล่าวคือ

$$H_0(k,\ell): Y(k,\ell) = N(k,\ell) \tag{2.46}$$

$$H_1(k,\ell): Y(k,\ell) = S(k,\ell) + N(k,\ell) \tag{2.47}$$

โดยหากคิดรวมถึงการมีอยู่หรือไม่มีของเสียงพูดเช่นนี้ด้วยแล้ว จะทำให้สมการที่ (2.17) สามารถเขียนใหม่ในรูปทั่วไปที่ไม่ขึ้นกับค่าผิดเพี้ยนที่เลือกใช้ได้เป็น

$$\hat{S}(k,\ell) = f_d^{-1}\left[\begin{array}{r} E\{\, f_d[S(k,\ell)] \,|\, Y(k,\ell), H_1(k,\ell)\} P(H_1(k,\ell)|\, Y(k,\ell)) + \\ E\{\, f_d[S(k,\ell)] \,|\, Y(k,\ell), H_0(k,\ell)\} P(H_0(k,\ell)|\, Y(k,\ell)) \end{array}\right] \tag{2.48}$$

โดยที่ $f_d[\cdot]$ คือ ฟังก์ชันที่ขึ้นกับค่าผิดเพี้ยนที่เลือกใช้ เช่น $f_d[\cdot]$ คือ $\log[\cdot]$ ในกรณีที่เลือก d_{LSA} เป็นต้น
$f_d^{-1}[\cdot]$ คือ ฟังก์ชันผกผันของ $f_d[\cdot]$ เช่น $f_d^{-1}[\cdot]$ คือ $\exp[\cdot]$ ในกรณีที่ $f_d[\cdot] = \log[\cdot]$ เป็นต้น
$E\{S(k,\ell)\,|\,Y(k,\ell), H_1(k,\ell)\}$ คือ ค่าประมาณสเปกตรัมเสียงพูดเมื่อทราบว่ามีเสียงพูดอยู่
$P(H_1(k,\ell)|\,Y(k,\ell))$ คือความน่าจะเป็นที่จะมีเสียงพูดอยู่เมื่อให้มาด้วยสเปกตรัมเสียงพูดที่ถูกรบกวน
$E\{S(k,\ell)\,|\,Y(k,\ell), H_0(k,\ell)\}$ คือ ค่าประมาณเสียงพูดเมื่อทราบว่าไม่มีเสียงพูดอยู่ และ
$P(H_0(k,\ell)|\,Y(k,\ell))$ คือ ความน่าจะเป็นที่ไม่มีเสียงพูดอยู่เมื่อให้มาด้วยสเปกตรัมเสียงพูดที่รบกวน

ค่าประมาณสเปกตรัมเสียงพูดเมื่อทราบว่ามีเสียงพูดอยู่ สามารถหามาได้โดยวิธีวิเคราะห์เช่นเดียวกับที่ผ่านมาในหัวข้อ 2.1.1.1 ดังนั้นผลเฉลยของพจน์นี้จึงเป็นเช่นเดียวกับที่ได้มาในหัวข้อดังกล่าวด้วย เช่น ในกรณีที่ใช้ค่าผิดเพี้ยน d_{LSA} และการแจกแจงความน่าจะเป็นแบบเกาส์เซียนทั้งกับสเปกตรัมเสียงพูดและเสียงรบกวน จะได้ว่า

$$\exp\left(E\{\log[S(k,\ell)]\,|\,Y(k,\ell), H_1(k,\ell)\}\right) = G_{\mathrm{LSA}}(k,\ell)Y(k,\ell) \tag{2.49}$$

เป็นต้น

ความน่าจะเป็นที่จะมีเสียงพูดอยู่เมื่อให้มาด้วยสเปกตรัมเสียงพูดที่ถูกรบกวน $P(H_1(k,\ell)|\,Y(k,\ell))$ สามารถคำนวณได้จาก Bayes' rule โดยจากสมการที่ (2.46) และ (2.47) และกำหนดการแจกแจงความน่าจะเป็นแบบเกาส์เซียนให้ทั้งกับสเปกตรัมเสียงพูดและเสียงรบกวน จะได้ว่า

$$p(Y(k,\ell)\,|\,H_0(k,\ell)) = \frac{1}{\pi\lambda_N(k,\ell)}\exp\left(-\frac{|\,Y(k,\ell)\,|^2}{\lambda_N(k,\ell)}\right) \tag{2.50}$$

$$p(Y(k,\ell)\,|\,H_1(k,\ell)) = \frac{1}{\pi(\lambda_N(k,\ell)+\lambda_S(k,\ell))}\exp\left(-\frac{|\,Y(k,\ell)\,|^2}{(\lambda_N(k,\ell)+\lambda_S(k,\ell))}\right) \tag{2.51}$$

เมื่อ $\lambda_S(k,\ell) = E\{|\,S(k,\ell)\,|^2\,|\,H_1(k,\ell)\}$ และ $\lambda_N(k,\ell) = E\{|\,N(k,\ell)\,|^2\}$ คือค่าความแปรปรวนสเปกตรัมเสียงพูดและค่าความแปรปรวนสเปกตรัมเสียงรบกวนตามลำดับ ซึ่งจะเห็นว่าค่าความแปรปรวนสเปกตรัมเสียงพูดในที่นี้ต่างจากในสมการที่ (2.13) ตรงที่ในที่นี้มีการคำนึงถึงช่วงการมีเสียงพูดอยู่ด้วยเท่านั้น

จาก Bayes' rule จะได้ว่า

$$P(H_1(k,\ell)\,|\,Y(k,\ell)) = \frac{\Lambda(k,\ell)}{\Lambda(k,\ell)+1} \triangleq p(k,\ell) \tag{2.52}$$

เมื่อ

$$\Lambda(k,\ell) = \frac{1-q(k,\ell)}{q(k,\ell)}\frac{p(Y(k,\ell)\,|\,H_1(k,\ell))}{p(Y(k,\ell)\,|\,H_0(k,\ell))} \tag{2.53}$$

เมื่อ $q(k,\ell)$ คือ ความน่าจะเป็นก่อนประสบของการไม่มีสเปกตรัมเสียงพูด (A priori probability for speech absnce) อาศัยสมการที่ (2.50) และ (2.51) จะได้ว่า

$$p(k,\ell) = \left[1+\frac{q(k,\ell)}{q(k,\ell)+1}(1+\xi(k,\ell))\exp(-v(k,\ell))\right]^{-1} \tag{2.54}$$

และ

$$P(H_0(k,\ell)\,|\,Y(k,\ell)) = 1-p(k,\ell) \tag{2.55}$$

ค่าประมาณสเปกตรัมเสียงพูดเมื่อทราบว่าไม่มีเสียงพูดอยู่ ถูกกำหนดให้เป็นดังนี้

$$E\{S(k,\ell)\,|\,Y(k,\ell),H_0(k,\ell)\} = G_{\min}(k,\ell)Y(k,\ell) \tag{2.56}$$

โดย $G_{\min}(k,\ell)$ เป็นค่าคงที่น้อยๆ ค่าหนึ่งซึ่งสามารถเลือกหาค่าที่เหมาะสมได้จากผลความเป็นธรรมชาติของเสียงรบกวนตกค้างในเสียงพูดที่ถูกปรับปรุง

ตัวอย่างการกดเสียงรบกวนที่อาศัยความน่าจะเป็นในการมีอยู่ของสเปกตรัมเสียงพูด ที่ใช้ค่าผิดเพี้ยนแบบ d_{LSA} ได้แก่ จากสมการที่ (2.48) จะได้ว่า

$$\hat{A}(k,\ell) = \exp\left[\begin{array}{l} E\{\log[A(k,\ell)]\,|\,Y(k,\ell),H_1(k,\ell)\}p(k,\ell)+ \\ \qquad E\{\log[A(k,\ell)]\,|\,Y(k,\ell),H_0(k,\ell)\}(1-p(k,\ell)) \end{array}\right] \tag{2.57}$$

$$\hat{A}(k,\ell) = \left\{\exp[E\{\log(A(k,\ell)) \mid Y(k,\ell), H_1(k,\ell)\}]\right\}^{p(k,\ell)} \times \left\{\exp[E\{\log(A(k,\ell)) \mid Y(k,\ell), H_0(k,\ell)\}]\right\}^{(1-p(k,\ell))} \tag{2.58}$$

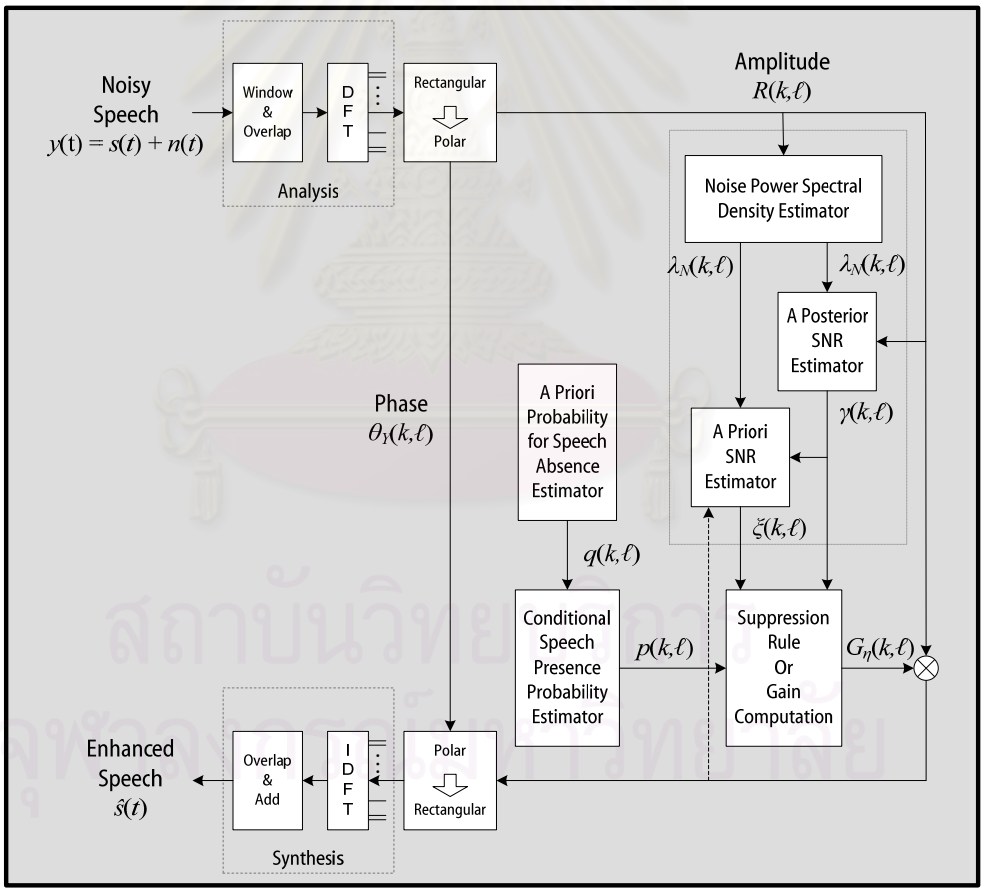
$$\hat{A}(k,\ell) = \left\{G_{\mathrm{LSA}}(k,\ell)R(k,\ell)\right\}^{p(k,\ell)} \times \left\{G_{\min}(k,\ell)R(k,\ell)\right\}^{(1-p(k,\ell))} \tag{2.59}$$

$$\hat{A}(k,\ell) = \left(G_{\mathrm{LSA}}(k,\ell)\right)^{p(k,\ell)} \left(G_{\min}(k,\ell)\right)^{(1-p(k,\ell))} R(k,\ell) \tag{2.60}$$

สมการที่ (2.60) นำมาซึ่ง Spectral gain ที่เรียกว่า Optimally-Modified Log-Spectral Amplitude, OMLSA [11]

$$\hat{A}(k,\ell) = G_{\mathrm{OMLSA}}(k,\ell)R(k,\ell) \tag{2.61}$$

ระบบการกดเสียงรบกวนโดยมีการอาศัยความน่าจะเป็นของสเปกตรัมสียงพูดถูกแสดงในรูปที่ 2.8



รูปที่ 2.8 ระบบ NS ที่อาศัยความน่าจะเป็นในการมีอยู่ของเสียงพูด

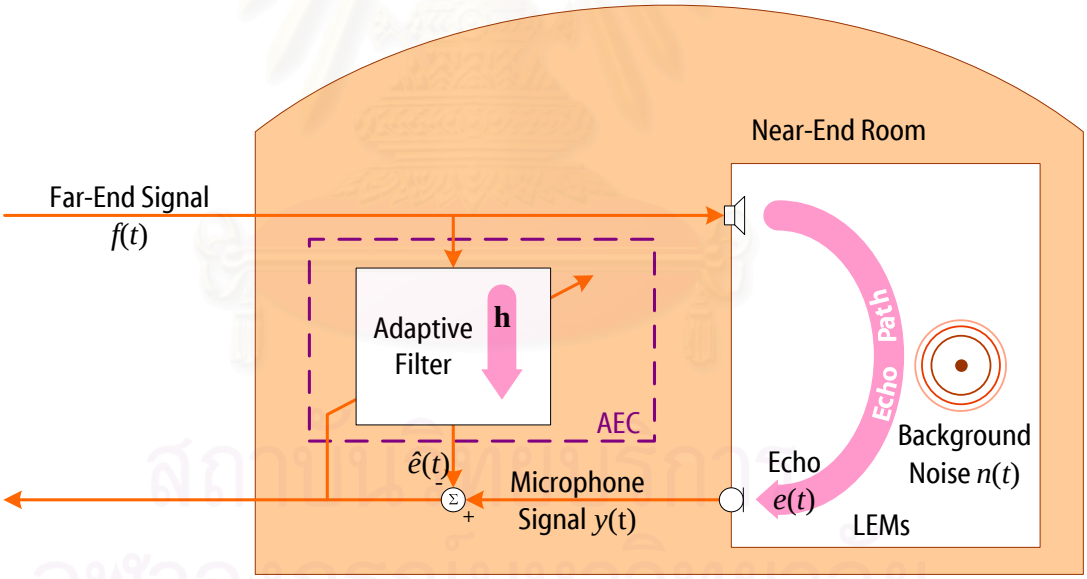
2.2. การลดเสียงสะท้อน

ในหัวข้อนี้จะกล่าวโดยสังเขปถึงวิธีการลดเสียงสะท้อนที่นิยมใช้กันอย่างแพร่หลาย อันได้แก่ การหักล้างเสียงสะท้อน (Acoustic Echo Cancellation, AEC) เพื่อใช้อ้างอิงและเปรียบเทียบประสิทธิภาพในด้านต่างๆ กับวิธีการกดเสียงสะท้อน (Acoustic Echo Suppression, AES) ซึ่งเป็นเทคนิคที่จะได้รับการปรับใช้กับวิธีการลดเสียงสะท้อนที่นำเสนอในวิทยานิพนธ์ฉบับนี้ เนื้อหาของหัวข้อนี้ประกอบไปด้วย 2 ส่วนหลักได้แก่ ส่วนของการหักล้างเสียงสะท้อน และส่วนของการกดเสียงสะท้อน

2.2.1. การหักล้างเสียงสะท้อน (Acoustic Echo Cancellation, AEC)

2.2.1.1. หลักการของ AEC

หลักการของ AEC คือ การใช้ค่าประมาณสัญญาณเสียงสะท้อน $\hat{e}(t)$ ไปหักลบออกจากสัญญาณเสียงสะท้อนจริง $e(t)$ ที่ปะปนอยู่ในสัญญาณไมโครโฟน $y(t)$ ดังรูปที่ 2.9 การกระทำเช่นนี้ส่งผลให้เกิดการสื่อสารแบบสองทางเต็มอัตราที่ปราศจากเสียงสะท้อนขึ้นได้ [33] ประสิทธิภาพของ AEC ขึ้นอยู่กับความสามารถในการประมาณค่าสัญญาณเสียงสะท้อน ทั้งนี้หากสามารถประมาณค่าสัญญาณเสียงสะท้อนได้ดีเพียงใด เสียงสะท้อนจริงก็จะถูกหักลบออกไปได้มากเท่านั้น



รูปที่ 2.9 AEC ในระบบการสนทนาแบบแฮนด์ฟรี

ในทางปฏิบัติจะเห็นได้ว่าสัญญาณเสียงสะท้อน $e(t)$ เกิดจากการที่สัญญาณเสียงทางห้องไกล $f(t)$ ถูกขยายด้วยลำโพงและสะท้อนกลับไปมาในสภาพแวดล้อมของห้องใกล้ก่อนจะถูกรับได้โดยไมโครโฟนในห้องใกล้ โดยระบบที่ทำให้เกิดเหตุการณ์สะท้อนดังกล่าวนี้ถูกเรียกว่า ระบบลำโพงใกล้ไมโครโฟน (Loudspeaker Encloser Microphone system, LEMs) และวิถีทางที่แปรเปลี่ยนสัญญาณเสียงทางห้องไกลให้กลายเป็นสัญญาณเสียงสะท้อนดังกล่าวถูกเรียกว่า “วิถีสะท้อนทางเสียง” (Echo path) เนื่องจากสัญญาณเสียงทางห้องไกลเป็นข้อมูลซึ่ง

สามารถรับมาได้ดังนั้นหากสามารถรู้ หรือประมาณวิถีสะท้อนทางเสียงดังกล่าวได้แล้ว ก็จะสามารถทำการสังเคราะห์ค่าประมาณสัญญาณเสียงสะท้อนขึ้นมาได้โดยการนำสัญญาณเสียงทางห้องไกลมาผ่านวิถีสะท้อนทางเสียงที่ประมาณขึ้นนั่นเอง

โดยมาก[5]วิถีสะท้อนทางเสียงถูกสมมติให้เป็นระบบ LTI และมัก[6]ถูกจำลองด้วยวงจรกรองที่มีผลตอบสนองอิมพัลส์จำกัด (FIR Filter) $\mathbf{h}$ ดังนั้นค่าประมาณสัญญาณเสียงสะท้อน $\hat{e}(t)$ จึงสามารถหาได้ดังนี้

$$\hat{e}(t) = \mathbf{h}^T \mathbf{f}(t) \tag{2.62}$$

เมื่อ t เป็นตัวบ่งชี้เวลาแบบไม่ต่อเนื่อง (Discrete time index) $\mathbf{h} = [h(0) \quad h(1) \quad \cdots \quad h(L-1)]^T$ คือวงจรกรองที่มีผลตอบสนองอิมพัลส์จำกัด $\cdot^T$ คือ เครื่องหมายเมทริกซ์สลับเปลี่ยน (Transpose of matrix) $\mathbf{f}(t) = [f(t) \quad f(t-1) \quad \cdots \quad f(t-L+1)]^T$ คือเวกเตอร์ของสัญญาณทางห้องไกล และ L คือจำนวนสัมประสิทธิ์ของวงจรกรอง $\mathbf{h}$

จากการจำลองที่กล่าวมาจะเห็นว่าปัญหาการประมาณค่าสัญญาณเสียงสะท้อนถูกลดลงเหลือเพียงปัญหาการประมาณค่า วิถีสะท้อนทางเสียง หรือวงจรกรอง $\mathbf{h}$ ที่เหมาะสมเท่านั้น วิธีการประมาณค่าวิถีสะท้อนทางเสียงดังกล่าวได้รับการพัฒนาจากนักวิจัยจำนวนมากมาเป็นระยะเวลายาวนาน และจะถูกกล่าวถึงในหัวข้อต่อไปนี้

2.2.1.2. การประมาณค่าวิถีสะท้อนทางเสียง

ในทางปฏิบัติ*วงจรกรองปรับตัว*จะถูกใช้ในการประมาณหาวิถีสะท้อนทางเสียงดังกล่าว โดยเมื่อสมมติให้วิถีสะท้อนทางเสียงเป็นวงจรกรอง FIR แล้ว วงจรกรองปรับตัวที่ใช้จึงมีลักษณะเป็น FIR ด้วย วงจรกรองปรับตัวจะปรับตัวอย่างอัตโนมัติเพื่อลู่เข้าสู่คำตอบหนึ่งๆ โดยขั้นตอนวิธีการที่ใช้ในการปรับตัวของวงจรกรองปรับตัวถูกพัฒนามาจากขั้นตอนวิธีการหาค่าที่เหมาะสมที่สุดสำหรับสัญญาณเชิงกำหนดแบบวนรอบ (Iterative Deterministic Optimization, IDO) ซึ่งอาศัยฟังก์ชันต้นทุน (Cost Function) ที่เป็นค่าผิดพลาดกำลังสอง (Square Error) ดังนี้

$$J_{\text{IDO}} = (x - \hat{x})^2 \tag{2.63}$$

เมื่อเปลี่ยนจากการอาศัยข้อมูลจากสัญญาณเชิงกำหนด (Deterministic) ชุดเดียวกันในแต่ละรอบการปรับตัวของขั้นตอนวิธี IDO เป็นการอาศัยข้อมูลชุดใหม่ที่เข้าถึงได้ ณ ขณะเวลาที่ทำการปรับตัว จะทำให้ได้วิธีการชนิดใหม่ซึ่งถูกเรียกว่า ขั้นตอนวิธีการประมาณเชิงสโทแคสติกแบบวนรอบ (Recursive Stochastic Approximation

---

[5] นอกจากระบบ LTI แล้ว ยังมีงานวิจัยจำนวนหนึ่งที่สมมติให้วิถีสะท้อนทางเสียงเป็นระบบไม่ชิงเส้น แต่อย่างไรก็ตามงานวิจัยดังกล่าวอยู่นอกเหนือขอบเขตของวิทยานิพนธ์ฉบับนี้

[6] โดยส่วนใหญ่แล้วจะทำการจำลองวิถีสะท้อนทางเสียงด้วย FIR แต่นอกจากนี้ยังมีงานวิจัยที่ใช้วงจรกรองที่มีผลตอบสนองอิมพัลส์อนันต์ (IIR Filter) ในการจำลองวิถีสะท้อนทางเสียง ซึ่งอยู่นอกเหนือขอบเขตของวิทยานิพนธ์ฉบับนี้

Algorithm, RSA) [34] การอาศัยข้อมูลชุดใหม่ในทุกรอบการปรับตัวนี้ ทำให้ฟังก์ชันต้นทุนของ RSA เปลี่ยนเป็นค่าเฉลี่ยของค่าผิดพลาดกำลังสอง (Mean Square Error, MSE) ดังนี้

$$J_{\mathrm{RSA}} = E\{(x - \hat{x})^2\} \tag{2.64}$$

แต่เดิมปริมาณที่สำคัญ และถูกใช้เพื่อหาเส้นทางการปรับตัวด้วยเทคนิคต่างๆ ใน IDO ได้แก่ ค่าของฟังก์ชันต้นทุน และ/หรือ ค่าเกรเดียนต์ของฟังชันต้นทุน และ/หรือ ค่าเฮสเซียนเมทริกซ์ของฟังก์ชันต้นทุน ซึ่งปริมาณต่างๆ เหล่านี้ไม่สามารถหาได้โดยตรง เมื่อฟังก์ชันต้นทุนดังกล่าวอยู่ในรูปของปริมาณเชิงสโตแคสติก ดังเช่นในกรณีของ RSA ดังนั้นการประมาณค่าฟังก์ชันต้นทุนที่เป็นปริมาณเชิงสโตแคสติกให้กลายเป็นปริมาณเชิงกำหนดจะทำให้สามารถนำเทคนิคสำหรับ IDO ที่มีอยู่เดิมมาใช้กับ RSA ได้

การประมาณค่าฟังก์ชันต้นทุนในรูปแบบต่างๆ รวมทั้งการเลือกใช้เทคนิคที่แตกต่างกันในการปรับตัว ทำให้ได้มาซึ่งขั้นตอนวิธีสำหรับ RSA จำนวนมาก อาทิเช่น NLMS [3], FRLS [34] และ APA [35] เป็นต้น โดยในส่วนต่อไปจะเป็นการบรรยายถึงขั้นตอนวิธี RSA ได้แก่ LMS และ NLMS ที่ถูกพัฒนาขึ้นจากเทคนิคการปรับบตัวตามเส้นทางที่ชันที่สุด (Steepest-descent) ของขั้นตอนวิธี IDO

ขั้นตอนวิธีที่อาศัยหลักการของ Steepest-Descent

ในขั้นตอนวิธี IDO นั้น หลักการของ Steepest-Descent คือ อาศัยเส้นทางซึ่งฟังก์ชันต้นทุนมีความชันมากที่สุดเป็นเส้นทางในการปรับตัวเข้าสู่ค่าต่ำสุดหรือค่าสูงสุดของฟังก์ชันต้นทุน ทำให้ Steepest-Descent เป็นเทคนิคในการปรับตัวที่มีความเรียบง่ายที่สุดวิธีการหนึ่ง เนื่องจากอาศัยเพียงค่าเกรเดียนต์ของฟังก์ชันต้นทุนๆ ณ ตำแหน่งที่พิจารณาในการบอกทิศทางการปรับตัวเท่านั้น

การนำหลักการของ Steepest-Descent ใน IDO มาประยุกต์ใช้ใน RSA ก็เพียงเปลี่ยนรูปแบบฟังก์ชันต้นทุนให้อยู่ในรูปของค่าเชิงสโตแคสติกเท่านั้น โดยในวิทยานิพนธ์ฉบับนี้จะขออธิบายโดยอิงการทำงานและตัวแปรของวิธีการหักล้างเสียงสะท้อน (ดูรูปที่ 2.9 ประกอบ) จุดมุ่งหมายของการหักล้างเสียงสะท้อนคือ การหาค่าประมาณสัญญาณเสียงสะท้อน $\hat{e}(t)$ ให้ได้ใกล้เคียงกับ สัญญาณเสียงสะท้อนจริง $e(t)$ มากที่สุด หรือกล่าวคือหาจุดต่ำสุดของฟังก์ชันต้นทุน

$$J = E\{(e(t) - \hat{e}(t))^2\} \tag{2.65}$$

เมื่อละตัวห้อย RSA ไว้ เพื่อความกระชับ

จากแบบจำลองของวิถีสะท้อนทางเสียงในสมการที่ (2.62) ทำให้สมการที่ (2.65) สามารถเขียนได้ดังนี้

$$J(\mathbf{h}) = E\{(e(t) - \mathbf{h}^T \mathbf{f}(t))^2\} \tag{2.66}$$

จุดมุ่งหมายของการหักล้างเสียงสะท้อนการก็คือ หา $\mathbf{h}$ ซึ่งทำให้ $J_{\mathrm{RSA}}(\mathbf{h})$ มีค่าต่ำที่สุด (และเรียก $\mathbf{h}$ นั้นว่า $\mathbf{h}_{opt}$) โดยสามารถเขียนเป็นสมการทางคณิตศาสตร์ได้ว่า

$$\mathbf{h}_{opt} = \arg\min_{\mathbf{h}} J(\mathbf{h}) \tag{2.67}$$

จากสมการที่ (2.66) สามารถแสดงได้ว่า [3]

$$\nabla_{\mathbf{h}} J(\mathbf{h}) = \mathbf{R}\,\mathbf{h} - \mathbf{p} \tag{2.68}$$

โดยที่ $\nabla_{\mathbf{h}}$ คือเครื่องหมายเกรเดียนต์ที่ถูกหาเทียบกับ $\mathbf{h}$

$$\mathbf{R} = E\{\mathbf{f}(t)\,\mathbf{f}(t)^T\} \tag{2.69}$$

คือ เมทริกซ์สหสัมพันธ์อัตโนมัติของสัญญาณเสียงทางห้องไกล และ

$$\mathbf{p} = E\{\mathbf{f}(t)\,e(t)\} \tag{2.70}$$

คือ เวกเตอร์สหสัมพันธ์ข้ามระหว่าง สัญญาณทางห้องไกล และสัญญาณเสียงสะท้อน ทำให้สามารถหาค่า $\mathbf{h}_{opt}$ ได้ดังนี้

$$\nabla_{\mathbf{h}} J(\mathbf{h})\Big|_{\mathbf{h}_{opt}} = 0 = \mathbf{R}\,\mathbf{h}_{opt} - \mathbf{p} \tag{2.71}$$

$$\mathbf{h}_{opt} = \mathbf{R}^{-1}\mathbf{p} \tag{2.72}$$

จากหลักการของ Steepest-Descent สมการปรับตามกาล (Updated equation) จึงสามารถเขียนได้เป็น

$$\mathbf{h}(t+1) = \mathbf{h}(t) + \mu\left(-\nabla J(\mathbf{h})\Big|_{\mathbf{h}(t)}\right) \tag{2.73}$$

เมื่อ $\mathbf{h}(t) = [h_t(0) \quad h_t(1) \quad \cdots \quad h_t(L-1)]^T$ คือสัมประสิทธิ์ของระบบเชิงเส้นผลตอบจำกัด ณ เวลา t (เนื่องจาก $\mathbf{h}$ ปรับตัวไปตามเวลา จึงต้องอาศัยตัวบ่งชี้เวลาเพื่อแบ่งแยกความแตกต่างกันในแต่ละเวลานี้)

$$\mu \in [0, \frac{2}{\lambda_{\max}}) \tag{2.74}$$

คือค่าช่วงก้าว (Step-size) เมื่อ $\lambda_{\max}$ คือค่าเจาะจงที่มากที่สุดของเมทริกซ์ $\mathbf{R}$ และ $\nabla_{\mathbf{h}} J(\mathbf{h})\Big|_{\mathbf{h}(t)} = \mathbf{R}\,\mathbf{h}(t) - \mathbf{p}$ คือ เกรเดียนต์ของฟังก์ชันต้นทุนที่ตำแหน่ง $\mathbf{h}(t)$

สมการที่ (2.73) สามารถอธิบายได้อย่างง่ายคือ ณ $\mathbf{h}(t)$ ใดๆ Steepest-Descent จะอาศัยทิศทางของเกรเดียนต์ของฟังก์ชันต้นทุน ณ ตำแหน่ง $\mathbf{h}(t)$ นั้นๆ (หรือทิศทางที่มีความชันมากที่สุดนั่นเอง) ในการก้าวต่อ จาก $\mathbf{h}(t)$ ไปสู่ตำแหน่งต่อไป $\mathbf{h}(t+1)$ ด้วยค่าช่วงก้าว μ ที่เลือกใช้ และ ณ $\mathbf{h}_{opt}$ เกรเดียนต์ของฟังก์ชันต้นจะมีค่าเท่ากับ 0 ดังนั้นระบบจะไม่ทำการปรับตัวอีก

สังเกตว่าสมการที่ (2.73) ใช้ทิศทางที่ตรงข้ามกับเกร์เดียนท์ในการก้าวหรือปรับตัว ทั้งนี้เนื่องจากจุดมุ่งหมายคือต้องการหาค่าที่ต่ำที่สุดของฟังก์ชันต้นทุนในขณะที่เกรเดียนต์ของฟังก์ชันต้นทุนนั้นจะชี้ในทิศทางที่ชันที่สุดที่ฟังก์ชันต้นทุนมีค่าเพิ่มขึ้นนั่นเอง

ตัวแปรซึ่งมีความสำคัญในเทคนิค Steepest-Descent นี้คือค่าช่วงก้าว ซึ่งหากดูอย่างสังเขป จะสามารถสรุปได้ว่า การเลือกค่าช่วงก้าวที่มีค่ามาก จะทำให้ได้ผลการลู่เข้าสู่คำตอบที่รวดเร็ว แต่ให้ค่าผิดพลาดจาก $\mathbf{h}_{opt}$ ในช่วงสภาวะคงตัวสูง เนื่องจากเกิดการแกว่งไปมารอบจุดที่ต่ำที่สุดอย่างมากด้วยเช่นกัน ส่วนการเลือกใช้ค่าช่วงก้าวที่ต่ำก็จะให้ผลตรงกันข้ามคือให้อัตราการลู่เข้าสู่คำตอบที่ช้ากว่า แต่ความผิดเพลาดจาก $\mathbf{h}_{opt}$ ในช่วงสภาวะคงตัวน้อยกว่า นอกจากนี้หากเลือกค่าช่วงก้าวที่มีค่ามากเกินกว่า $2/\lambda_{\max}$ จะทำให้การปรับตัวเกิดการลู่ออกได้ การวิเคราะห์เกี่ยวกับค่าช่วงก้าว μ ถูกบรรยายอย่างละเอียดใน [3]

ขั้นตอนวิธีการปรับตัวที่อาศัยหลักการของเทคนิค Steepest-Descent ที่มีชื่อเสียงเป็นอย่างมาก ได้แก่ LMS และ NLMS รวมทั้งเทคนิค VSNLMS ซึ่งจะถูกใช้ในการทดลองในบทที่ 4 จะถูกกล่าวถึงในหัวข้อย่อยต่อไปนี้

I. ขั้นตอนวิธี LMS

ขั้นตอนวิธีการปรับตัวที่อาศัยเทคนิค Steepest-Descent นำมาซึ่งสมการปรับตามกาลดังในสมการที่ (2.73) โดยจะเห็นว่าค่าเกร์เดียนท์ของฟังก์ชันต้นทุน $\nabla_{\mathbf{h}} J(\mathbf{h})\big|_{\mathbf{h}(t)} = \mathbf{R}\,\mathbf{h}(t) - \mathbf{p}$ ยังคงประกอบด้วยตัวแปรซึ่งเป็นปริมาณเชิงสโตแคสติกอยู่ 2 ตัวได้แก่ $\mathbf{R}$ และ $\mathbf{p}$ การประมาณตัวแปรทั้งสองโดยแทนด้วยค่า ณ ขณะนั้น (Instantaneous) กล่าวคือ

$$\mathbf{R} = \mathbf{f}(t)\,\mathbf{f}(t)^T \tag{2.75}$$

$$\mathbf{p} = \mathbf{f}(t)\,e(t) \tag{2.76}$$

ทำให้ได้มาซึ่งขั้นตอนวิธีการปรับตัวที่มีชื่อเสียงซึ่งรู้จักกันในชื่อ Least Mean Square (LMS)

สมการปรับตามกาลของ LMS ได้แก่

$$\mathbf{h}(t+1) = \mathbf{h}(t) + \mu\,\mathbf{f}(t)\Big(e(t) - \mathbf{h}^T(t)\,\mathbf{f}(t)\Big) \tag{2.77}$$

โดยสามารถหาได้จากการแทนค่า $\mathbf{R}$ และ $\mathbf{p}$ ในสมการที่ (2.75) และ (2.76) ลงในสมการที่ (2.73)

II. ขั้นตอนวิธี NLMS

หากทำการหาค่าช่วงก้าวที่เหมาะสมด้วยเทคนิค Optimal Line Search [34] ดังเช่นที่เคยมีมาใน IDO จะทำให้ได้สมการปรับตามกาลของขั้นตอนวิธีการที่เรียกว่า Normalized Least Square (NLMS) [3] ดังนี้

$$\mathbf{h}(t+1) = \mathbf{h}(t) + \frac{\mu\,\mathbf{f}(t)\Big(e(t) - \mathbf{h}^T(t)\,\mathbf{f}(t)\Big)}{\Big(\|\mathbf{f}(t)\|^2 + \delta\Big)} \tag{2.78}$$

เมื่อ

$$\mu \in [0,2) \tag{2.79}$$

คือ ค่าช่วงก้าว

$$\left\|\mathbf{f}(t)\right\|^2 = \mathbf{f}^T(t)\mathbf{f}(t) \tag{2.80}$$

คือ กำลังสองสัญญาณทางห้องไกล และ δ คือ ค่าคงที่น้อยๆ เพื่อป้องกันกรณีการเกิดส่วนเป็น 0

ซึ่งเห็นได้ว่าค่าช่วงก้าวของ NLMS จะไม่ขึ้นอยู่กับกำลังของสัญญาณทางห้องไกล เหมือนดังเช่นในกรณีของ LMS (สมการที่ (2.74)) อีกต่อไป ทำให้ NLMS เหมาะสมกับการใช้งานในทางปฏิบัติมากกว่า LMS เป็นอย่างยิ่ง เนื่องจากในทางปฏิบัติเสียงจากทางห้องไกลอาจมีกำลังไม่คงที่ตลอดการสนทนา

III. Variable Step-Size NLMS (VSNLMS) X[43]

เนื่องจากในส่วนของผลการทดลองในบทที่ 4 Variable Step-Size NLMS (VSNLMS) [43] ถูกนำไปใช้เปรียบเทียบกับวิธีที่นำเสนอ ดังนั้นจึงขอกล่าวถึง VSNLMS ไว้ ณ ที่นี้ด้วย

VSNLMS ใน [43] ถูกออกแบบมาเพื่อจุดประสงค์ที่จะทำให้วงจรกรองปรับตัวสามารถทำงานได้ภายใต้สภาพแวดล้อมที่มีสัญญาณเสียงรบกวน โดยทำการปรับปรุงค่าช่วงก้าว μ ใน NLMS ให้สามารถปรับเปลี่ยนค่าได้อย่างอัตโนมัติขึ้นกับ ค่าประมาณสหสัมพันธ์ข้ามระหว่างสัญญาณเสียงทางห้องไกล $f(t)$ (สัญญาณขาเข้าของวงจรกรองปรับตัว) และสัญญาณผิดพลาด (Error signal)

$$d(t) = \left(e(t) - \mathbf{h}^T(t)\mathbf{f}(t)\right) \tag{2.81}$$

ซึ่ง ในกรณี STS สัญญาณผิดพลาดนี้จะประกอบไปด้วย สัญญาณเสียงรบกวนพื้นหลัง $n(t)$ และสัญญาณเสียงสะท้อนตกค้างนั่นเอง

หลักการของ VSNLMS คือ เมื่อใดที่ค่าประมาณสหสัมพนธ์ข้ามระหว่าง $f(t)$ และ $d(t)$ มีค่าสูง ซึ่งส่อความถึงว่า เสียงสะท้อนตกค้างยังคงมีค่ามากอยู่ จะใช้ค่าช่วงก้าวที่มีขนาดใหญ่ ทั้งนี้เพื่อเพิ่มอัตราการลู่เข้าให้กับวงจรกรองปรับตัว และเมื่อ ค่าประมาณสหสัมพนธ์ข้ามระหว่าง $f(t)$ และ $d(t)$ มีค่าต่ำ ซึ่งส่อความถึงว่า เสียงสะท้อนตกค้างมีปริมาณที่น้อยแล้ว จะใช้ค่าช่วงก้าวที่มีขนาดเล็ก ทั้งนี้เพื่อให้วงจรกรองสามารถรักษาสถานะอยู่ตัวเอาไว้ได้แม้กระทั่งในสภาพที่สัญญาณเสียงรบกวนมีพลังงานสูง และยังส่งผลให้ความผิดพลาดจากวิถีสะท้อนทางเสียงจริง (Misadjustment) มีค่าลดลงอีกด้วย

สมการปรับตามกาลของ VSNLMS สามารถสรุปได้ดังนี้

$$\mathbf{h}(t+1) = \mathbf{h}(t) + \frac{\alpha \mathrm{ACC(t)}}{\left(\left\|\mathbf{f}(t)\right\|^2 + \delta\right)} d(t)\mathbf{f}(t) \tag{2.82}$$

เมื่อ $\alpha \in (0,2)$ คือ ค่าสเกลช่วงก้าว และ

$$\mathrm{ACC(t)} = \frac{1}{N}\sum_{i=0}^{N-1} \mathrm{C_i}(t) \tag{2.83}$$

เมื่อ N คือจำนวนตัวอย่างที่ใช้ในการคำนวณหาค่าอัตสหสัมพันธ์ และ

$$\mathrm{C}_i(\mathrm{t}) = \frac{P_{di}(t)}{\sqrt{P_{dd}(t)P_{ii}(t)}} \tag{2.84}$$

$$P_{di}(t) = \lambda P_{di}(t-1) + (1-\lambda)d(t)f(t-i) \tag{2.85}$$

$$P_{dd}(t) = \lambda P_{dd}(t-1) + (1-\lambda)d^2(t) \tag{2.86}$$

$$P_{ii}(t) = \lambda P_{ii}(t-1) + (1-\lambda)f^2(t-i) \tag{2.87}$$

ขั้นตอนวิธีอื่นๆ

นอกจากเทคนิค Steepest-Descent แล้ว IDO ยังมีเทคนิคอื่นๆ อีกจำนวนมากซึ่ง อาศัยปริมาณที่ไม่ใช่เพียงเกร์เดียนท์ของฟังก์ชันต้นทุน อาทิเช่น เฮสเซียนเมทริกซ์ของฟังก์ชันต้นทุน เป็นต้น ในการหาทิศทางและค่าช่วงก้าวที่เหมาะสมในการปรับตัวเข้าสู่คำตอบ ตัวอย่างของขั้นตอนวิธี RSA ที่อาศัยเทคนิคต่างๆ ใน IDO ได้แก่

เทคนิค Newton-Rapson ของ IDO ถูกพัฒนาไปเป็นตระกูล Recursive Least Square (RLS) [34]
เทคนิค Quasi-Newton-Rapson ของ IDO ถูกพัฒนาไปเป็นตระกูล Affine Projection (APA) [35]
เทคนิค Conjugated Gradient [4] ของ IDO ถูกใช้อธิบายการทำงานของ Momentum LMS (MLMS) [36]
เป็นต้น

อย่างที่ได้กล่าวไว้ในบทนำ สัมประสิทธิ์ของวงจรกรองปรับตัวในระบบการสนทนาแบบแฮนด์ฟรีจำเป็นต้องมีจำนวนมาก เพื่อให้สอดรับกับวิถีสะท้อนทางเสียง ณ บริเวณที่ทำการพิจารณา ทำให้ความซับซ้อนในการคำนวณมีมากตามไปด้วย ดังนั้นจึงมีขั้นตอนวิธีจำนวนไม่น้อยที่ออกแบบมาเพื่อลดความซับซ้อนในการคำนวณลง อาทิเช่น Frequency-Domain LMS (FLMS) [38] ซึ่งอาศัย Fast Fourier Transform (FFT) ในการลดความซับซ้อนของการคอนโวลูชัน (Convolution) โดยแลกมาด้วยเวลาประวิงช่วงหนึ่ง นอกจากจะสามารถลดความซับซ้อนในการคำนวณให้กับขั้นตอนวิธีการปรับตัวแล้ว FLMS ยังสามารถใช้คุณสมบัติการตั้งฉากในตัวเอง (Self-Orthogonal) ในการเพิ่มอัตราการลู่เข้าให้กับขั้นตอนวิธีได้อีกด้วย [38] จากข้อดีของโครงสร้าง FLMS ทำให้ต่อมา Multidelay Block Frequency Domain (MDF) [39] ได้ถูกพัฒนาขึ้นเพื่อลดเวลาประวิงของขั้นตอนวิธี FLMS ลง และจากนั้น [40] ทำการปรับปรุงขั้นตอนวิธี MDF จนกระทั่งได้ขั้นตอนวิธีที่อาศัยโครงสร้างของ FLMS โดยที่ปราศจากเวลาประวิงได้

การลดความซับซ้อนในการคำนวณที่ได้รับความสนใจจากนักวิจัยจำนวนมากอีกทางหนึ่งได้แก่ การพิจารณาวงจรกรองปรับตัวในแถบความถี่ย่อย (Subband Adaptive Filter) ด้วยขั้นตอนการลดอัตราชักตัวอย่าง (Down

Sampling) ในกระบวนการวิเคราะห์แถบความถี่ย่อย (Subband Analysis) ทำให้อัตราในการปรับตัวของวงจรกรองปรับตัวลดลง และสามารถลดความซับซ้อนในการคำนวณลงได้ เมื่อใช้กับวงจรกรองที่มีจำนวนสัมประสิทธิ์ที่สูงพอ (เนื่องจากต้องคำนึงถึงความซับซ้อนของการวิเคราะห์และสังเคราะห์แถบความถี่ย่อยด้วย) ไม่เพียงลดความซับซ้อนในการคำนวณลงเท่านั้น คุณสมบัติของสัญญาณเข้าของวงจรกรองแบบปรับตัวในแต่ละแถบความถี่ย่อยยังเปลี่ยนไปอย่างวิคราะห์ได้อีกด้วย ทำให้นักวิจัยพยายามเลือกความแตกต่างดังกล่าวให้เอื้อต่อการลู่ข้าวของวงจรกรองปรับตัวในแต่ละแถบความถี่ย่อยมากที่สุด [41]

นอกจากนี้ยังมีการอาศัยเทคนิคอื่นเข้ามาช่วยในการออกแบบการทำงานเฉพาะอย่างอีกด้วยเช่น การอาศัยการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis, PCA) ของระบบที่พิจารณาและทำการปรับตัวบนโดเมนองค์ประกอบหลักนั้นๆ ด้วยขั้นตอนวิธี Proportionate NLMS [45] และเรียกเทคนิคดังกล่าวนี้ว่า PCP ซึ่งเป็นขั้นตอนวิธีที่ให้อัตราการลู่เข้าที่สูง ความซับยซ้อนในการคำนวณต่ำ แต่แลกมาด้วยปริมาณหน่วยความจำที่ต้องใช้ [37] และจำกัดงานประยุกต์ที่ใช้อยู่เพียงระบบที่นำพิจารณาหาองค์ประกอบหลักเท่านั้น

2.2.1.3. คุณสมบัติของ AEC

หากวิถีสะท้อนทางเสียงจริงมีความเป็นเชิงเส้นและมีผลตอบสนองอิมพัลส์จำกัดแล้ว จะสามารถกล่าวได้ว่าผลตอบสนองอิมพัลส์ของวงจรกรองปรับตัวที่เหมาะสมที่สุด $\mathbf{h}_{opt}$ จะมีค่าเท่ากับ ผลตอบสนองอิมพัลส์ของวิถีสะท้อนทางเสียงจริงนั่นเอง

ในความเป็นจริงแล้ววิถีสะท้อนทางเสียง อาจมีความไม่เป็นเชิงเส้น หรือ มีจำนวนสัมประสิทธิ์ไม่เท่ากับ ผลตอบสนองอิมพัลส์ของวงจรกรองปรับตัว $\mathbf{h}$ ที่จำลองเอาไว้ หรือแม้แต่อาจมีจำนวนสัมประสิทธิ์เป็นอนันต์ (Infinite Impulse Response, IIR) กล่าวคือแบบจำลองที่ทำการสมมติไว้อาจเกิดความไม่เหมาะสมขึ้นได้ ในกรณีเช่นนี้วงจรกรองปรับตัวจะไม่สามารถเลียนแบบวิถีสะท้อนทางเสียงได้ทุกประการ (แต่อย่างไรก็ตาม $\mathbf{h}_{opt}$ จะสามารถหาได้เสมอ) ผลเสียที่ได้รับจากความไม่เท่ากันระหว่าง $\mathbf{h}_{opt}$ ที่ประมาณได้กับวิถีสะท้อนทางเสียงได้แก่ เสียงสะท้อนตกค้าง (Residual Echo) กล่าวคือเมื่อนำ $\hat{e}(t)$ ซึ่งไม่สามารถประมาณให้เท่ากับ $e(t)$ ได้ ไปหักลบออกจากสัญญาณไมโครโฟน $y(t)$ จะมีสัญญาณเสียงสะท้อนบางส่วนไม่ได้ถูกหักล้างและยังคงถูกส่งกลับไปยังคู่สนทนาทางห้องไกล [33]

ในทางปฏิบัติ วิถีสะท้อนทางเสียงมีการเปลี่ยนแปลงอยู่ตลอดเวลา เนื่องด้วยสาเหตุต่างๆ เช่น ผู้คนเดินไปเดินมาภายในห้อง และประตูห้องที่ถูกเปิดเข้าเปิดออก เป็นต้น ทำให้วงจรกรองแบบปรับตัวต้องมีประสิทธิภาพการลู่เข้าสู่คำตอบที่ดีพอเพื่อสามารถติดตามการเปลี่ยนแปลงของวิถีสะท้อนทางเสียงได้ทัน

นอกจากความไม่เหมาะสมของแบบจำลองแล้ว ยังมีอีกปัจจัยหนึ่งที่ทำให้ประสิทธิภาพการทำงานของ AEC ลดลงได้ นั่นคือการที่สัญญาณไมโครโฟน $y(t)$ มีสัญญาณอื่นๆ นอกเหนือจากสัญญาณเสียงสะท้อน $e(t)$ ปะปนอยู่ด้วย [3], [52] โดยถึงแม้ว่าวิถีสะท้อนทางเสียงจะมีผลตอบสนองอิมพัลส์จำกัดซึ่งเหมาะสมกับผลตอบสนองอิมพัลส์จำกัด $\mathbf{h}$ ที่จำลองไว้ในวงจรกรองปรับตัวแล้วก็ตาม สัญญาณอื่นที่ปะปนอยู่เหล่านี้จะทำให้วงจรกรองปรับตัวไม่สามารถปรับตัวเข้าสู่ผลตอบสนองอิมพัลส์ของวิถีสะท้อนทางเสียงจริงได้ ตัวอย่างของเสียงปะปนดังกล่าว ได้แก่ เสียงรบกวนพื้นหลัง (Background Noise) ที่มีอยู่ภายในห้องใกล้ โดยเสียงรบกวนยิ่งมีพลังงานมากขึ้นเพียงใด วงจรกรองปรับตัวก็จะลู่เข้าสู่ค่าที่ห่างจากผลตอบสนองอิมพัลส์ของวิถีสะท้อนทางเสียง

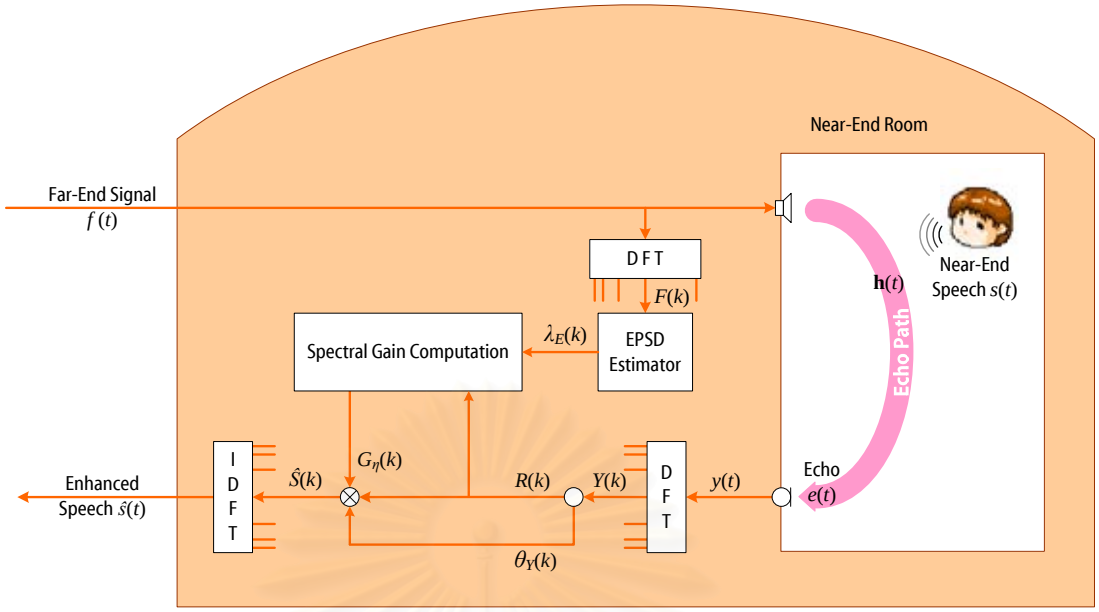
จริงมากขึ้นเทานั้นจนกระทั่งอาจเกิดการลู่ออกของวงจรกรองปรับตัวได้ ดังนั้น AEC ซึ่งทำงานภายใต้สภาพแวดล้อมที่มีเสียงรบกวนพื้นหลังจึงต้องได้รับการปรับปรุงให้สามารถทนต่อเสียงรบกวนได้ ตัวอย่างงานวิจัยเกี่ยวกับขั้นตอนวิธีการปรับตัวที่ทนทานต่อเสียงรบกวนพื้นหลังเหล่านี้ได้แก่ DUET Algorithm [42] และ VSNLMS [43] (ขั้นตอนวิธีที่ III ของหัวข้อ ขั้นตอนวิธีที่อาศัยหลักการของ Steepest-Descent ในหัวข้อย่อยที่ 2.2.1.2) เป็นต้น

นอกจากเสียงรบกวนพื้นหลังแล้ว สัญญาณเสียงที่ปะปนอีกประเภทหนึ่งซึ่งเกิดขึ้นในกรณีที่คู่สนทนาทั้งสองพูดพร้อมกันหรือที่เรียกว่า สถานการณ์ดับเบิลทอล์ก (Double-Talk Situation, DTS) โดยแตกต่างจากสถานการณ์ปกติที่พิจารณาในระบบ AEC คือ คู่สนทนาพูดทีละฝ่าย หรือที่เรียกว่าสถานการณ์ซิงเกิ้ลทอล์ก (Single-Talk Situation, STS) เนื่องจากสัญญาณเสียงดับเบิลทอล์กหรือสัญญาณเสียงของผู้พูดในห้องใกล้มีพลังงานที่สูงมากเมื่อเทียบกับสัญญาณเสียงสะท้อนจนอาจทำให้วงจรกรองปรับตัวเกิดลู่ออกจากสถานะอยู่ตัว (Steady State) และนอกจากนี้เนื่องจากสัญญาณเสียงดับเบิลทอล์กมีสหสัมพันธ์กับสัญญาณเสียงสะท้อนค่อนข้างมากทำให้ขั้นตอนวิธีปรับตัวที่ใช้รับมือกับเสียงรบกวนพื้นหลังไม่สามารถใช้ได้ใน DTS อย่างไรก็ตามเนื่องจากสถานการณ์ดับเบิ้ลทอล์กจะเกิดในช่วงสั้นๆ การแก้ไขปัญหา DTS จึงทำได้โดยการสั่งให้วงจรกรองปรับตัวทำการคงค่าวิถีสะท้อนทางเสียงเดิมที่ประมาณเอาไว้เมื่อก่อนเกิดสถานการณ์ดับเบิ้ลทอล์ก โดยอาศัยการทำงานของตัวตรวจหาสถานการณ์ดับเบิ้ลทอล์ก (Double-Talk Detector, DTD) เพื่อระบุว่าช่วงเวลาใดเกิดสถานการณ์เบิ้ลทอล์ก หลังจากนั้นจึงสามารถทำการปรับค่าสัมประสิทธิ์ของวงจรกรองปรับตัวเสียใหม่เพื่อติดตามการเปลี่ยนแปลงของวิถีสะท้อนทางเสียงที่อาจเกิดขึ้น [44]

2.2.2. การกดเสียงสะท้อน (Acoustic Echo Suppression, AES)

2.2.2.1. หลักการของ AES

การกดเสียงสะท้อน [46], [47] ดังรูปที่ 2.10 ถูกพัฒนาขึ้นโดยใช้แนวคิดของเทคนิคการกดทางสเปกตรัมซึ่งบรรยายไว้ในหัวข้อที่ 2.1.1 กล่าวคือ ในขณะที่ระบบ AEC พยายามหักล้างสัญญาณเสียงสะท้อนในระดับของรูปคลื่น (Waveform) ระบบ AES กลับทำการกดสัญญาณเสียงสะท้อนลงในโดเมน STFT ดังนั้นในขณะที่ AEC พยายามหาค่าประมาณสัญญาณเสียงสะท้อนเพื่อนำไปหักล้างกับสัญญาณเสียงสะท้อนจริง AES กลับต้องการเพียงค่าความหนาแน่นสเปกตรัมกำลังเสียงสะท้อน (Echo Power Spectral Density, EPSD) เท่านั้น และด้วยรูปแบบการทำงานในโดเมน STFT นี้เอง ทำให้ AES สามารถลดสัญญาณเสียงสะท้อนลงได้โดยไม่จำเป็นต้องทำการประมาณวิถีสะท้อนทางเสียงอย่างแม่นยำ เป็นเหตุให้ระบบ AES มีความทนทานต่อการเปลี่ยนแปลงวิถีสะท้อนทางเสียง (Echo path change) [46] และมีความซับซ้อนในการคำนวณที่ต่ำกว่า AEC มาก แต่ทั้งนี้แลกมาด้วยความผิดเพี้ยนของสัญญาณเสียงพูดทางฝั่งห้องใกล้ที่ถูกปรับปรุง



รูปที่ 2.10 ระบบการกดเสียงเสียงสะท้อน

AES ทำการประมาณสัญญาณเสียงพูดในลักษณะเดียวกับ NS กล่าวคือ ในแต่ละองค์ประกอบทางความถี่ ค่าประมาณขนาดสเปกตรัมเสียงพูด $\hat{A}(k,\ell)$ เกิดจากการคูณขนาดสเปกตรัมสัญญาณไมโครโฟน $R(k,\ell)$ ด้วย Spectral Gain $G_\eta(k,\ell)$ ก่อนที่จะนำเฟสสเปกตรัมสัญญาณไมโครโฟน $\theta_Y(k,\ell)$ มารวมเข้า กลายเป็นค่าประมาณสเปกตรัมเสียงพูด $\hat{S}(k,\ell)$ ดังรูปที่ 2.10 สังเกตได้ว่า AES ต่างจาก NS เพียงแค่สัญญาณเสียงก่อกวน ในสถานการณ์ของ AES ได้แก่ สัญญาณเสียงสะท้อน แทนที่จะเป็นสัญญาณเสียงรบกวนดังใน NS

2.2.2.2. เทคนิคการกดเสียงสะท้อน

ในหัวข้อย่อยนี้จะกล่าวถึงเทคนิคการลดเสียงสะท้อนที่จัดเป็น AES 2 เทคนิคได้แก่ Short Time Spectral Based Echo Suppression (STSES) [46] และ Perceptual Acoustic Echo Suppression (PAES) [47]

Short Time Spectral Based Echo Suppression (STSES)

[46] เสนอให้ใช้ Spectral Gain แบบหักลบทางความถี่ (Spectral Subtraction, SpS) $G_{\mathrm{SpS}}(k,\ell)$ ดังนี้

$$G_{\mathrm{SpS}}(k,\ell)=\left[1-\beta_{\mathrm{SpS}}\left(\frac{\varepsilon(k,\ell)}{\psi(k,\ell)}\right)^{\frac{\alpha_{\mathrm{SpS}}}{2}}\right]^{\frac{1}{\alpha_{\mathrm{SpS}}}} \tag{2.88}$$

เมื่อ $\varepsilon(k,\ell)$[7] คือ ค่าสเปกตรัมกำลังเสียงสะท้อน ณ ขณะนั้น

[7] ค่าสเปกตรัมกำลังเสียงสะท้อน ณ ขณะนั้น $\varepsilon(k,\ell)$ ไม่ใช่ค่าเฉลี่ยสเปกตรัมกำลังเสียงสะท้อน $\lambda_E(k,\ell)$ แต่มีความสัมพันธ์กัน คือ $\lambda_E(k,\ell)=E\{\varepsilon(k,\ell)\}$

$\psi(k,\ell) = R^2(k,\ell)$ คือ ค่าสเปกตรัมกำลังสัญญาณไมโครโฟน ณ ขณะนั้น
ณ องค์ประกอบทางความถี่ที่ k และเฟรมเวลาที่ ℓ
β_{SpS} คือ ค่าคงที่การลบแบบเกิน (Over Subtraction)
α_{SpS} คือ ค่างคงที่การกด หากมีค่ามากจะส่งผลให้กฎการลดทอนสามารถลดสัญญาณเสียงสะท้อนหรือเสียงรบกวนได้มาก โดยแลกมาซึ่งความผิดเพี้ยนของสัญญาณเสียงพูดทางห้องใกล้

จะเห็นว่า Spectral Gain แบบหักลบทางความถี่ G_{SpS} ต้องการค่าสเปกตรัมกำลังเสียงสะท้อน ณ ขณะนั้น $\varepsilon(k,\ell)$ ซึ่ง STSES เสนอให้ใช้การประมาณค่าสเปกตรัมกำลังเสียงสะท้อน ณ ขณะนั้น ด้วยขนาดกำลังสองของค่าประมาณสเปกตรัมเสียงสะท้อน $\hat{E}(k,\ell)$

$$\varepsilon(k,\ell) \approx |\hat{E}(k,\ell)|^2 \tag{2.89}$$

และประมาณค่า $\hat{E}(k,\ell)$ ดังนี้

$$\hat{E}(k,\ell) = F(k,\ell) * \hat{H}_{\mathrm{C}}(k,\ell) \tag{2.90}$$

เมื่อ $\hat{E}(k,\ell)$ คือ ค่าประมาณเชิงซ้อนของสเปกตรัมเสียงสะท้อน ณ ขณะนั้น
$F(k,\ell)$ คือ ค่าสเปกตรัมสัญญาณทางห้องไกล ณ ขณะนั้น
$\langle * \rangle$ คือ สัญลักษณ์แทนการคอนโวลูชัน
$\hat{H}_{\mathrm{C}}(k,\ell)$ [8] คือ วงจรกรองแบบปรับตัวที่มีสัมประสิทธิ์เป็นจำนวนเชิงซ้อน

เนื่องจากจำนวนของสัมประสิทธิ์ของวงจรกรองปรับตัวในแต่ละองค์ประกอบทางความถี่มีจำนวนน้อยลงมากเมื่อเปรียบเทียบกับวงจรกรองปรับตัวใน AEC ดังนั้นจึงสามารถเลือกใช้ขั้นตอนวิธีการปรับตัวแบบ RLS เพื่อให้ระบบมีอัตราการลู่เข้าที่รวดเร็วยิ่งขึ้นได้

Perceptual Acoustic Echo Suppression (PAES)

เนื่องจาก Spectral Gain แบบหักลบทางความถี่ G_{SpS} มีการแกว่งตัวสูง ทำให้ STSES ได้รับผลกระทบจาก Musical Noise [46] ดังนั้นการกดเสียงสะท้อนตามบนโดเมนซองสเปกตรัม (Perceptual Acoustic Echo Suppression, PAES) [47] จึงถูกเสนอขึ้น PAES ทำการกดสัญญาณเสียงสะท้อนบนโดเมนซองสเปกตรัม (Spectral Envelop) ซึ่งโดเมนซองสเปกตรัมดังกล่าว คือโดเมนที่พยายามเลียนแบบความละเอียดทางความถี่ (Frequency Resolution) ของการรับรู้ของหูมนุษย์ ดังนั้นจึงอาจเรียกว่า โดเมนซองสเปกตรัมทางโสต (Auditory Spectral Envelop) เมื่อย้ายมาพิจารณาในโดเมนดังกล่าว ค่าประมาณเชิงซ้อนของสเปกตรัมเสียงสะท้อน ณ

[8] ความสัมพันธ์ระหว่างวงจรกรองปรับตัวที่มีสัมประสิทธิ์เป็นจำนวนเชิงซ้อน และวิถีสะท้อนทางเสียงจริงถูกบรรยายไว้ใน [Avendano]

ขณะนั้น ณ ซองสเปกตรัมที่ κ [9] จะถูกประมาณขึ้นจากวงจรกรองแบบปรับตัวที่มีสัมประสิทธิ์เป็นจำนวนจริง ดังนี้

$$\varepsilon(\kappa,\ell) = \varphi(\kappa,\ell) * \hat{H}_{\mathrm{R}}(\kappa,\ell) \tag{2.91}$$

เมื่อ $\varphi(\kappa,\ell)$ คือ ค่าสเปกตรัมสัญญาณทางห้องไกล ณ ขณะนั้น
$\hat{H}_{\mathrm{R}}(\kappa,\ell)$ คือ วงจรกรองปรับตัวที่มีสัมประสิทธิ์เป็นจำนวนจริง
ณ ซองสเปกตรัมที่ κ และเฟรมเวลาที่ ℓ

PAES ได้รับผลกระทบจาก Musical Noise น้อยลงกว่า STSES ทั้งนี้เนื่องจากบนโดเมนซองสเปกตรัม Spectral Gain มีการแกว่งตัวที่ค่อนข้างเรียบ (Smooth) [47] และพร้อมกันกับที่ PAES ยังคงรักษาข้อดีของ STSES ไว้ คือความทนต่อการเปลี่ยนแปลงวิถีสะท้อนทางเสียง ความซับซ้อนในการคำนวณของ PAES ก็ยังลดลงเป็นอย่างมากอีกด้วย ทำให้ระบบ AES เหมาะสำหรับงานประยุกต์ที่ไม่สามารถใช้ความซับซ้อนที่สูงได้ เช่น ในเครื่องคอมพิวเตอร์ส่วนบุคคล เป็นต้น แต่อย่างไรก็ตาม ความผิดเพี้ยนของค่าประมาณเสียงพูดซึ่งสังเกตุได้ในช่วงสถานการณ์ดับเบิลทอล์กก็มีมากด้วยเช่นกัน

2.3. การลดเสียงสะท้อนและเสียงรบกวน

เนื่องจากระบบการสนทนาแบบแฮนด์ฟรีเผชิญกับปัญหาเสียงก่อกวนทั้งสองชนิดได้แก่ เสียงสะท้อนและเสียงรบกวน การเพิ่มสมรรถนะเสียงพูดในระบบดังกล่าว จึงต้องอาศัยทั้งวิธีการลดเสียงสะท้อน และวิธีการลดเสียงรบกวนทำงานควบคู่กันไป และดังที่ได้กล่าวให้เห็นไปแล้วในหัวข้อที่ผ่านมาว่า ไม่ว่าจะเป็นวิธีการลดเสียงสะท้อนก็ดี หรือวิธีการลดเสียงรบกวนก็ดี ต่างมีวิธีการที่ถูกนำเสนอเป็นจำนวนมาก ดังนั้นสิ่งแรกที่ต้องทำเมื่อต้องการออกแบบให้วิธีการลดเสียงก่อกวนทั้งสองทำงานร่วมกัน ได้แก่ วิธีการใดที่ควรถูกเลือกมาทำงานร่วมกัน เช่น เลือกพิจารณาการหักล้างเสียงสะท้อนทำงานร่วมกับการกดเสียงรบกวน หรือ จะเลือกพิจารณาการกดเสียงสะท้อนทำงานร่วมกับการกดเสียงรบกวน หรือ เลือกพิจารณาการลดเสียงสะท้อนและการลดเสียงรบกวนที่ขึ้นกับเทคนิคการกดบนโดเมนเวปเล็ต เป็นต้น

สำหรับในวิทยานิพนธ์ฉบับนี้นำเสนอวิธีการเพิ่มประสิทธิภาพของระบบที่เลือกใช้การกดเสียงสะท้อนและการกดเสียงรบกวน หรือที่เรียกว่า “การกดเสียงสะท้อนและเสียงรบกวน” (Acoustic Echo and Noise Suppression, AENS) กล่าวคือ เป็นการลดเสียงสะท้อนและเสียงรบกวนที่ขึ้นกับเทคนิคการกดทางสเปกตรัม (Acoustic Echo and Noise Reduction based on Spectral Suppression Technique) นั่นเอง

อย่างไรก็ตามในหัวข้อนี้จะทำการกล่าวถึงความเป็นมาอย่างสังเขป ของการออกแบบการลดเสียงสะท้อนและเสียงรบกวน (Acoustic Echo and Noise Reduction, AENR) เพื่อความครบถ้วนและการสามารถเปรียบเทียบได้

---

[9] κ ถูกใช้เป็นตัวบ่งชี้องค์ประกอบทางซองสเปกตรัม ในขณะที่ k ถูกใช้เป็นตัวบ่งชี้องค์ประกอบทางความถี่

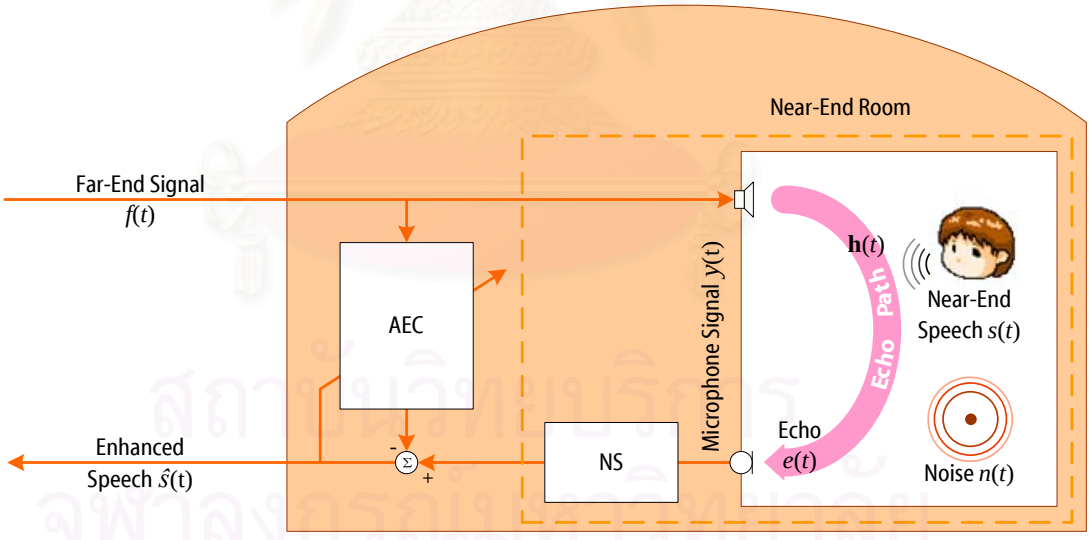
ในช่วงต้นของการออกแบบนั้น นักวิจัยได้มุ่งความสนใจไปที่ AEC ทำงานร่วมกับ NS เนื่องจากทั้งสองวิธีให้ประสิทธิภาพการทำงนที่ดีและความซับซ้อนในการคำนวณที่เหมาะสมในการแก้แต่ละปัญหาอย่างแยกจากกันมากที่สุด วิธีการที่เรียบง่ายที่สุดในการนำทั้งสองกระบวนการมาทำงานร่วมกันได้แก่ การนำมาต่อเรียงกัน (Cascading) โดยหาก AEC สามารถทำการหักล้างเสียงรบกวนได้หมด และ NS สามารถกดเสียงรบกวนได้อย่างมีประสิทธิภาพแล้ว ระบบร่วม (Combined System) ดังกล่าวก็จะสามารถทำงานได้เป็นอย่างเป็นที่พอใจ

การออกแบบระบบร่วม AEC และ NS ดังกล่าวสามารถแบ่งออกได้เป็น 3 แนวทางใหญ่ได้แก่

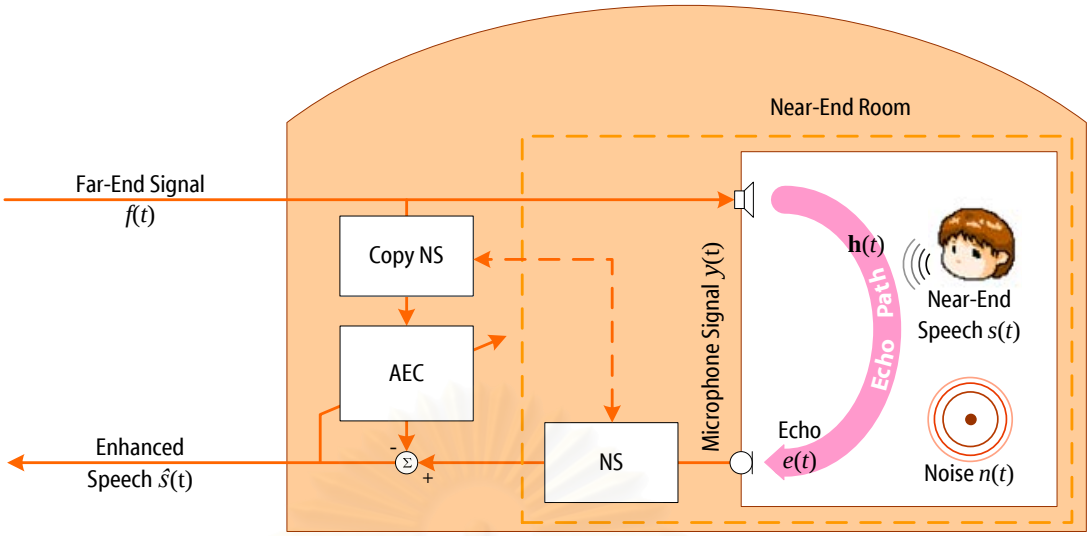
2.3.1. NS ต่อด้วย AEC (NSAEC)

ในสถานการณ์จริง ประสิทธิภาพของ AEC จะลดลงอย่างมากเมื่อทำงานภายใต้สภาพแวดล้อมที่ระดับเสียงรบกวนพื้นหลังมีค่าสูง ในขณะที่ประสิทธิภาพของ NS ไม่ลดลงมากนักเมื่อทำงานภายใต้สภาพแวดล้อมที่มีสัญญาณเสียงสะท้อน ดังนั้นจึงควรลดเสียงรบกวนลงก่อนทำการลดเสียงสะท้อนด้วย AEC ดังบล็อกไดอะแกรมในรูปที่ 2.11

อย่างไรก็ตาม ถึงแม้ว่าสัญญาณที่ออกมาจากวงจรกรองของ NS จะมีระดับสัญญาณเสียงรบกวนที่ต่ำลง แต่ปัญหาที่ตามมาคือ วิถีสะท้อนทางเสียงซึ่งวงกรองปรับตัวใน AEC ต้องทำการประมาณนั้น ได้รบผลจากความไม่เป็นเชิงเส้นของ NS ด้วย ทำให้ AEC ไม่สามารถเลียนแบบวิถีสะท้อนทางเสียงดังกล่าวได้อย่างถูกต้อง จึงมีผู้สนอให้นำวงจรกรองที่เหมือนกับวงจรกรองของ NS ไปต่อเรียงไว้ก่อนหน้า วงจรกรองใน AEC เพื่อลดผลของความไม่เป็นเชิงเส้นดังกล่าวลง ดังรูปที่ 2.12



รูปที่ 2.11 NSAEC



รูปที่ 2.12 MNSAEC

จะเห็นว่า NSAEC เป็นการนำระบบ NS มาต่อเรียงกับระบบ AEC ดังนั้นความซับซ้อนในการคำนวณของระบบร่วมจะเพิ่มขึ้นและยังคงขึ้นอยู่กับจำนวนสัมประสิทธิ์ของวงจรกรองแบบปรับตัวของระบบ AEC ที่ต้องมากเพียงพอสำหรับการประมาณเสียงสะท้อนทางเสียง

ถึงแม้ว่ารูปแบบการเรียงตัวตาม NSAEC จะสมเหตุสมผลในเชิงความคิดทั่วไป แต่ยังไม่มีคำยืนยันทางคณิตศาสตร์ที่บ่งชี้ว่า NS เป็นวิธีการที่ควรถูกใช้ก่อน AEC ดังนั้น [49] จึงพยายามทำการวิเคราะห์ปัญหาลำดับวางตัวของ NS และ AEC โดยการหาคำตอบที่เหมาะสม (Optimal Solution) ในเชิงค่าเฉลี่ยของค่าผิดพลาดกำลังสองน้อยสุด (MMSE) ระหว่างสัญญาณเสียงพูดทางห้องใกล้และค่าประมาณสัญญาณเสียงพูดทางห้องใกล้ บนสมมติฐานที่ว่าสัญญาณที่พิจารณาในระบบทุกสัญญาณมีความเป็นจุดนิ่ง และสัญญาณเสียงพูดทางฝั่งห้องใกล้ $s(t)$ ถูกประมาณได้จากการคอนโวลูชันของวงจรกรองเชิงเส้นกับสัญญาณไมโครโฟน $y(t)$ และสัญญาณเสียงทางฝั่งห้องไกล $f(t)$ ดังนี้

$$\hat{s}(t) = w_1(t) * y(t) + w_2(t) * f(t) \tag{2.92}$$

โดยที่ $w_1(t)$ และ $w_2(t)$ เป็นวงจรกรองเชิงเส้นที่มีจำนวนสัมประสิทธิ์เป็นอนันต์ (IIR) หรือ มีจำนวนสัมประสิทธิ์จำกัด (FIR) ก็ได้

คำตอบที่ได้หลังจากการหาคำตอบที่เหมาะสมของค่าเฉลี่ยของค่าผิดพลาดกำลังสองน้อยสุด $E\{|s(t)-\hat{s}(t)|^2\}$ มีความสอดคล้องกับระบบรวมที่มีโครงสร้าง AEC ต่อ ด้วย NS ดังนี้

$$\hat{s}(t) = \left(\Im^{-1}\left\{ \frac{R_{ss}(k,\ell)}{R_{yy}(k,\ell) - R_{yf}(k,\ell) R_{ff}^{-1}(k,\ell) R_{fy}(k,\ell)} \right\} \right) * \left(y(t) + f(t) * \Im^{-1}\left\{ \frac{R_{fy}(k,\ell)}{R_{ff}(k,\ell)} \right\} \right) \tag{2.93}$$

เมื่อ $\Im^{-1}\{\cdot\}$ แทนการแปลงฟูริเยร์ผกผันแบบไม่ต่อเนื่อง (Inverse Discrete Fourier Transform)

$R_{ff}(k,\ell)$ เป็นความหนาแน่นสเปกตรัมกำลังข้าม (Cross Power Spectral Density) ระหว่างสัญญาณเสียงทางฝั่งห้องไกล $f(t)$ และสัญญาณเสียงทางฝั่งห้องไกล $f(t)$

$R_{fy}(k,\ell)$ เป็นความหนาแน่นสเปกตรัมกำลังข้ามระหว่างสัญญาณเสียงทางฝั่งห้องไกล $f(t)$ และสัญญาณไมโครโฟน $y(t)$

$R_{yf}(k,\ell)$ เป็นความหนาแน่นสเปกตรัมกำลังข้ามระหว่างสัญญาณเสียงทางฝั่งห้องไกล $y(t)$ และสัญญาณไมโครโฟน $f(t)$

$R_{yy}(k,\ell)$ เป็นความหนาแน่นสเปกตรัมกำลังข้ามระหว่างสัญญาณไมโครโฟน $y(t)$ และสัญญาณไมโครโฟน $y(t)$

$R_{ss}(k,\ell)$ เป็นความหนาแน่นสเปกตรัมกำลังข้ามระหว่างสัญญาณสัญญาณเสียงพูดทางฝั่งห้องใกล้ $s(t)$ และสัญญาณเสียงพูดทางฝั่งห้องใกล้ $s(t)$

จะเห็นได้อย่างชัดเจนว่าคำตอบที่เหมาะสมในสมการที่ (2.93) สามารถแบ่งการทำงานออกได้เป็น 2 ส่วน ได้แก่ ส่วนแรกคือส่วนของการตัดสัญญาณเสียงสะท้อน โดยมีผลตอบเชิงความถี่ของระบบเป็น

$$C(k,\ell) = \frac{R_{fy}(k,\ell)}{R_{ff}(k,\ell)} \tag{2.94}$$

ซึ่งตรงกับผลตอบเชิงความถี่ของ $\mathbf{h}_{opt}$ ในสมการที่ (2.72) นั่นเอง และส่วนที่สองคือส่วนที่ทำหน้าที่ลดสัญญาณเสียงสะท้อนตกค้างและสัญญาณเสียงรบกวน โดยมีผลตอบเชิงความถี่เป็น

$$H(k,\ell) = W_1(k,\ell) = \frac{R_{ss}(k,\ell)}{R_{yy}(k,\ell) - R_{yf}(k,\ell)R_{ff}^{-1}(k,\ell)R_{fy}(k,\ell)} \tag{2.95}$$

ความหมายของสมการที่ (2.95) นำมาซึ่งแนวทางการพัฒนาระบบรวมอีก 2 แนวทางดังต่อไปนี้

2.3.2. AEC_RF[10]F ต่อด้วย NS (AECNS)

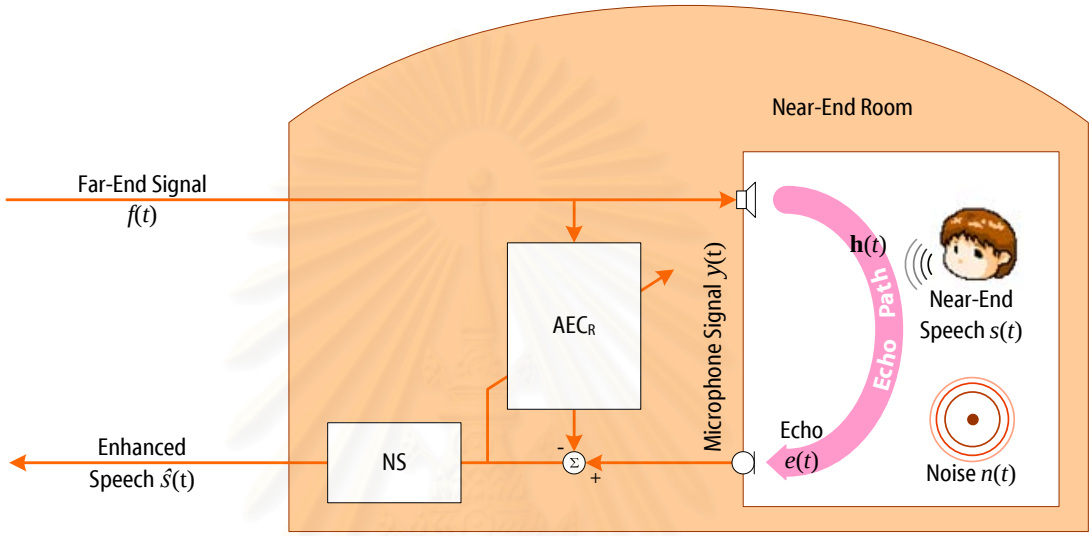
หากมองว่าระบบ AEC ซึ่งถูกวางไว้ก่อนหน้าระบบ NS สามารถทำงานได้อย่างมีประสิทธิภาพ กล่าวคือ สัญญาณสะท้อนถูกกำจัดไป ทำให้สัญญาณไมโครโฟนและสัญญาณลำโพงที่มาจากห้องใกล้ไม่มีสหสัมพันธ์กันแล้ว $H(k,\ell)$ ในสมการที่ (2.95) จะลดรูปลงเหลือเพียง

$$H(k,\ell) = W_1(k,\ell) = \frac{R_{ss}(k,\ell)}{R_{yy}(k,\ell)} = \frac{R_{ss}(k,\ell)}{R_{ss}(k,\ell) + R_{nn}(k,\ell)} \tag{2.96}$$

ซึ่งตรงกับผลเฉลยวีเนอร์ (Wiener gain) ในสมการที่ (2.21) สำหรับระบบ NS เมื่อ $R_{ss}(k,\ell) = \lambda_S(k,\ell)$ และ $R_{nn}(k,\ell) = \lambda_N(k,\ell)$ นั่นเอง กล่าวคือหากระบบ AEC สามารถทำงานได้อย่างมีประสิทธิภาพแล้ว ส่วนที่ทำหน้าที่ลดสัญญาณเสียงสะท้อนตกค้างและสัญญาณเสียงรบกวนตามสมการ (2.95) จะสามารถถูกแทนที่ด้วยระบบ

[10] AEC_R หมายถึง ระบบ AEC ที่ได้รับการปรับปรุงให้มีความทนทานต่อสัญญาณเสียงรบกวน

NS ได้ดังรูปที่ 2.13 เนื่องจากข้อสังเกตนี้เองทำให้การพัฒนาระบบรวม ในแนวทางที่ 2 มุ่งเน้นไปที่การเพิ่มประสิทธิภาพการทำงานของ AEC ให้สามารถทำงานภายใต้สภาพแวดล้อมที่มีสัญญาณรบกวนได้ โดยในที่นี้จะเรียกว่า AEC ที่ถูกปรับปรุงให้สามารถทำงานภายใต้สภาพแวดล้อมที่มีสัญญาณรบกวนว่า AEC_R ตัวอย่างของ AEC_R ไดแก่ VNLMS และ M&NLMS เป็นต้น ความซับซ้อนในการคำนวณของระบบรวมในแนวทางที่ 2 มีค่าสูงเนื่องจากยังคงขึ้นอยู่กับจำนวนสัมประสิทธิ์ของวงจรกรองปรับตัวของ AEC_R ที่สูงอยู่ (AEC_R มีความซับซ้อนในการคำนวณที่สูงกว่า AEC ธรรมดา)

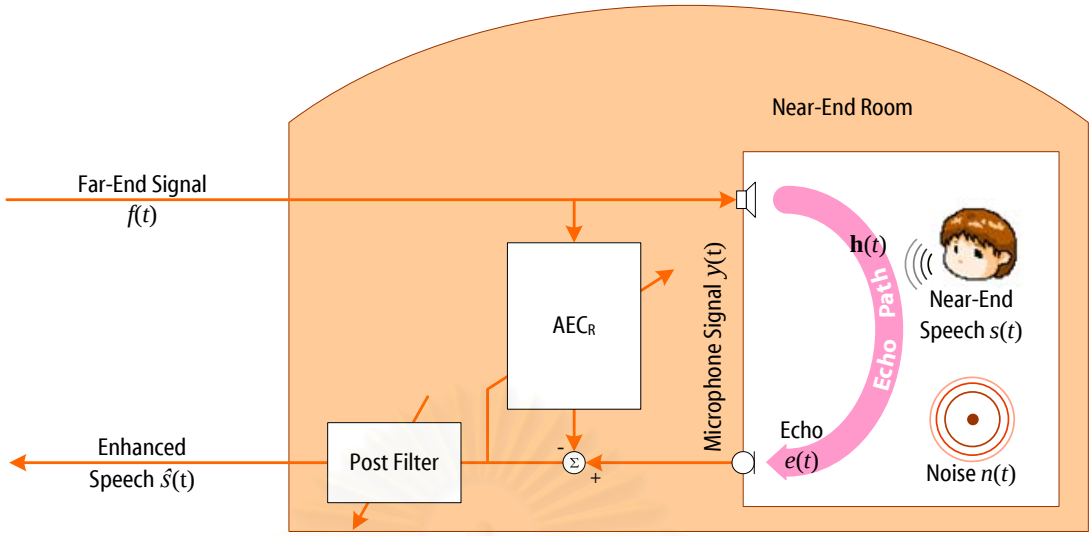


รูปที่ 2.13 AECNS

2.3.3. AEC_R ต่อด้วยวงจรกรองคัดท้าย (AEC_R with Post-Filtering, AECP)

จากสมการที่ (2.95) จะเห็นได้ว่าประสิทธิภาพสูงสุดของ AEC_R ในการตัดสัญญาณเสียงสะท้อนภายใต้สภาวะแวดล้อมที่มีสัญญาณรบกวนน จำเป็นต้องอาศัยวงจรกรองอีกตัวหนึ่งเพื่อทำการกรองทั้งสัญญาณเสียงสะท้อนตกค้าง และสัญญาณเสียงรบกวนออกจากสัญญาณออกของ AEC_R โดยวงจรกรองตัวที่เพิ่มขึ้นมานี้ถูกพัฒนามาจาก NS ซึ่งจะเรียกว่า วงจรกรองคัดท้าย (Post Filter) บล็อกไดอะแกรมของระบบรวมแนวทางที่ 3 นี้ถูกแสดงไว้ในรูปที่ 2.14

เนื่องจากวงจรกรองคัดท้ายสามารถลดได้ทั้งสัญญาณเสียงสะท้อนตกค้างและสัญญาณเสียงรบกวน จึงสามารถลดจำนวนสัมประสิทธิ์ของวงจรกรองแบบปรับตัวในส่วนของ AEC_R ลงได้ ทำให้อัตราการลู่เข้าของระบบ AEC_R เร็วขึ้น และมีความทนทานต่อสัญญาณรบกวนที่สูงขึ้น [50] งานวิจัยระบบรวมตามแนวทางที่ 3 จึงมุ่งเน้นไปที่การพัฒนาวงจรกรองคัดท้ายนี้เพื่อให้สามารถกำจัดได้ทั้งสัญญาณเสียงรบกวนและสัญญาณเสียงสะท้อนตกค้างอย่างมีประสิทธิภาพ ตัวอย่างของแนวทางนี้ได้แก่ [49] และ [50] สำหรับความซับซ้อนในการคำนวณของระบบรวมในแนวทางนี้ จะมีแนวโน้มลดลงจาก 2 แนวทางแรก



รูปที่ 2.14 AEC_R with Post-Filtering

2.3.4. การกดเสียงสะท้อนและเสียงรบกวน (AENS)

AENS ถูกเสนอขึ้นครั้งแรกโดย [53] ในขณะที่ทั้ง 3 แนวทางที่กล่าวมาแล้วอาศัยหลักการทำงานของวงจรกรอง 2 ตัว กล่าวคือ $w_1(t)$ และ $w_2(t)$ ดังสมการที่ (2.92) ในการหาค่าประมาณสัญญาณเสียงพูด แนวทางที่ 4 (AENS) นี้เริ่มจากการอาศัยวงจรกรองเพียงหนึ่งตัวในการหาค่าประมาณสัญญาณเสียงพูดดังนี้ (รูปที่ 2.15)

$$\hat{s}(t) = g(t) * y(t) \tag{2.97}$$

เมื่อ $g(t)$ เป็นวงจรกรองเชิงเส้นใดๆ เช่นเดียวกับ $w_1(t)$ และ $w_2(t)$ ทำให้สมการที่ (2.97) สามารถเขียนในโดเมนความถี่ได้ดังนี้

$$\hat{S}(k,\ell) = G(k,\ell)Y(k,\ell) \tag{2.98}$$

เมื่อทำการหาค่าที่เหมาะสมที่สุดจากการหา MMSE

$$\hat{S}(k,\ell) = \arg\min_{\hat{S}(k,\ell)} E\{|S(k,\ell) - \hat{S}(k,\ell)|^2 | Y(k,\ell)\} = \arg\min_{G(k,\ell)} E\{|S(k,\ell) - G(k,\ell)Y(k,\ell)|^2\} \tag{2.99}$$

จะได้ว่า

$$\hat{S}(k,\ell) = G(k,\ell)Y(k,\ell) = \frac{\lambda_S(k,\ell)}{\lambda_S(k,\ell) + \lambda_D(k,\ell)} Y(k,\ell) = \frac{\xi^D(k,\ell)}{\xi^D(k,\ell) + 1} Y(k,\ell) \tag{2.100}$$

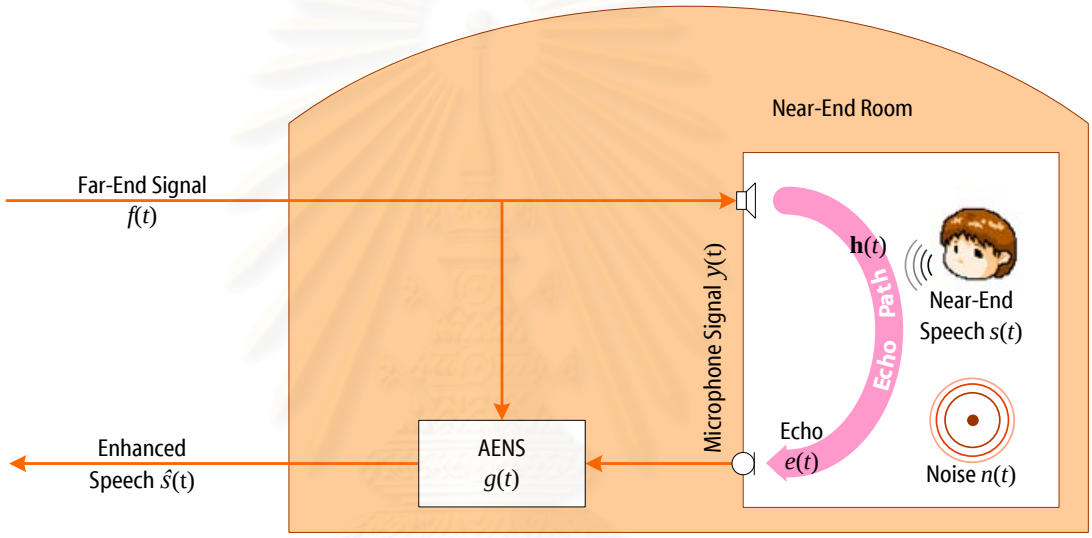
เมื่อ

$$\xi^D(k,\ell) = \frac{\lambda_S(k,\ell)}{\lambda_D(k,\ell)} \tag{2.101}$$

คืออัตราสัญญาณต่อสัญญาณก่อกวนก่อนประสบ (a priori Signal to Disturbance Ratio, a priori SDR ξ^D) ณ องค์ประกอบทางความถี่ที่ k และเฟรมเวลาที่ ℓ และ $\lambda_D(k,\ell)$ คือค่าความแปรปรวนสเปกตรัมเสียงก่อกวน โดยในที่นี้ได้แก่

$$\lambda_D(k,\ell) = E\{|N(k,\ell)|^2\} + E\{|E(k,\ell)|^2\} = \lambda_N(k,\ell) + \lambda_E(k,\ell) \tag{2.102}$$

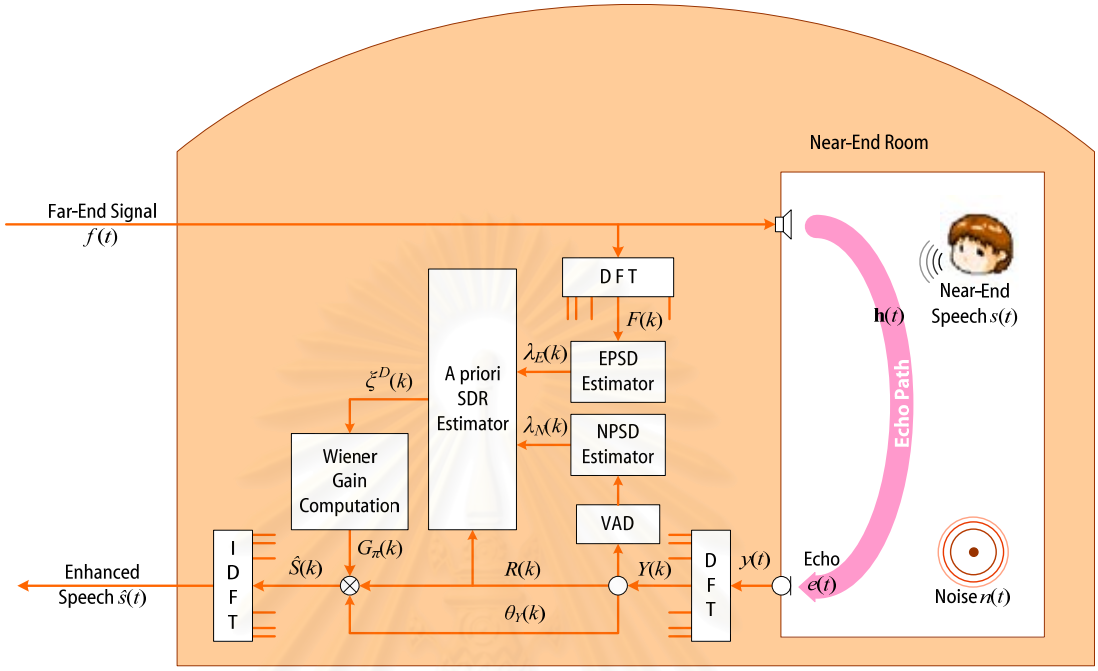
เมื่อ $\lambda_E(k,\ell) = E\{|E(k,\ell)|^2\}$ คือค่าความแปรปรวนสเปกตรัมเสียงสะท้อน



รูปที่ 2.15 AENS

2.3.4.1. การนำการกดเสียงสะท้อนและเสียงรบกวนไปปฏิบัติ

การกดเสียงสะท้อนและเสียงรบกวนใน [53] มีลักษณะการทำงานดังสรุปไว้ในรูปที่ 2.16



รูปที่ 2.16 AENS ใน [53]

การประมาณค่า NPSD $\lambda_N(k,\ell)$ สามารถกระทำได้ในช่วงที่ไม่มีทั้งเสียงพูดและเสียงสะท้อน โดยใช้สมการที่ (2.34) ส่วนการประมาณค่า EPSD $\lambda_E(k,\ell)$ สามารถกระทำได้โดยการหาค่าอัตราส่วนระหว่างความหนาแน่นสเปกตรัมกำลังข้าม (Cross Power Spectral Density) ของสัญญาณเสียงพูดทางห้องไกล $f(t)$ และสัญญาณไมโครโฟน $y(t)$ กับความหนาแน่นสเปกตรัมกำลังของสัญญาทางฝั่งห้องไกล $f(t)$ ดังนี้

$$\lambda_E(k,\ell) = \frac{|\lambda_{fy}(k,\ell)|^2}{\lambda_{ff}(k,\ell)} \tag{2.103}$$

เมื่อ $\lambda_{fy}(k,\ell)$ คือความหนาแน่นสเปกตรัมกำลังข้ามของ $f(t)$ และ $y(t)$

$\lambda_{ff}(k,\ell)$ คือความหนาแน่นสเปกตรัมกำลังของ $f(t)$

โดยค่า $\lambda_{ff}(k,\ell)$ และ $\lambda_{ff}(k,\ell)$ ถูกประมาณได้โดยใช้วงจรกรองแบบอิมพัลส์ไม่จำกัดอันดับหนึ่ง และเรียกวิธีการประมาณ EPSD นี้ว่า Coherence Function Method (CFM)

ค่า a priori SDR ถูกนิยามใหม่เพื่อความเหมาะสมในการหาค่าประมาณดังนี้

$$\xi^D(k,\ell) = \frac{1}{1/\xi^E(k,\ell) + 1/\xi^N(k,\ell)} \tag{2.104}$$

เมื่อ

$$\xi^E(k,\ell) = \frac{\lambda_S(k,\ell)}{\lambda_E(k,\ell)} \tag{2.105}$$

คือ ค่าอัตราสัญญาณต่อสัญญาณเสียงสะท้อนก่อนประสบ (a priori Signal to Echo Ratio, a priori SER) และ

$$\xi^N(k,\ell) = \frac{\lambda_S(k,\ell)}{\lambda_N(k,\ell)} \tag{2.106}$$

คือ ค่าอัตราสัญญาณต่อสัญญาณเสียงรบกวนก่อนประสบ (a priori Signal to Noise Ratio, a priori SNR)

โดย a priori SER และ a priori SNR จะถูกทำการประมาณโดยอาศัย DD อย่างแยกจากกันดังนี้

$$\hat{\xi}^E(k,\ell) = \alpha_{\mathrm{DD}}\,\tilde{\xi}^E(k,\ell-1) + (1-\alpha_{\mathrm{DD}})\,\delta^E(k,\ell) \tag{2.107}$$

และ

$$\hat{\xi}^N(k,\ell) = \alpha_{\mathrm{DD}}\,\tilde{\xi}^N(k,\ell-1) + (1-\alpha_{\mathrm{DD}})\,\delta^N(k,\ell) \tag{2.108}$$

เมื่อ

$$\tilde{\xi}^E(k,\ell-1) = \frac{\hat{A}^2(k,\ell-1)}{\lambda_E(k,\ell-1)} \tag{2.109}$$

$$\tilde{\xi}^N(k,\ell-1) = \frac{\hat{A}^2(k,\ell-1)}{\lambda_N(k,\ell-1)} \tag{2.110}$$

$$\delta^E(k,\ell) = \max[\gamma^E(k,\ell)-1, 0] \tag{2.111}$$

$$\delta^N(k,\ell) = \max[\gamma^N(k,\ell)-1, 0] \tag{2.112}$$

โดย

$$\gamma^E(k,\ell) = \frac{|Y(k,\ell)|^2}{\lambda_E(k,\ell)} \tag{2.113}$$

$$\gamma^N(k,\ell) = \frac{|Y(k,\ell)|^2}{\lambda_N(k,\ell)} \tag{2.114}$$

จากนั้นค่าประมาณ a priori SDR จึงสามารถประมาณได้ดังนี้

$$\hat{\xi}^{D}(k,\ell)=\frac{1}{1/\hat{\xi}^{E}(k,\ell)+1/\hat{\xi}^{N}(k,\ell)} \tag{2.115}$$

ค่าประมาณสเปกตรัมเสียงพูดสามารถหาได้จากสมการที่ (2.100) โดยใช้ $\hat{\xi}^{D}(k,\ell)$ แทน $\xi^{D}(k,\ell)$ จากนั้นค่าประมาณสเปกตรัมเสียงพูดจะถูกแปลงกลับไปสู่โดเมนเวลาผ่านทางวิธีการ Overlap-Add ต่อไป

หน้าที่การทำงานต่างๆ ของ AENS ที่ถูกนำเสนอใน [53] ถูกสรุปไว้ในตารางที่ 2.1

ตารางที่ 2.1 สรุปหน้าที่การทำงนต่างๆ ของ AENS ใน [53]

กระบวนการทำงานต่างๆ ของ AENS ใน [53]	
การประมาณค่า NPSD	ประมาณในช่วงที่ไม่มีเสียงพูด
การประมาณค่า EPSD	CFM
กาสรประมาณค่า a priori SDR	Decision Direct
Spectral Gain G_{η}	Wiener Gain

ความซับซ้อนในการคำนวณของ AENS ใน [53] มีค่าต่ำกว่าในสามแนวทางแรกเป็นอย่างมาก อย่างไรก็ตามจากผลการทดลองและการวิเคราะห์ของผู้พัฒนาเองชี้ว่า AENS ทำให้เกิดการผิดเพี้ยนของสัญญาณเสียงพูดทางห้องใกล้ที่มากกว่าทั้ง 3 แนวทางแรก โดยเฉพาะอย่างยิ่งในช่วงสถานการณ์ดับเบิลทอล์ก

2.3.4.2. รูปแบบทั่วไปของ AENS

AENS ใน [53] ถูกพัฒนาขึ้นจากความต้องการที่จะใช้วงจรกรองเพียงตัวเดียวในการลดเสียงสะท้อนและเสียงรบกวน อย่างไรก็ตาม AENS สามารถขยายมุมมองออกได้ว่าเป็นการลดเสียงก่อกวนใดๆ โดยอาศัยเทคนิคการกดทางสเปกตรัม กล่าวคือ เป็นการมองปัญหาการลดเสียงก่อกวนใดๆ บนโดเมนความถี่ โดยสเปกตรัมเสียงพูดที่ถูกก่อกวน $Y(k,\ell)$ เกิดจากการบวกกันระหว่าง สเปกตรัมเสียงพูด $S(k,\ell)$ และสเปกตรัมเสียงก่อกวน $D(k,\ell)$ ดังนี้

$$Y(k,\ell)=S(k,\ell)+D(k,\ell) \tag{2.116}$$

โดยที่ สเปกตรัมเสียงก่อกวน $D(k,\ell)$ เกิดจากเสียงก่อกวนชนิดต่างๆ รวมกัน โดยสำหรับในวิทยานิพนธ์ฉบับนี้คือ การรวมกันของสเปกตรัมเสียงสะท้อน $E(k,\ell)$ และสเปกตรัมเสียงรบกวน $N(k,\ell)$ ดังนี้

$$D(k,\ell)=E(k,\ell)+N(k,\ell) \tag{2.117}$$

ค่าประมาณขนาดสเปกตรัมเสียงพูด $\hat{A}(k,\ell)$ สามารถหาได้จากผลคูณระหว่าง Spectral Gain $G_{\pi}(k,\ell)$ และขนาดสเปกตรัมเสียงพูดที่ถูกก่อกวน $R(k,\ell)$

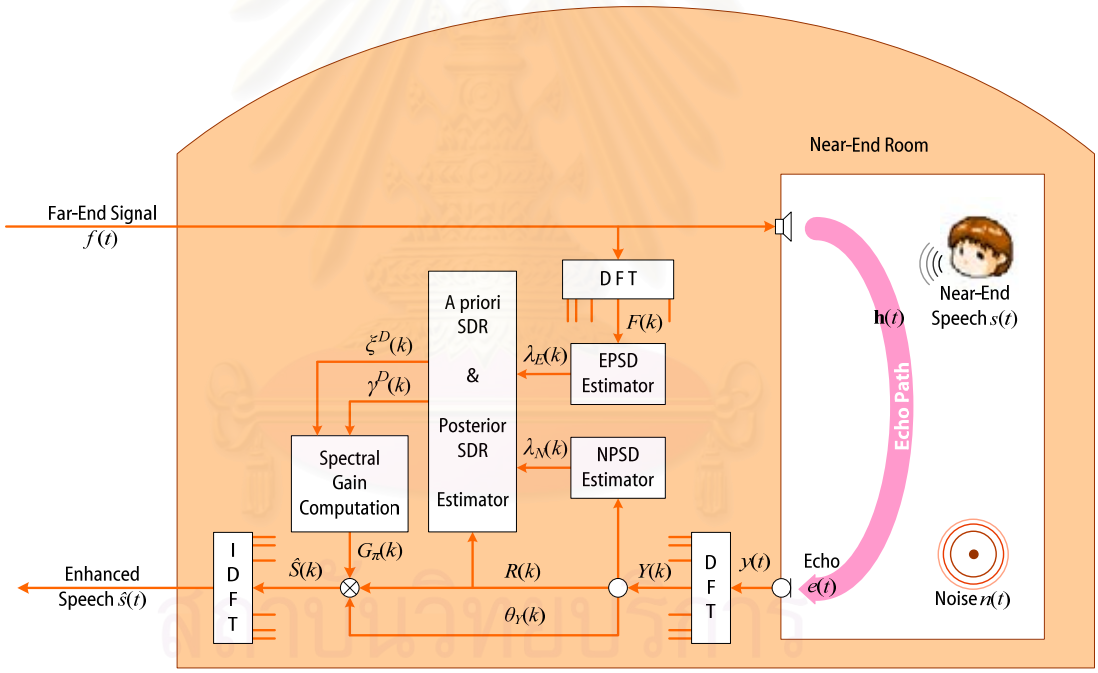
$$\hat{A}(k,\ell)=G_{\pi}(k,\ell)R(k,\ell) \tag{2.118}$$

และจะถูกนำมารวมกับเฟสสเปกตรัมเสียงพูดที่ถูกก่อกวน $\theta_Y(k,\ell)$ กลายเป็นค่าประมาณสเปกตรัมเสียงพูด หรือสเปกตรัมเสียงพูดที่ถูกปรับปรุง $\hat{S}(k,\ell)$ ก่อนจะถูกแปลงกลับไปสู่โดเมนเวลาผ่านทางวิธีการ Overlap Add ต่อไป

ซึ่ง Spectral Gain $G_\pi(k,\ell)$ สามารถหามาได้จากความผิดเพี้ยนชนิดต่างๆ ดังที่เคยได้กล่าวไว้แล้วในหัวข้อที่ 2.1.1.1 และเช่นกัน Spectral Gain เหล่านี้จะติดค่าตัวแปรสำคัญสองตัวได้แก่ ค่า a priori SDR ซึ่งถูกนิยามตามสมการที่ (2.101) และค่าอัตราสัญญาณต่อสัญญาณก่อกวนหลังประสบ (Posterior SDR) ซึ่งนิยามดังนี้

$$\gamma^D(k,\ell) = \frac{|Y(k,\ell)|^2}{\lambda_D(k,\ell)} \tag{2.119}$$

หากทำการเลือกใช้ Spectral Gain G_{SE} ที่ได้จากการใช้ความผิดเพี้ยนแบบ d_{SE} จะทำให้ได้วิธีการ AENS ดังที่เสนอใน [53] นั่นเอง และวิธีการประมาณ a priori SDR และ Posterior SDR จึงเป็นสิ่งที่สำคัญสำหรับวิธีการ AENS เช่นเดียวกัน รูปแบบทั่วไปของ AENS ถูกแสดงไว้ใน



รูปที่ 2.17 รูปแบบทั่วไปของ AENS

วิทยานิพนธ์ฉบับนี้จะอาศัยรูปแบบทั่วไปของ AENS ในการวิเคราะห์และแก้ไขปัญหาการลดเสียงสะท้อนและเสียงรบกวนต่อไป

บทที่ 3

การพัฒนาประสิทธิภาพของการกดเสียงสะท้อนและเสียงรบกวน

จากบทที่ 2 จะเห็นว่า AENS เป็นวิธีการลดเสียงสะท้อนและเสียงรบกวนซึ่งมีความซับซ้อนในการคำนวณต่ำที่สุด ดังนั้นจึงมีความเหมาะสมกับงานประยุกต์ที่ไม่สามารถรองรับวิธีการที่มีความซับซ้อนในการคำนวณสูงได้ เช่น การสนทนาแบบแฮนด์ฟรีซึ่งดำเนินควบคู่ไปกับกิจกรรมอื่นๆ สำหรับเครื่องคอมพิวเตอร์ส่วนบุคคล และการสนทนาแบบแฮนด์ฟรีสำหรับโทรศัพท์ติดรถยนต์ เป็นต้น

จากผลการทดสอบใน [53] ชี้ให้เห็นว่าความสามารถในการลดเสียงสะท้อน (Echo Attenuation, EA) ของ AENS อยู่ในระดับปานกลาง ในขณะที่คุณภาพของเสียงพูดที่ถูกปรับปรุงด้วย AENS มีความผิดเพี้ยนค่อนข้างมากโดยเฉพาะอย่างยิ่งในช่วง DTS เมื่อเปรียบเทียบกับ ระบบร่วมอื่นๆ ได้แก่ NSAEC AECNS และ AECP ดังแสดงในตารางที่ 3.1

ตารางที่ 3.1 เปรียบการทำงานโดยสังเขปของ AENR ทั้ง 4 แนวทาง [51] ในด้านการลดเสียงรบกวน (Noise Attenuation, NA) การลดเสียงสะท้อน (Echo Attenuation, EA) ความผิดเพี้ยนของสัญญาณเสียงพูด (Speech Distortion) และความซับซ้อนในการคำนวณ (Computational Complexity)

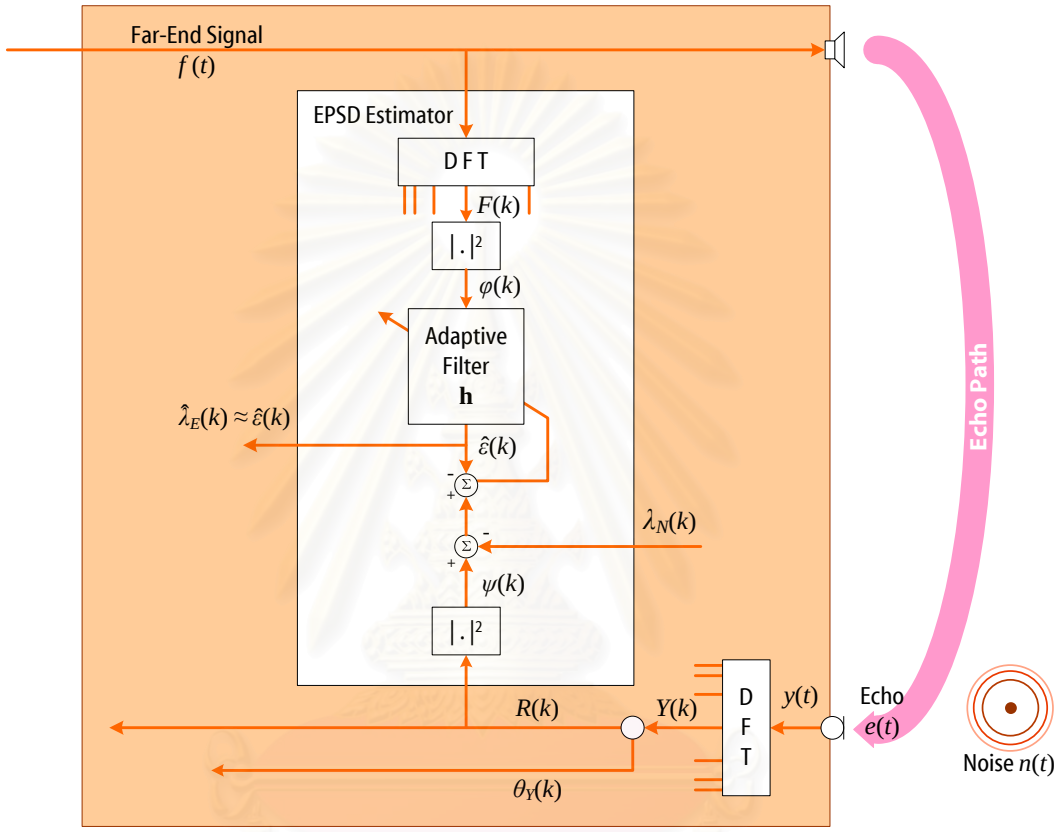
	NSAEC [52]		AECNS [52]		AECP [49]-[50]		AENS [53]	
	STS	DTS	STS	DTS	STS	DTS	STS	DTS
การลดลงของเสียงรบกวน	ดี	ดี	ดี	ดี	ดี	ดี	ดี	ดี
การลดลงของเสียงสะท้อน	ดี	ดี	ดี	ดี	ดี	ดี	ปานกลาง	ปานกลาง
ความผิดเพี้ยนของสัญญาณเสียงพูด	น้อย	น้อย	น้อย	น้อย	น้อย	ปานกลาง	น้อย	สูง
ความซับซ้อนในการคำนวณ	สูง		สูง		ปานกลาง		ต่ำ	

ดังนั้นจากข้อดีและข้อเสียของ AENS ข้างต้น วิทยานิพนธ์ฉบับนี้จึงเลือกที่จะทำการศึกษาและพัฒนา AENS โดยยังคงข้อดี อันได้แก่ ความซับซ้อนในการคำนวณที่ต่ำ เอาไว้ และทำการเพิ่มประสิทธิภาพในด้านการลดเสียงสะท้อน และความผิดเพี้ยนของสัญญาณเสียงพูด ให้มีมากยิ่งขึ้น โดยนำเสนอการเพิ่มประสิทธิภาพให้กับ 2 กระบวนการดังต่อไปนี้

1. เพิ่มความสามารถในการกดเสียงสะท้อนให้กับ AENS โดยนำเสนอการประมาณค่า EPSD ที่มีความแม่นยำเพิ่มสูงขึ้น และ
2. เพิ่มคุณภาพเสียงพูดที่ถูกปรับปรุงด้วย AENS โดยการพัฒนาวิธีการประมาณค่า a priori SDR

3.1. การประมาณค่า EPSD ที่นำเสนอ

การประมาณค่า EPSD ที่แม่นยำมากขึ้นจะช่วยให้ AENS สามารถกดเสียงสะท้อนลงได้ดีและแม่นยำยิ่งขึ้น การกดเสียงสะท้อนที่แม่นยำยิ่งขึ้นนี้ยังอาจส่งผลให้เกิดความผิดเพี้ยนที่ต่ำลงของเสียงพูดของผู้พูดในห้องใกล้ในช่วงสถานการณ์ดับเบิลทอล์กได้อีกด้วย วิทยานิพนธ์ฉบับนี้เสนอให้ใช้ วงจรกรองปรับตัวที่มีผลตอบอิมพัลส์แบบจำกัดในการประมาณค่า EPSD ดังรูปที่ 3.1



รูปที่ 3.1 การประมาณ EPSD ที่นำเสนอ

เริ่มจากเขียนฟังก์ชันต้นทุนดังสมการ

$$J(\mathbf{h}) = E\{|\psi(k,\ell) - \mathbf{h}^T(k,\ell)\mathbf{F}_{\text{pow}}(k,\ell)|^2\} \tag{3.1}$$

เมื่อ $\psi(k,\ell)$ คือ ค่าสเปกตรัมกำลังสัญญาณไมโครโฟน ณ ขณะนั้น ดังเช่นที่เคยได้นิยามไว้แล้วในหัวข้อย่อยที่ 2.2.2

$\mathbf{h}(k) = [h(k,0) \quad h(k,1) \quad \cdots \quad h(k,P-1)]^T$ คือ เวกเตอร์สัมประสทธิ์ของวงจรกรองปรับตัว ณ องค์ประกอบทางความถี่ที่ k

P คือ จำนวนสัมประสิทธิ์ของวงจรกรองปรับตัว ณ แต่ละองค์ประกอบทางความถี่ที่ k

$\mathbf{F}_{\text{pow}}(k,\ell) = [\varphi(k,\ell) \quad \varphi(k,\ell-1) \quad \cdots \quad \varphi(k,\ell-P+1)]^T$ คือ เวกเตอร์สเปกตรัมกำลังสัญญาณทางห้องไกล ณ องค์ประกอบทางความถี่ที่ k และเฟรมเวลาที่ ℓ

$\varphi(k,\ell) = |F(k,\ell)|^2$ คือสเปกตรัมกำลังสัญญาณทางห้องไกล ณ ขณะนั้น ณ องค์ประกอบทางความถี่ที่ k และเฟรมเวลาที่ ℓ

ดังนั้น ค่าติดลบของเกรเดียนต์ของฟังก์ชันต้นทุน สามารถหาได้เป็น [3]

$$-\nabla_{\mathbf{h}} J(\mathbf{h}) = \mathbf{p} - \mathbf{R}\mathbf{h} \tag{3.2}$$

เมื่อ เวกเตอร์สหสัมพันธ์ข้าม (Cross-Correlation Vector) ระหว่างค่าสเปกตรัมกำลังสัญญาณไมโครโฟน และเวกเตอร์สเปกตรัมกำลังสัญญาณทางห้องไกล นิยามเป็น

$$\begin{aligned}\mathbf{p} &= E\{\psi(k,\ell)\mathbf{F}_{\text{pow}}(k,\ell)\} \\ &= \lambda_N(k,\ell)E\{\mathbf{F}_{\text{pow}}(k,\ell)\} + E\{\varepsilon(k,\ell)\mathbf{F}_{\text{pow}}(k,\ell)\}\end{aligned} \tag{3.3}$$

$\varepsilon(k,\ell)$ คือ ค่าสเปกตรัมกำลังเสียงสะท้อน ณ ขณะนั้น เช่นเดียวกับที่เคยกล่าวไว้ในหัวข้อย่อยที่ 2.2.2 และเมทริกซ์สหสัมพันธ์อัตโนมัติ (Autocorrelation Matrix) ของเวกเตอร์สเปกตรัมกำลังสัญญาณทางห้องไกล

$$\mathbf{R} = E\{\mathbf{F}_{\text{pow}}(k,\ell)\mathbf{F}_{\text{pow}}^T(k,\ell)\} \tag{3.4}$$

จากวิธีการ Steepest-Descent [3] ทำให้ได้สมการปรับให้ทันกาล (Update Equation) ของวงจรกรองปรับตัวดังนี้

$$\mathbf{h}(k,\ell+1) = \mathbf{h}(k,\ell) + \mu(-\nabla J(\mathbf{h})\big|_{\mathbf{h}(k,\ell)}) \tag{3.5}$$

ทำการประมาณค่าเชิงสโตแคสติกของ $\mathbf{p}$ และ $\mathbf{R}$ ในสมการที่ (3.2) โดยใช้ค่า ณ ขณะเฟรมเวลานั้น (Instantaneous) ดังนี้

$$\hat{\mathbf{p}} = \lambda_N(k,\ell)\mathbf{F}_{\text{pow}}(k,\ell) + \varepsilon(k,\ell)\mathbf{F}_{\text{pow}}(k,\ell) \tag{3.6}$$

และ

$$\hat{\mathbf{R}} = \mathbf{F}_{\text{pow}}(k,\ell)\mathbf{F}_{\text{pow}}^T(k,\ell) \tag{3.7}$$

สมการปรับให้ทันกาลในสมการที่ (3.5) จึงเขียนได้เป็น

$$\mathbf{h}(k,\ell+1) = \mathbf{h}(k,\ell) + \mu\mathbf{F}_{\text{pow}}(k,\ell)\Big(\lambda_N(k,\ell) + \varepsilon(k,\ell) - \mathbf{F}_{\text{pow}}^T(k,\ell)\mathbf{h}(k,\ell)\Big) \tag{3.8}$$

ซึ่งก็คือสมการปรับตามกาลแบบ LMS ในโดเมนสเปกตรัมกำลังนั่นเอง ทำการปรับปรุงให้กลายเป็น NLMS ดังนี้

$$\mathbf{h}(k,\ell+1) = \mathbf{h}(k,\ell) + \mu\frac{\mathbf{F}_{\text{pow}}(k,\ell)\Big(\lambda_N(k,\ell) + \varepsilon(k,\ell) - \mathbf{F}_{\text{pow}}^T(k,\ell)\mathbf{h}(k,\ell)\Big)}{\mathbf{F}_{\text{pow}}^T(k,\ell)\mathbf{F}_{\text{pow}}(k,\ell) + \delta} \tag{3.9}$$

ในทางปฏิบัตินั้นพจน์ $\lambda_N(k,\ell)+\varepsilon(k,\ell)$ ในสมการที่ (3.9) ควรมีค่าเป็น $\varepsilon(k,\ell)$ เท่านั้น เนื่องจากต้องการหาค่าประมาณของเฉพาะ $\varepsilon(k,\ell)$ นั่นเอง ดังนั้นจึงเสนอให้แทนพจน์ดังกล่าวด้วย

$$\varepsilon_\psi(k,\ell) \approx \psi(k,\ell) - \lambda_N(k,\ell) \tag{3.10}$$

ทั้งนี้เพื่อลดผลของ NPSD ที่จะมาก่อกวนการปรับตัว และสมการปรับตามกาลเป็น

$$\mathbf{h}(k,\ell+1) = \mathbf{h}(k,\ell) + \mu \frac{\mathbf{F}_{\text{pow}}(k,\ell)\left(\varepsilon_\psi(k,\ell) - \mathbf{F}_{\text{pow}}^T(k,\ell)\mathbf{h}(k,\ell)\right)}{\mathbf{F}_{\text{pow}}^T(k,\ell)\mathbf{F}_{\text{pow}}(k,\ell)+\delta} \tag{3.11}$$

ในทางปฏิบัติ $\lambda_N(k,\ell)$ สามารถประมาณได้จากช่วงที่ไม่มีทั้งเสียงพูดทางห้องใกล้และเสียงสะท้อน ดังจะกล่าวถึงต่อไปในบทที่ 4

จากนั้น EPSD $\lambda_E(k,\ell)$ จะถูกประมาณโดย

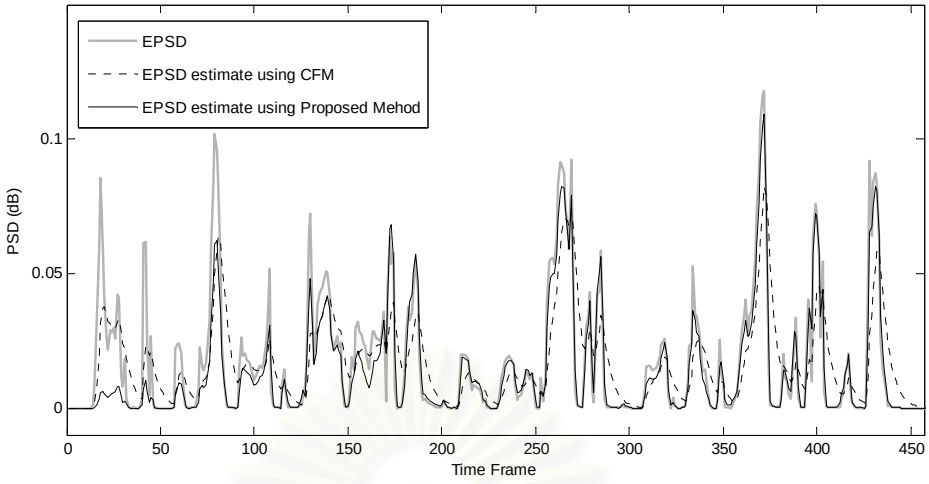
$$\hat{\lambda}_E(k,\ell) \approx \hat{\varepsilon}(k,\ell) = \mathbf{h}^T(k,\ell)\mathbf{F}_{\text{pow}}(k,\ell) \tag{3.12}$$

การประมาณเช่นนี้เปรียบได้กับ AEC ที่อาศัยวงจรกรองแบบปรับตัวซึ่งทำงานบนโดเมนสเปกตรัมกำลังนั่นเอง แต่อย่างไรก็ตามความสัมพันธ์ระหว่าง $\mathbf{h}^T(k,\ell)$ กับวิถีสะท้อนทางเสียงในห้องใกล้นั้นเป็นเรื่องที่ซับซ้อนและต้องได้รับการศึกษาต่อไป การทำงานบนโดเมนสเปกตรัมกำลังนี้ทำให้ความซับซ้อนในการคำนวณของวิธีการที่นำเสนอลดลงจาก AEC เป็นอย่างมาก ทั้งนี้เนื่องจากการประมาณค่า EPSD ไม่มีการพิจารณาข้อมูลทางเฟสของแต่ละสัญญาณที่ทำการพิจารณา ในขณะที่การประมาณสัญญาณเสียงสะท้อนใน AEC ต้องทำการประมาณทั้งขนาดและเฟสสเปกตรัมเสียงสะท้อนเพื่อสามารถนำค่าประมาณสัญญาณเสียงสะท้อนที่ได้ไปหักลบกับสัญญาณเสียงสะท้อนจริงได้อย่างแม่นยำ[11] การคิดคำนึงถึงปริมาณข้อมูลทางเฟสนี้เอง ที่ทำให้การประมาณสัญญาณเสียงสะท้อนใน AEC มีความซับซ้อนที่มากกว่าการประมาณค่า EPSD ที่นำเสนอ

ผลการเปรียบเทียบการประมาณค่า EPSD ระหว่าง CFM และวิธีที่นำเสนอถูกแสดงในรูปที่ 3.2 ซึ่งเห็นได้ว่าวิธีการประมาณค่า EPSD ที่นำเสนอให้ผลการประมาณค่า EPSD ที่ตรงกับค่า EPSD ณ ขณะนั้นมากกว่า CFM แต่อย่างไรก็ตามวิธีการที่นำเสนอต้องอาศัยระยะเวลาในการปรับตัวระยะหนึ่งก่อนที่จะสามารถประมาณค่า EPSD ได้อย่างเหมาะสม (ในรูปที่ 3.2 การประมาณค่า EPSD ที่นำเสนอใช้เวลาประมาณ 100 เฟรม ในการปรับตัว หรือประมาณ 1.6 วินาที ณ ความถี่ชักตัวอย่าง 8 kHz ซึ่งถือว่าน้อยมากเมื่อเทียบกับระยะที่ใช้ในการลู่เข้าสู่สถานะคงตัวในระบบ AEC ซึ่งใช้ขั้นตอนการปรับตัวในตระกูล LMS) ในขณะที่ CFM ไม่ต้องอาศัยช่วงระยะเวลาดังกล่าวเลย

---

[11] ในทางปฏิบัตินั้น AEC ที่ทำงานบนโดเมนเวลาไม่ได้ทำการประมาณค่า ขนาดสเปกตรัม และเฟสสเปกตรัมของสัญญาณเสียงสะท้อน ออกมาโดยตรง แต่ทั้งนี้เนื่องจากค่าประมาณสัญญาณเสียงสะท้อนที่ AEC ประมาณออกมาได้ในทางเวลานั้น เป็นสัญญาณเสียงสะท้อนที่คล้ายกับสัญญาณเสียงสะท้อนจริงในทางเวลา ทำให้ในขนาดและเฟสในโดเมนความถี่ของค่าประมาณสัญญาณเสียงสะท้อนดังกล่าว ไปเหมือนกับค่าขนาดและเฟสของเสียงสะท้อนจริงโดยปริยาย และอาจกล่าวได้ว่า AEC ต้องทำการประมาณทั้งขนาดและเฟสของสัญญาณเสียงสะท้อนได้



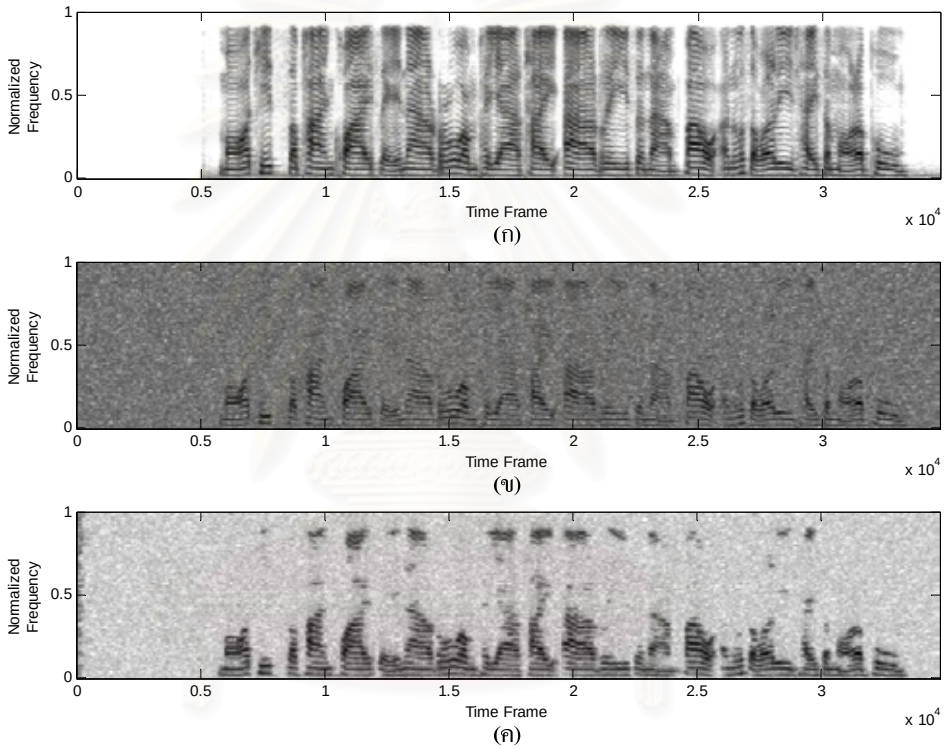
รูปที่ 3.2 เปรียบเทียบการประมาณค่า EPSD ณ องค์ประกอบทางความถี่ที่ 10 ที่ระดับ Global SNR 20 dB ระหว่าง CFM และวิธีการที่นำเสนอ

ในกรณี DTS การประมาณค่า EPSD ที่นำเสนอ จะต้องอาศัย DTD เพื่อตรวจหาช่วง DTS เช่นเดียวกับระบบ AEC โดยทำการคงค่าสัมประสิทธิ์ของวงจรกรองปรับตัวในแต่ละองค์ประกอบทางความถี่เอาไว้

การประมาณค่า EPSD ที่นำเสนออาศัยหลักการประมาณเช่นเดียวกับ PAES [47] เพียงแต่ไม่ได้ทำการย้ายโดเมนไปพิจารณาในโดเมนซองสเปกตรัม โดยที่สามารถกระทำเช่นนี้ได้โดยไม่เกิดผลของ Musical noise ในสัญญาณที่ถูกปรับปรุง เป็นเพราะว่าวิธีการประมาณ EPSD ที่นำเสนอจะถูกใช้ควบคู่ไปกับเทคนิค SS ซึ่งอาศัย a priori SDR เป็นตัวแปรหลัก [19] ดังนั้นผลการแกว่งของ EPSD จะถูกปรับปรุงในขั้นตอนการประมาณ a priori SDR อีกครั้งหนึ่ง

3.2. การปรับปรุงการประมาณค่า a priori SDR

เมื่อพิจารณาว่า AENS เป็นวิธีการลดเสียงสะท้อนและเสียงรบกวนที่อาศัยเทคนิคการกดทางสเปกตรัมดังที่ได้กล่าวไว้ในหัวข้อที่ 2.3.4.2 ทำให้สามารถวิเคราะห์ได้ว่าความผิดเพี้ยนของเสียงพูด[12] ซึ่งเป็นสัญญาณขาออกของ AENS มีต้นเหตุมาจากเทคนิคการกดทางสเปกตรัมนั่นเอง โดยความผิดเพี้ยนของเสียงพูดดังกล่าวสามารถรับรู้ได้โดยการฟังเสียงพูด (ที่ไม่ถูกรบกวน) เปรียบเทียบกับสัญญาณเสียงพูดขาออกของ AENS (หรือเสียงพูดที่ถูกปรับปรุง) ตัวอย่าง สเปกโทรแกรม (Spectrogram) ซึ่งถูกใช้บรรยายสเปกตรัมสัญญาณที่เวลาต่างๆ ของเสียงพูดสะอาด ของเสียงพูดที่ถูกก่อกวนด้วยเสียงรบกวนพื้นหลังที่ระดับ SNR 5 dB และของค่าประมาณเสียงพูด หรือเสียงพูดที่ถูกปรับปรุงที่ได้จากการใช้เทคนิคการกดทางสเปกตรัม[13] ถูกแสดงในรูปที่ 3.3 (ก)-(ค) ตามลำดับ



รูปที่ 3.3 สเปกโทรแกรมของ ก) เสียงพูดสะอาด ข) เสียงพูดที่ถูกรบกวนที่ระดับ SNR 5 dB และ ค) เสียงพูดที่ถูกปรับปรุงจากวิธีการ NS

จากสเปกโทรแกรมในรูปที่ 3.3 จะเห็นถึงความแตกต่างระหว่างเสียงพูดและเสียงพูดที่ถูกปรับปรุงประการหนึ่งซึ่งอาจแสดงถึงความผิดเพี้ยนของเสียงพูดที่ถูกปรับปรุง นั่นคือการที่สเปกตรัมเสียงพูดในบางองค์ประกอบ

---

[12] สำหรับการสนทนาแบบแฮนด์ฟรี เสียงพูดในที่นี้ ได้แก่ เสียงพูดของผู้พูดทางห้องใกล้

[13] ในที่นี้เสียงก่อกวนได้แก่ เสียงรบกวนพื้นหลัง เพียงชนิดเดียวดังนั้น a priori SDR จะเท่ากับ a priori SNR (a priori SER มีค่าเป็นอนันต์) หรือกล่าวคือ AENS กลายเป็นวิธี NS ไปโดยปริยาย

ทางความถี่เกิดสูญหายไป โดยเฉพาะอย่างยิ่งในองค์ประกอบทางความถี่ที่ สเปกตรัมเสียงพูดสะอาดมีพลังงานค่อนข้างน้อยเมื่อเทียบกับเสียงรบกวน (องค์ประกอบทางความถี่ดังกล่าวได้แก่องค์ประกอบทางความถี่ที่เสียงพูดถูกบดบังด้วยเสียงรบกวนจนไม่สามารถมองเห็นได้ในสเปกโทรแกรมของเสียงพูดที่ถูกรบกวนในรูปที่ 3.3 ข.) วิทยานิพนธ์ฉบับนี้เชื่อว่าหากสามารถรักษาเสียงพูดในแต่ละองค์ประกอบทางความถี่ไว้ได้มากยิ่งขึ้นก็จะทำให้สามารถลดความผิดเพี้ยนของเสียงพูดที่ถูกปรับปรุงลงได้

ความสามารถในการรักษาเสียงพูดในแต่ละองค์ประกอบทางความถี่ เกี่ยวข้องกับการประมาณค่า a priori SDR เป็นอย่างมาก เนื่องจาก a priori SDR $\xi^D(k,\ell)$ เป็นตัวแปรหลักของ Spectral Gain $G_\pi(k,\ell)$ กล่าวอย่างสังเขปคือ หาก $\xi^D(k,\ell)$ มีค่าสูง $G_\pi(k,\ell)$ ก็จะมีค่าประมาณ 1 (ไม่กดสัญญาณในองค์ประกอบทางความถี่นั้นๆ เนื่องจากมีส่วนของสัญญาณที่ต้องการอยู่มากกว่าสัญญาณก่อกวนมาก) และหาก $\xi^D(k,\ell)$ มีค่าต่ำแล้ว $G_\pi(k,\ell)$ ก็จะมีค่าต่ำไปด้วย (กดสัญญาณนั้นลงอย่างมากเนื่องจากมีส่วนของสัญญาณก่อกวนอยู่มาก) ดังนั้นการประมาณค่า a priori SDR ที่ผิดไปจากความเป็นจริง จึงเป็นสาเหตุหลักของการ “กดมากเกินไป” (Over Suppression) และ “การกดน้อยเกินไป” (Under Suppression) ผลลัพธ์ของการกดมากเกินไปในบางองค์ประกอบทางความถี่นี้เองที่ทำให้เสียงพูดในองค์ประกอบทางความถี่นั้นๆ สูญหายไป ส่วนการกดน้อยเกินไปในบางกรณีจะทำให้เกิดเสียงรบกวนตกค้างที่เรียกว่า Musical Noise ดังนั้นการประมาณค่า a priori SDR ที่ดีจะนำมาซึ่งการรักษาสเปกตรัมเสียงพูดไว้ได้มาก ในขณะเดียวกับที่ไม่ก่อให้เกิด Musical Noise ถ้าเช่นนั้นหากสามารถพัฒนาการประมาณค่า a priori SDR ที่มีประสิทธิภาพดียิ่งขึ้น เป็นผลให้สามารถรักษาส่วนของเสียงพูดในแต่ละองค์ประกอบทางความถี่ไว้มากยิ่งขึ้น ในขณะที่ไม่ก่อให้เกิดเสียงรบกวนตกค้างแบบ Musical Noise ที่มากยิ่งขึ้นแล้ว ก็จะส่งผลให้สามารถลดความผิดเพี้ยนของเสียงพูดที่ถูกปรับปรุงลงได้ ด้วยเหตุจูงใจนี้เอง ทำให้วิทยานิพนธ์ฉบับนี้เลือกที่จะทำการพัฒนาการประมาณค่า a priori SDR เพื่อลดความผิดเพี้ยนของเสียงพูดที่ถูกปรับปรุงลง โดยจากการนิยามค่า a priori SDR ดังสมการที่ (2.104) ทำให้สามารถประมาณค่า a priori SDR ได้จากค่าประมาณ a priori SER และค่าประมาณ a priori SNR ดังนั้นการพัฒนาการประมาณค่า a priori SDR จึงสามารถถูกแบ่งพิจารณาได้เป็น 2 ส่วนได้แก่ การประมาณค่า a priori SNR แบบต่างๆ จนกระทั่งถึงการประมาณค่า a priori SNR ที่นำเสนอและการประมาณค่า a priori SER ที่นำเสนอ ดังรายละเอียดต่อไปนี้

3.2.1. การประมาณค่า a priori SNR

การวิเคราะห์การประมาณค่า a priori SNR ในหัวข้อนี้ จะกระทำโดยสมมติว่าเสียงก่อกวนมีเพียงเสียงรบกวนพื้นหลังเพียงชนิดเดียว ดังนั้นเมื่ออยู่ในสถานการณ์เช่นนี้ ค่า a priori SDR จะเท่ากับค่า a priori SNR เนื่องจากค่า a priori SER มีค่าเป็นอนันต์ ($\lambda_E(k,\ell)=0\ ,\ \forall k,\ell$) จึงกลับไปสู่วิธีการ NS นั่นเอง โดยวัตถุประสงค์ในการพัฒนาการประมาณค่า a priori SNR ก็คือการได้มาซึ่งวิธีการประมาณค่า a priori SNR ที่สามารถรักษาสเปกตรัมเสียงพูดไว้ได้มากที่สุด ในขณะที่ไม่ก่อให้เกิดเสียงรบกวนตกค้าง Musical Noise

เนื่องจากความไม่เป็นจุดนิ่งของสัญญาณเสียงพูด วิทยานิพนธ์ฉบับนี้จะพิจารณาว่า ค่า a priori SNR ในช่วงที่มีเสียงพูด ซึ่งได้แก่ค่าอัตราส่วนระหว่างค่าคาดหวังของสเปกตรัมกำลังเสียงพูดและค่าคาดหวังของสเปกตรัมกำลังเสียงรบกวน ดังสมการที่ (2.22) (ซึ่งถูกเขียนใหม่ดังนี้)

$$\xi(k,\ell) = \frac{\lambda_S(k,\ell)}{\lambda_N(k,\ell)}$$

ควรถูกประมาณด้วยค่า Instantaneous SNR ซึ่งถูกมองว่าเป็นค่าอัตราส่วนระหว่างสเปกตรัมกำลังเสียงพูดและค่าคาดหวังของสเปกตรัมกำลังเสียงรบกวน ดังสมการที่ (2.38) (ซึ่งถูกเขียนใหม่ดังนี้)

$$\delta(k,\ell) = \max[\frac{|Y(k,\ell)|^2 - \lambda_N(k,\ell)}{\lambda_N(k,\ell)}, 0]$$

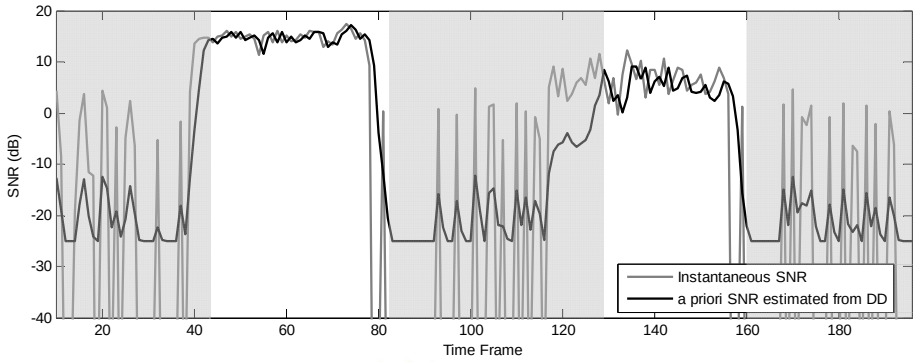
ทั้งนี้ค่า Instantaneous SNR ดังกล่าวจะไม่สามารถบ่งบอกถึงค่า a priori SNR ที่เหมาะสมในกรณีที่สเปกตรัมกำลังของเสียงพูดมีค่าน้อยมากเมื่อเทียบกับ ค่าสเปกตรัมมกำลังเสียงรบกวน ซึ่งจะตรงกับกรณีที่ ค่า Instantaneous SNR ถูกปัดให้เป็น 0 ในช่วงที่มีเสียงพูดในสมการที่ (2.38) แต่อย่างไรก็ตามวิทยานิพนธ์ฉบับนี้เชื่อว่าเสียงพูดที่มีพลังงานต่ำมากในกรณีดังกล่าวไม่มีความสำคัญต่อการได้ยินของมนุษย์มากนัก เพราะฉะนั้นจึงเป็นการเหมาะสมที่จะทำการประมาณค่า a priori SNR ด้วยค่า Instantaneous SNR ในช่วงที่มีเสียงพูด

สำหรับในช่วงที่ไม่มีเสียงพูดวิทยานิพนธ์ฉบับนี้จะพิจารณาว่าค่าประมาณ a priori SNR ควรมีค่าต่ำๆ และราบเรียบเพื่อให้ระดับเสียงรบกวนตกค้างมีค่าน้อย และลดผลของเสียงรบกวนตกค้างแบบ Musical Noise ตามลำดับ อย่างไรก็ตามเนื่องจากไม่สามารถบ่งชี้ได้แน่นอนว่าค่าประมาณ a priori SNR ควรมีค่าต่ำเพียงใด และราบเรียบเท่าใดจึงจะปลอดจากปัญหาเสียงรบกวนตกค้างแบบ Musical Noise ดังนั้นจึงอาศัยการเปรียบเทียบค่าประมาณ a priori SNR ในช่วงที่ไม่มีเสียงพูดนี้กับค่าประมาณจากวิธี DD ที่ได้ในช่วงเวลาเดียวกัน ทั้งนี้เนื่องจากวิธี DD ได้รับการยอมรับว่าเป็นวิธีที่มีความสามารถในการลดเสียงรบกวนตกค้างแบบ Musical Noise ได้เป็นอย่างดี จากที่กล่าวมาแล้วทำให้สามารถสรุปคุณสมบัติของวิธีการประมาณค่า a priori SNR ที่ควรจะเป็นได้ดังนี้

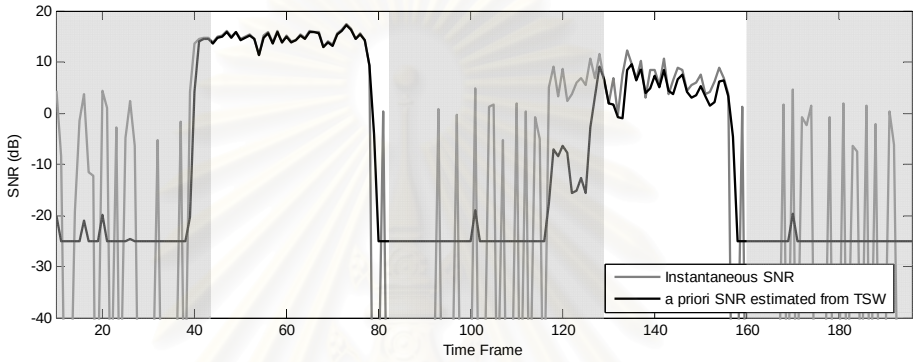
- **ในช่วงที่มีเสียงพูด** ค่าประมาณ a priori SNR ควรมีค่าเท่ากับค่า Instantaneous SNR
- **ในช่วงที่ไม่มีเสียงพูด** ค่าประมาณ a priori SNR ควรมีค่าน้อยๆ และราบเรียบ

$$\hat{\xi}(k,\ell) = \begin{cases} \delta(k,\ell), & \text{VAD is on} \\ 0, & \text{VAD is off} \end{cases} \qquad (3.13)$$

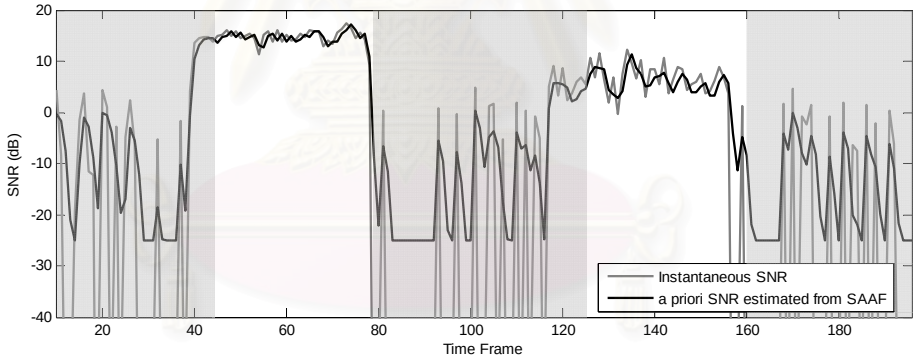
ลักษณะการประมาณค่า a priori SNR ของการประมาณแบบต่างๆ ที่ถูกศึกษาในวิทยานิพนธ์ฉบับนี้ (ได้แก่ วิธี DD วิธี TSW และวิธี SAAF) ถูกสังเกตผ่านทางการทดลองที่จำลองสัญญาณเสียงพูดที่ถูกรบกวนด้วยสัญญาณความถี่เดียวจำนวน 2 พัลส์สัญญาณ (พัลส์สัญญาณแรกอยู่ระหว่างเฟรมเวลาที่ 40 ถึง 80 และพัลส์สัญญาณหลังอยู่ระหว่างเฟรมเวลาที่ 115 ถึง 155 โดยจะถือว่าช่วงเวลาที่มีสัญญาณความถี่เดียวอยู่เป็นตัวแทนของช่วงเวลาที่มีเสียงพูด และช่วงเวลาที่ไม่มีสัญญาณความถี่เดียวอยู่เป็นตัวแทนของช่วงที่ไม่มีเสียงพูด) ณ ความถี่ 2000 Hz พัลส์สัญญาณแรกมีพลังงานมากกว่าพัลส์สัญญาณหลังประมาณ 7 dB ที่ระดับ Global SNR 0 dB ผลการทดลองในรูปที่ 3.4 แสดงให้เห็นว่า



(ก)



(ข)



(ค)

รูปที่ 3.4 ลักษณะของค่าประมาณ a priori SNR ที่ได้จากวิธีการประมาณแบบ
(ก) DD (ข) TSW และ (ค) SAAF
เมื่อจำลองสัญญาณเสียงที่ถูกรบกวนด้วยสัญญาณความถี่เดียวจำนวน 2 พัลส์สัญญาณ
ณ ความถี่ 2000 Hz ที่ระดับ Global SNR 0 dB

- ในช่วงที่สัญญาณความถี่เดียวมีพลังงานมากเมื่อเทียบกับพลังงานของเสียงรบกวน (ช่วงพัลส์สัญญาณแรก) การประมาณค่า a priori SNR แต่ละแบบต่างให้ค่าประมาณ a priori SNR ออกมาประมาณเท่ากับค่า Instantaneous SNR ซึ่งถือว่าเป็นลักษณะการประมาณที่ต้องการ

- ช่วงที่สัญญาณความถี่เดียวมีพลังงานค่อนข้างน้อยเมื่อเปรียบเทียบกับพลังงานของเสียงรบกวน (ช่วงพัลส์สัญญาณหลัง) ค่าประมาณ a priori SNR ที่ได้จากการประมาณแบบต่างๆ จะแตกต่างกันออกไป โดยประมาณช่วงเฟรมเวลาที่ 115 ถึงเฟรมเวลาที่ 125 ซึ่งเป็นช่วงที่มีเสียงพูดนั้น ค่าประมาณ a priori SNR ที่ได้จากวิธี DD และ TSW มีค่าน้อยกว่าค่า Instantanous SNR (น้อยกว่าที่ควรจะเป็น) ในขณะที่ค่าประมาณ a priori SNR ที่ได้จากวิธี SAAF มีค่าประมาณเท่ากับค่า Instantaneous SNR ทำให้เห็นได้ว่าวิธี SAAF เป็นวิธีการประมาณค่า a priori SNR ที่ดีที่สุดสำหรับช่วงที่มีเสียงพูด

- สำหรับช่วงที่ไม่มีเสียงพูดจะทำการเปรียบเทียบกับวิธี DD เนื่องจากเป็นวิธีที่ได้รับการยอมรับในเรื่องของการลดผลของเสียงรบกวนตกค้างแบบ Musical Noise โดยจะเห็นว่าวิธี TSW ให้ค่าประมาณ a priori SNR ที่ต่ำกว่าและราบเรียบกว่าวิธี DD ทำให้วิธี TSW นี้สามารถลดผลของเสียงรบกวนตกค้างแบบ Musical Noise ได้ดีกว่าวิธี DD (การทดลองฟังอย่างไม่เป็นทางการยืนยันผลการวิเคราะห์นี้) ในส่วนของวิธี SAAF จะเห็นว่าค่าประมาณ a priori SNR ในช่วงที่ไม่มีเสียงพูดอยู่ในระดับที่สูงกว่าและแกว่งมากกว่าวิธี DD ทำให้วิธี SAAF ได้รับผลของ Musical Noise มากกว่าวิธี DD ซึ่งสามารถได้ยินได้จากการทดลองฟังอย่างไม่เป็นทางการเช่นกัน จากที่ได้กล่าวมาแล้วในข้างต้นทำให้สามารถสรุปได้ว่า

<u>วิธี DD</u> สามารถประมาณค่า a priori SNR ได้อย่างเหมาะสมในช่วงที่มีเสียงพูดเมื่อเสียงพูดมีพลังงานมากเมื่อเปรียบเทียบกับเสียงรบกวน แต่เมื่อเสียงพูดมีพลังงานน้อยเมื่อเปรียบเทียบกับเสียงรบกวนแล้วค่าประมาณ a priori SNR ที่ได้จากการประมาณประเภทนี้จะไม่สามารถติดตาม (Tracking) ค่า Instantaneous SNR ได้อย่างรวดเร็ว กล่าวคือต้องอาศัยช่วงเวลาระยะหนึ่งในการลู่เข้า (Convergence) สู่ค่า Instantaneous SNR ในขณะที่ในช่วงที่ไม่มีเสียงพูดนั้นวิธี DD ได้รับการยอมรับในเรื่องของประสิทธิภาพการลดเสียงรบกวนตกค้างแบบ Musical Noise และถูกใช้เป็นตัวเปรียบเทียบกับวิธีการอื่นๆ

<u>วิธี TSW</u> สามารถประมาณค่า a priori SNR ได้อย่างแม่นยำมากกว่าวิธี DD ในช่วงที่มีเสียงพูด (สังเกตได้จากการที่ค่าประมาณ a priori SNR ที่ได้จากวิธี TSW มีค่าเท่ากับค่า Instantaneouds SNR ในขณะที่ค่าประมาณ a priori SNR ที่ได้จากวิธี DD มีค่าเท่ากับค่า Instantaneous SNR ในเฟรมก่อนหน้า) แต่ทั้งนี้เมื่อเสียงพูดมีพลังงานน้อยเมื่อเทียบกับเสียงรบกวนแล้ววิธี TSW ก็ให้การประมาณค่า a priori SNR ที่ต่ำกว่าค่า Instantaneous SNR เช่นกัน ส่วนในช่วงเวลาที่ไม่มีเสียงพูดวิธีการ TSW ให้การประมาณค่า a priori SNR ที่น้อยและราบเรียบกว่าวิธี DD ทำให้ได้ผลการลดลงของ เสียงรบกวนและเสียงรบกวนตกค้างแบบ Musical Noise ที่ลดลงมากกว่าวิธี DD

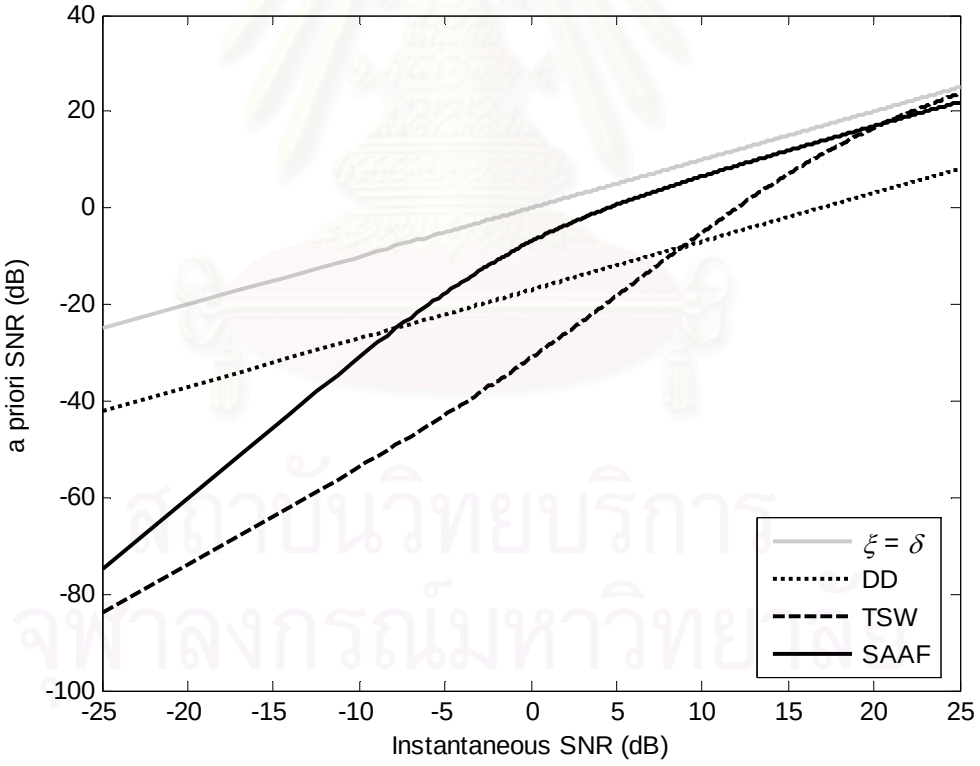
<u>วิธี SAAF</u> สามารถประมาณค่า a priori SNR ได้อย่างเหมาะสมในช่วงเวลาที่มีเสียงพูด แต่ในช่วงที่ไม่มีเสียงพูดวิธี SAAF ให้ค่าประมาณ a priori SNR ที่ค่อนข้างสูงและแกว่งทำให้วิธี SAAF ประสบกับปัญหาเสียงรบกวนตกค้างแบบ Musical Noise ที่มากกว่าวิธี DD และ TSW

ทั้งนี้การประมาณค่า a priori SNR ที่ต่ำกว่าค่า Instantaneous SNR ในช่วงที่มีเสียงพูดเป็นสาเหตุให้เกิดการกดมากเกินไปและนำมาซึ่งการสูญหายของส่วนของเสียงพูดในบางองค์ประกอบทางความถี่ และถึงแม้วิธี DD และ TSW จะสามารถประมาณค่า a priori SNR ได้อย่างเหมาะสมเมื่อให้เวลาระยะหนึ่ง แต่ในทางปฏิบัติค่า Instantaneous SNR ที่เกิดขึ้นในช่วงที่มีเสียงพูดจะกินเวลาเพียงไม่นาน (20-100 ms [28] หรือ 2-8 เฟรมเวลา) เนื่องจากความไม่เป็นจุดนิ่งของสัญญาณเสียงพูด ดังนั้นระยะเวลาดังกล่าวจะไม่เพียงพอสำหรับวิธี DD และ TSW

ลักษณะการประมาณค่า a priori SNR ของวิธีการประมาณแบบต่างๆ นอกจากจะสามารถสังเกตได้จากการทดลองที่ได้กล่าวมาแล้ว ยังสามารถสังเกตได้จาก Transition Equation ของการประมาณแบบต่างๆ ควบคู่กันไปได้อีกด้วย ดังจะได้กล่าวในรายละเอียดต่อไป

Transition Equation

Transition Equation (TE) เป็นสมการที่ได้มาจากการประมาณให้พจน์ของค่าสเปกตรัมเสียงพูดที่ถูกประมาณในเฟรมที่แล้วของวิธีการประมาณค่า a priori SNR ในตระกูล DD มีค่าเป็น 0 หรือน้อยๆ มากๆ จนตัดทิ้งได้ ทำให้สมการการประมาณค่า a priori SNR ของวิธีการต่างๆ ในตระกูล DD สามารถลดตัวแปรลงเหลือเพียง ค่า Instantaneous SNR เพียงตัวเดียว ทำให้สามารถวิเคราะห์พฤติกรรมของการประมาณค่า a priori SNR แบบต่างๆ ได้ง่ายยิ่งขึ้น แต่อย่างไรก็ตามการตัดส่วนของค่าสเปกตรัมเสียงพูดที่ถูกปรับปรุงที่ได้จากเฟรมก่อนหน้าทิ้งไปทำให้ TE สามารถใช้อธิบายลักษณะการประมาณค่า a priori SNR ได้เพียงสังเขป และเพียงในช่วงสถานการณ์ที่เฟรมก่อนหน้าเป็นเฟรมที่ไม่มีเสียงพูด ซึ่งได้แก่ช่วงที่แรเงาในรูปที่ 3.4 เท่านั้น ดังนั้นการวิเคราะห์ความหมายของ TE จึงต้องกระทำโดยอิงผลจากการทดลองในรูปที่ 3.4 ควบคู่ไปด้วย TE ของวิธี DD, TSW และ SAAF ซึ่งตรงกับสมการที่ (2.40), (2.42) และ (2.45) ตามลำดับ ถูกแสดงดังรูปที่ 3.5



รูปที่ 3.5 การเปรียบเทียบ Transition equation ของการประมาณ a priori SNR แบบต่างๆ

<u>กราฟ TE ของวิธี SAAF</u> นั้นลู่เข้าสู่กราฟ $\xi = \delta$ เร็วที่สุดซึ่งสามารถตีความหมายได้ว่า วิธี SAAF จะเริ่มประมาณค่า a priori SNR ด้วยค่า Instantaneous SNR ตั้งแต่ที่ค่า Instantaneous SNR ที่ต่ำ หรือกล่าวคือ วิธี SAAF มัก

ประมาณค่า a priori SNR ด้วยค่า Instantaneous SNR เหตุนี้เองทำให้วิธี SAAF ให้ค่าประมาณ a priori SNR ที่เหมาะสมในช่วงที่มีเสียงพูด โดยช่วงที่มีเสียงเสียงพูดในที่นี้หมายถึงเพียง เฟรมที่เริ่มมีเสียงพูด หรือ เฟรมที่เสียงพูดที่ถูกประมาณในเฟรมก่อนหน้ามีค่าน้อยๆ เท่านั้น เพราะ TE ไม่สามารถอธิบายครอบคลุมไปถึงช่วงที่เฟรมก่อนหน้ามีเสียงพูดอยู่ได้ แต่อย่างไรก็ตาม วิธี SAAF ก็ให้ค่าประมาณ a priori SNR ที่ไม่เหมาะสมในช่วงที่ไม่มีเสียงพูด

<u>กราฟ TE ของวิธี DD</u> ขนานไปกับกราฟ $\hat{\xi} = \delta$ แสดงให้เห็นว่าหากปราศจากพจน์ของสเปกตรัมเสียงพูดที่ถูกปรับปรุงในเฟรมก่อนหน้าแล้ว วิธี DD จะไม่ประมาณค่า a priori SNR ด้วยค่า Instantaneous SNR แต่จะประมาณค่า a priori SNR ด้วยค่า $(1-\alpha_{\mathrm{DD}})\delta(k,\ell)$ ทุกๆ ค่า Instantaneous SNR ซึ่งเป็นการแสดงให้เห็นถึงความเชื่องช้าในการติดตามค่า Instantaneous SNR ของวิธี DD เหตุนี้เองทำให้วิธี DD ไม่สามารถประมาณค่า a priori SNR ได้อย่างเหมาะสมในช่วงที่มีเสียงพูด โดยเฉพาะอย่างยิ่งในกรณีที่เสียงพูดมีพลังงานน้อยเมื่อเทียบกับเสียงรบกวน (Instantaneous SNR มีค่าต่ำ) แต่อย่างไรก็ดีลักษณะการทำงานเช่นนี้เป็นผลดีสำหรับช่วงที่ไม่มีเสียงพูด เนื่องจากในช่วงดังกล่าว ค่า Instantaneous SNR จะเกิดการแกว่งไปมาเนื่องจากสเปกตรัมกำลังเสียงรบกวนที่แกว่งไปมารอบค่าคาดหวังของมัน การที่วิธี DD เชื่องช้าต่อการติดตามค่า Instantaneous SNR ที่มีค่าน้อยทำให้วิธี DD ไม่ให้ค่าประมาณ a priori SNR ที่แกว่งไปมาตามค่า Instantaneous SNR นั่นเอง

<u>กราฟ TE ของวิธี TSW</u> ลู่เข้าสู่กราฟ $\hat{\xi} = \delta$ ที่ค่า Instantaneous SNR ที่สูงกว่าวิธี SAAF (และถือว่าต่ำกว่าวิธี DD) ซึ่งสามารถตีความได้ว่าวิธี TSW มีความไวต่อการติดตามค่า Instantaneous SNR อยู่ระหว่างวิธี DD ดังนั้นหากทำการวิเคราะห์อย่างสังเขปดังที่ผ่านมาแล้ว วิธี TSW ต้องสามารถประมาณค่า a priori SNR ในช่วงที่มีเสียงพูดได้อย่างเหมาะสมมากกว่าวิธี DD และได้แย่กว่าวิธี SAAF อย่างไรก็ตามจากผลการทดลองในรูปที่ 3.4 ในช่วงเฟรมเวลาที่ 115 ถึง 125 วิธี TSW ให้การประมาณค่า a priori SNR ที่ดีกว่าวิธี DD ในเฉพาะบางช่วงเท่านั้น โดยหากสังเกตให้ดีคือวิธี TSW จะให้ค่าประมาณ a priori SNR ในช่วงที่มีเสียงพูดที่ดีกว่าวิธี DD ก็ต่อเมื่อค่า Instantaneous SNR มีค่าเกิน 10 dB ขึ้นไป ซึ่งก็คือจุดที่กราฟ TE ของวิธี TSW เริ่มสูงขึ้นกว่าวิธี DD นั่นเอง

จากการวิเคราะห์ที่ผ่านมานี้ทำให้สามารถสรุปเป็นวิธีการตีความ ความหมายของ TE ได้ว่า การประมาณที่เริ่มให้ค่าประมาณ a priori SNR ลู่เข้าสู่กราฟ $\hat{\xi} = \delta$ ที่ค่า Instantaneous SNR ต่ำเท่าใดจะหมายถึงความไวในการติดตามค่า Instantaneous SNR ได้ดียิ่งขึ้นเท่านั้น แต่อย่างไรก็ตามการติดตามค่า Instantaneous SNR ที่ไวก็ไม่ได้เป็นที่ต้องการเสมอไปเพราะอาจนำมาซึ่งเสียงรบกวนตกค้างแบบ Musical Noise ดังเช่นในกรณีของ SAAF โดยสำหรับผลของเสียงรบกวนตกค้างแบบ Musical Noise นั้นจะสามารถทราบได้ก็ต่อเมื่อทำการทดลองเช่นเดียวกับในรูปที่ 3.4 และเปรียบเทียบกับวิธี DD เท่านั้น

3.2.2. การประมาณค่า a priori SNR ที่นำเสนอ

3.2.2.1. Two-Spectral Power (TSSP)

วิทยานิพนธ์ฉบับนี้เสนอให้ใช้วิธี TSNR โดยเลือก G_ψ ในสมการที่ (2.41) เป็น G_{SP} และเรียกว่า Two-Step Spectral Power (TSSP) โดยค่าประมาณ a priori SNR สำหรับวิธี TSSP สามารถหาได้จาก

$$\hat{\xi}^N_{\mathrm{TSSP}}(k,\ell) = \left(G_{\mathrm{SP}}(k,\ell)\Big|_{\hat{\xi}^N_{\mathrm{DD}}(k,\ell),\gamma^N(k,\ell)} \right)^2 \gamma^N(k,\ell) \tag{3.14}$$

เมื่อแทนค่า G_{SP} ในสมการที่ (2.28) ลงในสมการที่ (3.14) และทำการจัดรูปใหม่ (ภาคผนวก ก) จะได้

$$\hat{\xi}^N_{\mathrm{TSSP}}(k,\ell) = \hat{\xi}^N_{\mathrm{DD}}(k,\ell) + K_g\left(\delta^N(k,\ell) - \hat{\xi}^N_{\mathrm{DD}}(k,\ell)\right) \tag{3.15}$$

เมื่อ

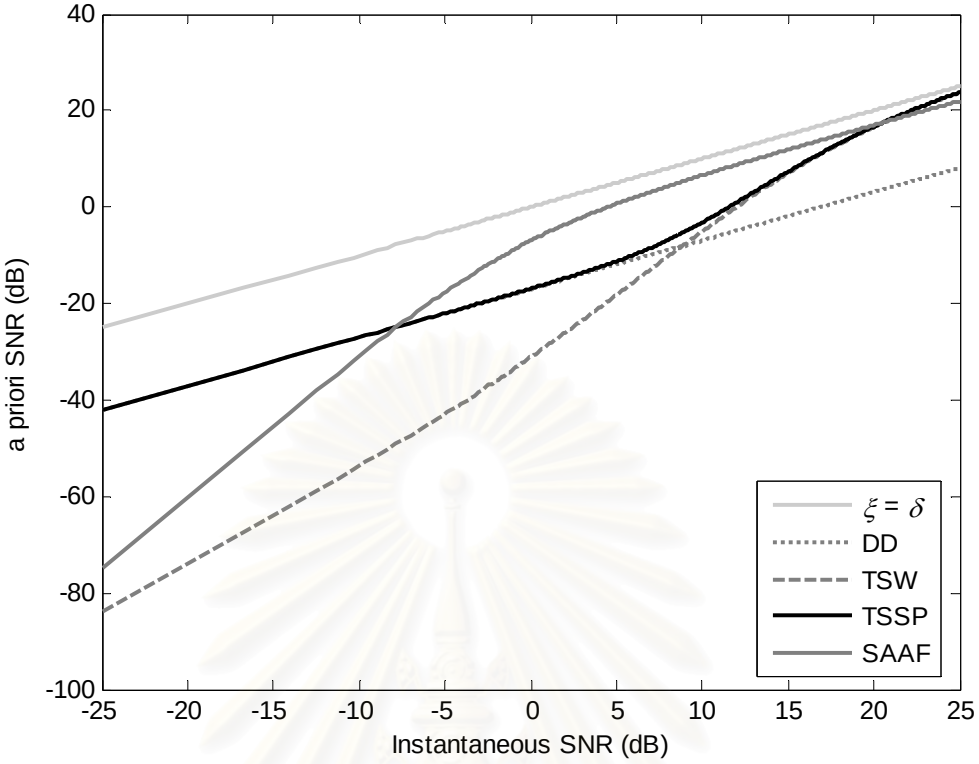
$$K_g = \left(\frac{\hat{\xi}^N_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^N_{\mathrm{DD}}(k,\ell)+1} \right)^2 \tag{3.16}$$

โดยมี TE เป็น

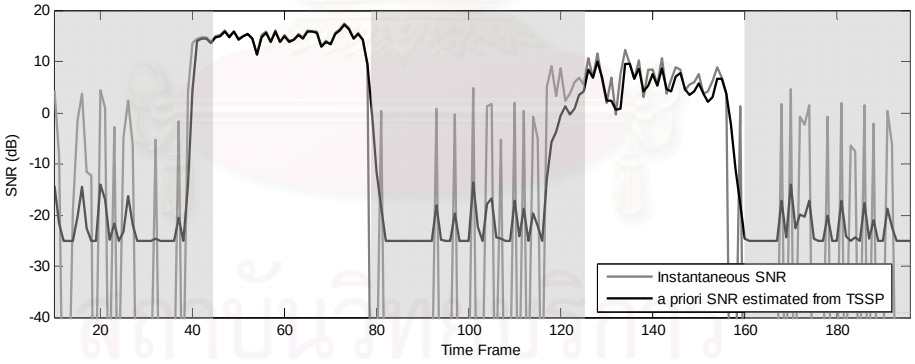
$$\hat{\xi}^N_{\mathrm{TSSP}}(k,\ell) = (1-\alpha_{\mathrm{DD}})\delta^N(k,\ell) + \left(\frac{(1-\alpha_{\mathrm{DD}})\delta^N(k,\ell)}{(1-\alpha_{\mathrm{DD}})\delta^N(k,\ell)+1} \right)^2 \alpha_{\mathrm{DD}}\delta^N(k,\ell) \tag{3.17}$$

โดย TE ในสมการที่ (3.17) ถูกแสดงเปรียบเทียบกับ TE ของวิธีการประมาณค่า a priori SNR แบบอื่นๆ ดังรูปที่ 3.6

จะเห็นว่า TSSP เริ่มให้ค่าประมาณ a priori SNR ที่สูงกว่าวิธี DD และวิธี TSW ที่ค่า Instantaneous SNR ที่ต่ำกว่า ทำให้สามารถทราบอย่างสังเขปว่า วิธี TSSP มีความไวในการติดตามค่า Instantaneous SNR มากยิ่งขึ้นกว่าวิธี DD และ TSW อย่างไรก็ตามเรื่องผลของเสียงรบกวนตกค้างแบบ Musical Noise สามารถสังเกตได้จากผลการทดลองที่อาศัยสัญญาณความถี่เดียวเช่นเดียวกับในรูปที่ 3.4 ของ TSSP ดังรูปที่ 3.7 กล่าวคือ ค่าประมาณ a priori SNR ในช่วงที่มีเสียงพูดของวิธี TSSP มีค่าใกล้เคียงกับของวิธี DD ดังนั้นจึงสามารถสรุปได้ว่า วิธี TSSP มีคุณสมบัติในการลดผลของ Musical Noise เช่นเดียวกับวิธี DD โดยยังสามารถติดตามค่า Instantaneous SNR ในช่วงที่มีเสียงพูดได้ดีกว่าวิธี DD และ TSW อีกด้วย โดยจากการทดลองฟังอย่างไม่เป็นทางการยืนยันผลการวิเคราะห์นี้



รูปที่ 3.6 เปรียบเทียบ Transition equation ของ TSSP เทียบกับการประมาณแบบอื่นๆ



รูปที่ 3.7 ลักษณะของค่าประมาณ a priori SNR ที่ได้จากการประมาณแบบ TSSP เมื่อจำลองสัญญาณเสียงที่ถูกรบกวนด้วยสัญญาณความถี่เดียวจำนวน 2 พัลส์สัญญาณ ณ ความถี่ 2000 Hz ที่ระดับ Global SNR 0 dB

สังเกตเห็นได้ว่าการประมาณค่า a priori SNR ที่ให้กราฟ TE ซึ่งลู่เข้าสู่กราฟ $\hat{\xi} = \delta$ ที่ค่า Instantaneous SNR ที่ต่ำเพียงใดจะยิ่งทำให้สามารถติดตามการเปลี่ยนแปลงค่า Instantaneous SNR ได้ดียิ่งขึ้นเท่านั้น แต่อย่างไรก็ตามหากมีความสามารถในการติดตามการเปลี่ยนแปลงที่ดีเกินไปก็อาจนำไปสู่ปัญหาเสียงรบกวนตกค้างแบบ Musical Noise ทั้งนี้ปัญหา Musical Noise สามารถสังเกตผลได้จากการทดลองโดยใช้สัญญาณความถี่เดียวดังในรูปที่ 3.4 และรูปที่ 3.7 นอกจากนี้เนื่องจากผลการทดลองได้ชี้ให้เห็นแล้วว่าวิธี SAAF ทำให้เกิดผลของเสียง

รบกวนตกค้างแบบ Musical Noise ดังนั้นจึงอาจใช้กราฟ TE ของวิธี SAAF เป็นตัวบ่งชี้อย่างสังเขปว่าวิธีการประมาณ a priori SNR ที่พิจารณาอยู่นั้นจะก่อให้เกิดเสียงรบกวนตกค้างแบบ Musical Noise หรือไม่โดยอาศัยเพียงการเปรียบเทียบ TE ของแต่ละวิธีเท่านั้นก็ได้

3.2.2.2. รูปแบบปรับปรุงของ TSW และ TSSP

หากอาศัยค่า K_g ตามสมการที่ (3.16) จะสามารถเขียนสมการของวิธี TSW ได้เป็น

$$\hat{\xi}^N_{\mathrm{TSW}}(k,\ell) = K_g\,\gamma(k,\ell) \tag{3.18}$$

ทำการปรับปรุงค่า K_g ในสมการที่ (3.15) และ (3.18) จะทำให้ได้การประมาณค่า a priori SNR ซึ่งจะถูกเรียกว่า Modified TSW (MTSW) และ Modified TSSP (MTSSP) ตามลำดับดังนี้

$$\hat{\xi}^N_{\mathrm{MTSSP}}(k,\ell) = \hat{\xi}^N_{\mathrm{DD}}(k,\ell) + M_g\left(\delta^N(k,\ell) - \hat{\xi}^N_{\mathrm{DD}}(k,\ell)\right) \tag{3.19}$$

$$\hat{\xi}^N_{\mathrm{MTSW}}(k,\ell) = M_g\,\gamma(k,\ell) \tag{3.20}$$

เมื่อ

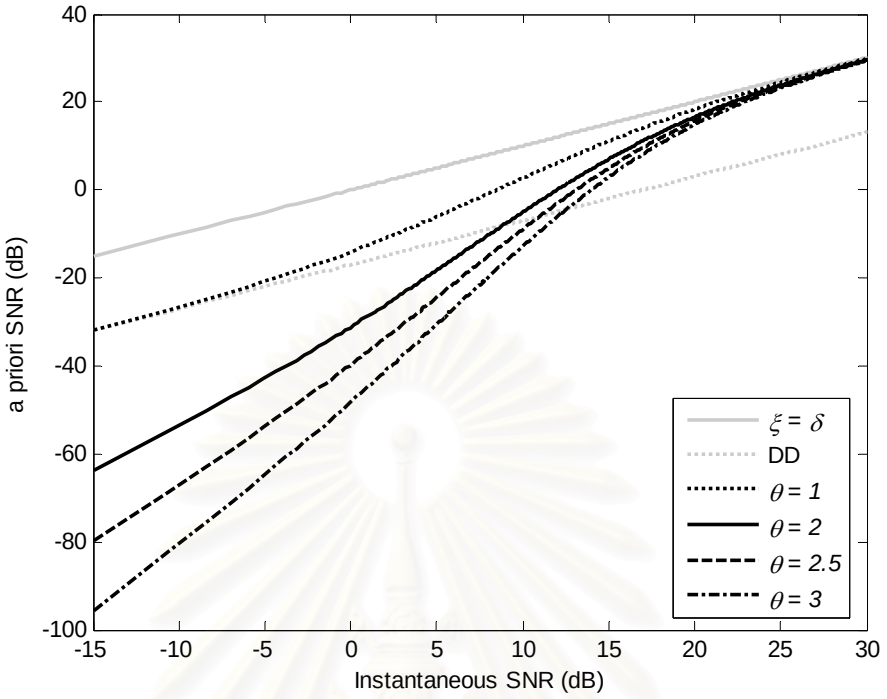
$$M_g = \left(\frac{\hat{\xi}^N_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^N_{\mathrm{DD}}(k,\ell)+\beta}\right)^{\theta} \tag{3.21}$$

โดย TE ของวิธี MTSSP และ MTSW สามารถเขียนได้เป็น

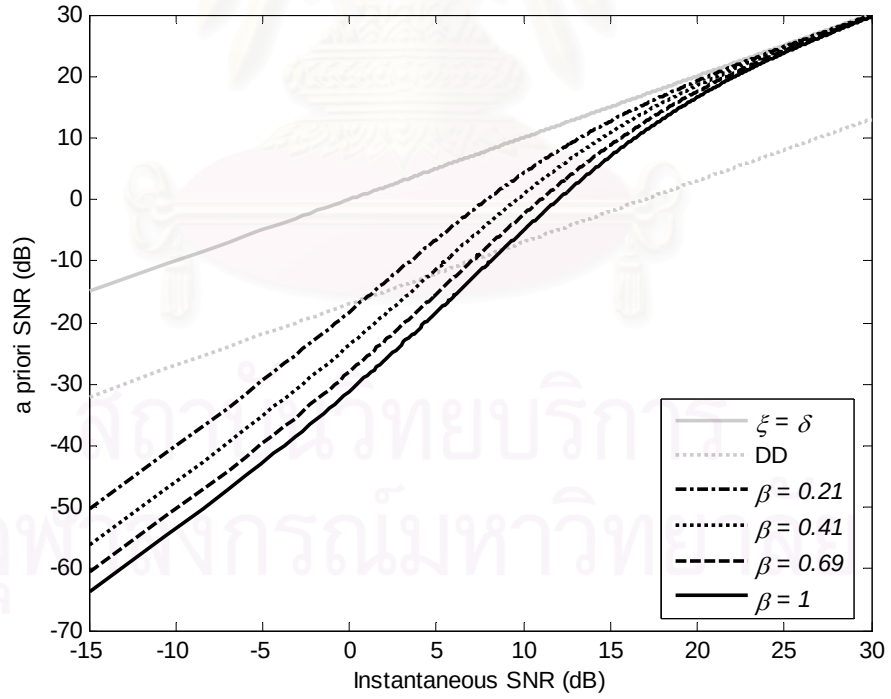
$$\hat{\xi}^N_{\mathrm{MTSSP}}(k,\ell) = (1-\alpha_{\mathrm{DD}})\delta^N(k,\ell) + \left(\frac{(1-\alpha_{\mathrm{DD}})\delta^N(k,\ell)}{(1-\alpha_{\mathrm{DD}})\delta^N(k,\ell)+\beta}\right)^{\theta}\alpha_{\mathrm{DD}}\delta^N(k,\ell) \tag{3.22}$$

$$\hat{\xi}_{\mathrm{MTSW}}(k,\ell) = \left(\frac{(1-\alpha_{DD})\delta(k,\ell)}{(1-\alpha_{DD})\delta(k,\ell)+\beta}\right)^{\theta}\gamma(k,\ell) \tag{3.23}$$

ทั้งนี้การเปลี่ยนแปลงค่า β และ θ จะทำให้กราฟ TE ของการประมาณค่า a priori SNR ทั้งสองเปลี่ยนแปลงไป ดังแสดงตัวอย่างในรูปที่ 3.8

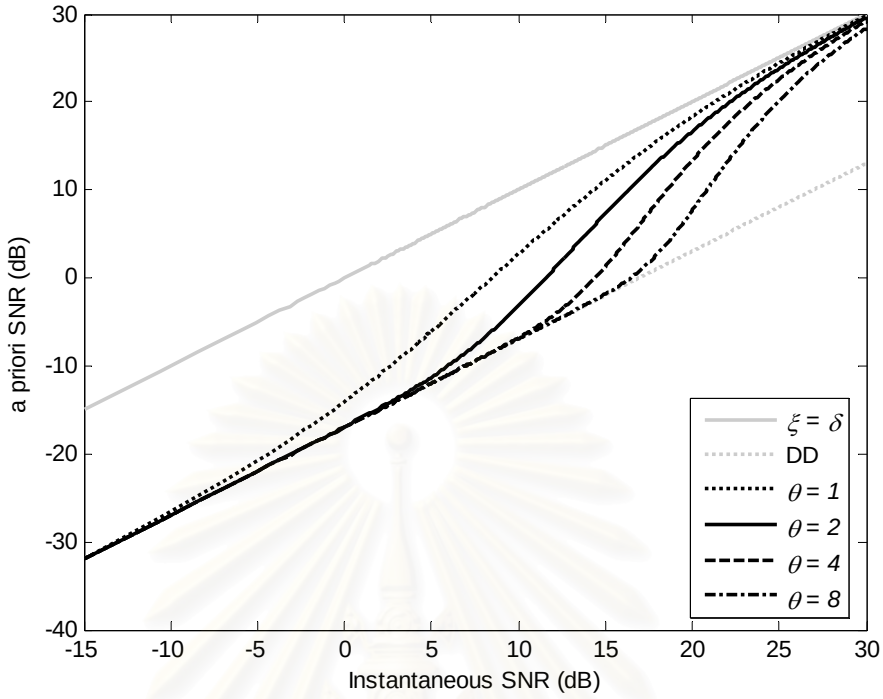


(ก)

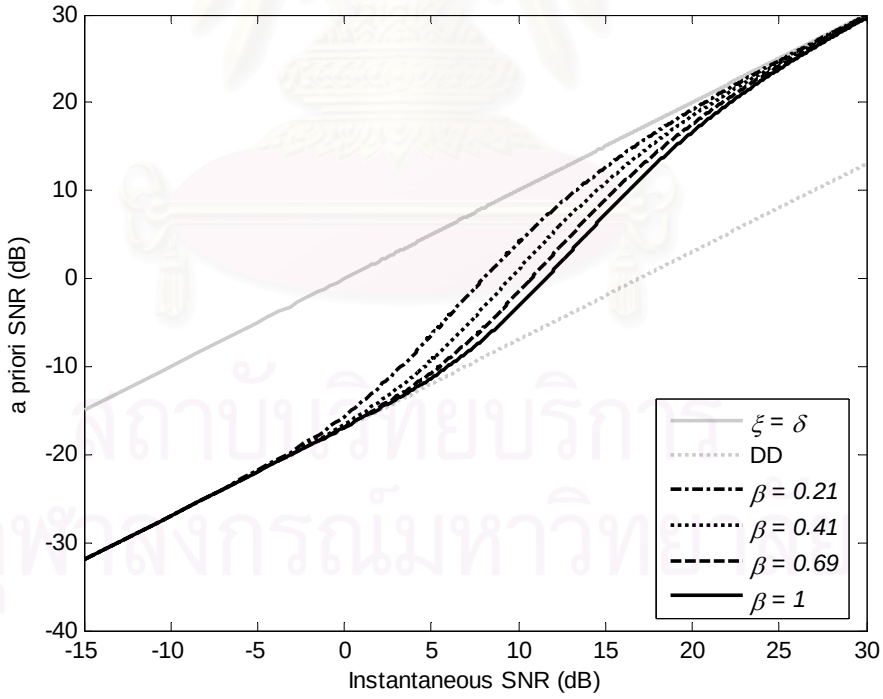


(ข)

รูปที่ 3.8 Transition equation ของ
(ก) MTSW เมื่อคงค่า $\beta = 1$ และปรับเปลี่ยนค่า θ และ (ข) MTSW เมื่อคงค่า $\theta = 2$ และปรับเปลี่ยนค่า β



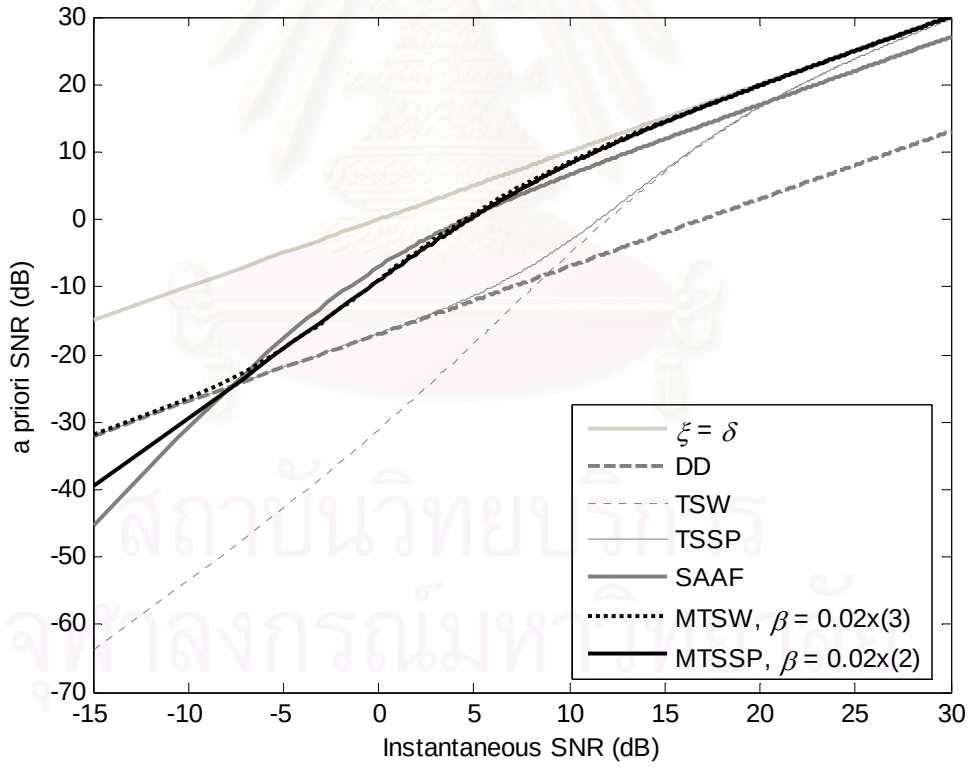
(ก)



(ข)

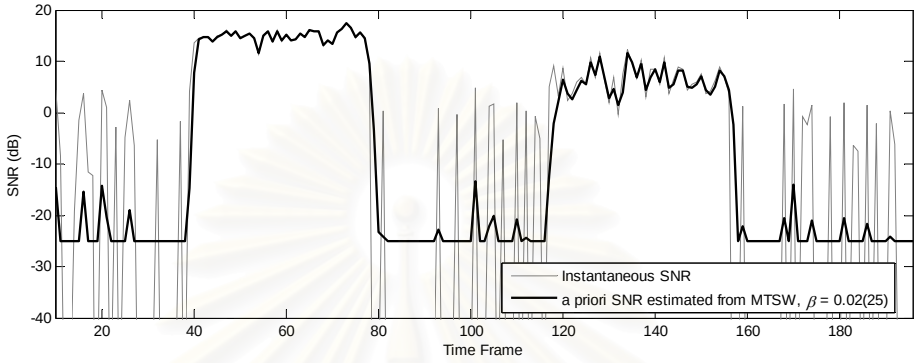
รูปที่ 3.9 Transition equation ของ
(ก) MTSSP เมื่อคงค่า $\beta = 1$ และปรับเปลี่ยนค่า θ และ (ข) MTSSP เมื่อคงค่า $\theta = 2$ และปรับเปลี่ยนค่า β

สังเกตได้ว่าผลที่เกิดจากการปรับเปลี่ยนค่า β จะทำให้จุดที่ค่าประมาณ a priori SNR เริ่มมีค่ามากกว่าค่าประมาณ a priori SNR ที่ได้จากวิธี DD มีค่าเปลี่ยนแปลงไป ในขณะที่การปรับเปลี่ยนค่า θ จะทำให้เกิดผลเช่นเดียวกับการปรับเปลี่ยนค่า β รวมทั้งยังส่งผลให้ TE มีอัตราการลู่เข้าสู่ กราฟ $\hat{\xi} = \delta$ ที่เร็วยิ่งขึ้น ดังนั้นจะเห็นว่าการปรับค่า θ เพื่อให้ได้ผลตามที่ต้องการ จะกระทำได้ยากกว่าการปรับค่า β เนื่องจากมีการเปลี่ยนแปลงเกิดขึ้นสองปรากฏการณ์ ดังนั้นในวิทยานิพนธ์ฉบับนี้จะเลือกใช้ค่า $\theta = 2$ กล่าวคือไม่เปลี่ยนแปลงค่าดังกล่าวจากวิธี TSW และ TSSP แต่จะทำการเปลี่ยนเฉพาะค่า β เท่านั้น จุดประสงค์ของการปรับปรุงวิธีการประมาณค่า a priori SNR คือต้องการเพิ่มความสามารถในการติดตามค่า Instantaneous SNR ในแต่ละวิธีการให้มากยิ่งขึ้น โดยที่ยังคงไม่ก่อให้เกิดผลของเสียงรบกวนตกค้างแบบ Musical Noise ดังนั้นจากแนวโน้มในรูปที่ 3.8 ค่า β ควรถูกเลือกให้มีขนาดน้อยลงกว่า 1 (หรือ 0.02×50) ทั้งนี้เพื่อให้กราฟ TE ของการประมาณทั้งสองแบบ ลู่เข้าสู่กราฟ $\hat{\xi} = \delta$ ที่ค่า Instantaneous SNR ที่ต่ำลงนั่นเอง นอกจากนี้พบว่าหากเลือกค่า $\beta = 0.02 \times 3$ และ $\beta = 0.02 \times 2$ สำหรับการประมาณ MTSW และ MTSSP ตามลำดับแล้ว กราฟ TE ของการประมาณทั้งสองแบบจะมีความใกล้เคียงกับ กราฟ TE ของวิธี SAAF ดังนั้นจึงกล่าวได้ว่าค่า β ดังกล่าวเป็นค่าต่ำสุด เพราะการเลือกค่า β ที่ต่ำกว่าค่าดังกล่าวนี้จะทำให้ TE ของการประมาณแต่ละแบบลู่เข้าสู่ กราฟ $\hat{\xi} = \delta$ ที่ค่า Insatantaneous SNR ที่ต่ำลงและจะนำมาซึ่งเสียงรบกวนตกค้างแบบ Musical Noise

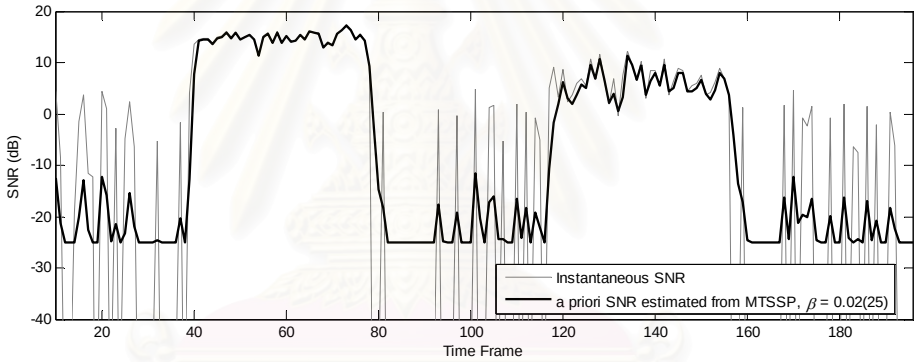


รูปที่ 3.10 เปรียบเทียบ Transition equation ระหว่าง
วิธี MTSW และวิธี MTSSP ที่ใช้ $\beta = 0.02 \times 3$ และ $\beta = 0.02 \times 2$ ตามลำดับ กับวิธี SAAF

จากผลการทดลองด้วยสถานการณ์จำลองดังในรูปที่ 3.4 และรูปที่ 3.7 พบว่าค่า β ที่เหมาะสมที่ยังคงไม่ก่อให้เกิดผลของเสียงรบกวนตกค้างแบบ Musical Noise ได้แก่ $\beta = 0.02 \times 25$ สำหรับทั้งสองวิธี โดยผลการทดลองถูกแสดงไว้ในรูปที่ 3.11 และ TE ของทั้งสองถูกแสดงไว้ในรูปที่ 3.12 โดยสามารถเห็นได้อย่างชัดเจนว่าวิธีการประมาณค่า a priori SNR ที่นำเสนอให้ผลการติดตามค่า Instantaneous SNR ได้ดียิ่งขึ้นในช่วงที่มีเสียงพูด ในขณะที่ยังคงรักษาระดับ a priori SNR ที่ไม่สูงและราบเรียบในช่วงที่ไม่มีเสียงพูดเอาไว้ได้

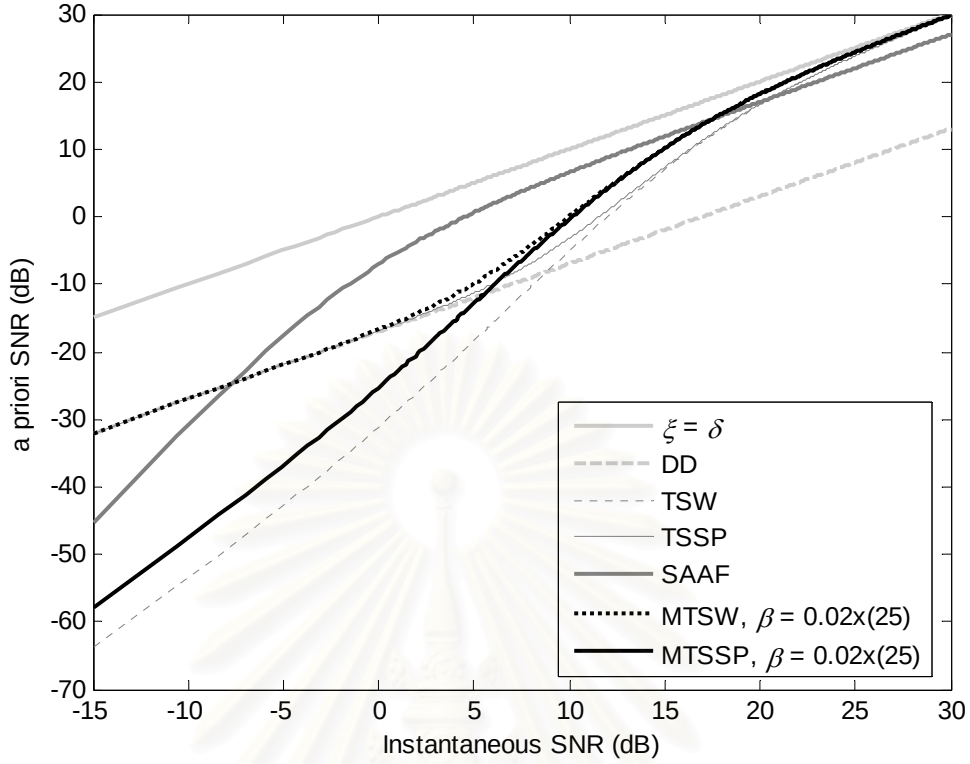


(ก)



(ข)

รูปที่ 3.11 ลักษณะของค่าประมาณ a priori SNR ที่ได้จากการประมาณแบบ
(ก) MTSW ที่ใช้ $\beta = 0.02 \times 25$ (ข) MTSSP ที่ใช้ $\beta = 0.02 \times 25$ เมื่อจำลองสัญญาณเสียงที่ถูกรบกวนด้วยสัญญาณความถี่เดียวจำนวน 2 พัลส์สัญญาณ ณ ความถี่ 2000 Hz ที่ระดับ Global SNR 0 dB



รูปที่ 3.12 เปรียบเทียบ TE ของการประมาณค่า a priori SNR แบบต่างๆ

ข้อสังเกตที่น่าสนใจเป็นอย่างยิ่งประการหนึ่งเกิดขึ้นเมื่อทดลองใช้การประมาณค่า a priori SNR ที่นำเสนอในกรณีของการลดเสียงรบกวนออกจากสัญญาณเสียงพูดพบว่า สัญญาณเสียงพูดที่ถูกปรับปรุงเกิดเสียงรบกวนตกค้างแบบ Musical Noise ขึ้น ซึ่งเสียงรบกวนตกค้างดังกล่าวไม่สามารถได้ยินในการทดลองที่อาศัยสัญญาณความถี่เดียว จากการวิเคราะห์โดยดูจากสเปกโทรแกรมของเสียงพูดที่ถูกปรับปรุงโดยใช้การประมาณค่า a priori SNR ที่นำเสนอดังรูปที่ 3.13 พบว่า เสียงรบกวนตกค้างแบบ Musical Noise ดังกล่าวไม่ได้เกิดจากการที่ไม่สามารถกดเสียงรบกวนลงได้หมด แต่เกิดจากการที่รักษาส่วนของเสียงพูดเอาไว้ได้อย่างไม่ต่อเนื่อง ที่สามารถทำการสรุปเช่นนี้ได้เป็นเพราะสามารถสังเกตได้ว่าในช่วงที่ไม่มีเสียงพูด (เห็นได้ชัดในช่วงต้นที่ยังไม่มีเสียงพูด) นั้นจะไม่ปรากฏเสียงรบกวนตกค้างแบบ Musical Noise แต่เสียงรบกวนตกค้างดังกล่าวจะเกิดขึ้นเฉพาะช่วงทีมีเสียงพูดอยู่เท่านั้น กล่าวคือ เสียง Musical Noise ที่ได้ยินเกิดมาจากส่วนของเสียงพูดเอง การที่ไม่สามารถรักษาส่วนของเสียงพูดไว้ได้อย่างต่อเนื่องนี้เป็นเพราะส่วนของเสียงพูดที่องค์ประกอบทางความถี่สูงมีพลังงานค่อนข้างน้อย เมื่อถูกสัญญาณเสียงรบกวนมาบดบังแล้วองค์ประกอบเหล่านั้นบางองค์ประกอบจะหายไปในขณะที่บางองค์ประกอบจะยังคงอยู่เป็นเหตุให้การประมาณค่า a priori SNR ที่ไวต่อการติดตามค่า Instantaneous SNR สามารถติดตามได้เฉพาะบางองค์ประกอบทางความถี่เท่านั้นจึงเกิดความไม่ต่อเนื่องขึ้นและนำมาซึ่งเสียงตกค้างแบบ Musical Noise การประมาณค่า a priori SNR ในแบบเดิมที่ค่อนข้างเชื่องช้าต่อค่า Instantaneous SNR จะไม่สามารถติดตามค่า Instantaneous SNR ได้ทันแม้ในองค์ประกอบทางความถี่ที่เสียงพูดยังไม่ถูกบดบัง (แต่มีค่าน้อย) ดังนั้นวิธีการประมาณค่า a priori SNR เหล่านั้นจึงไม่ได้รับผลจาก เสียงตกค้างแบบ Musical Noise ชนิดนี้ การแก้ปัญหาเสียงตกค้างดังกล่าวจึงอาจทำได้ โดยเลือกใช้ค่า $\beta(k)$ ที่มีค่าน้อยกว่า 1 ในเฉพาะองค์ประกอบทาง

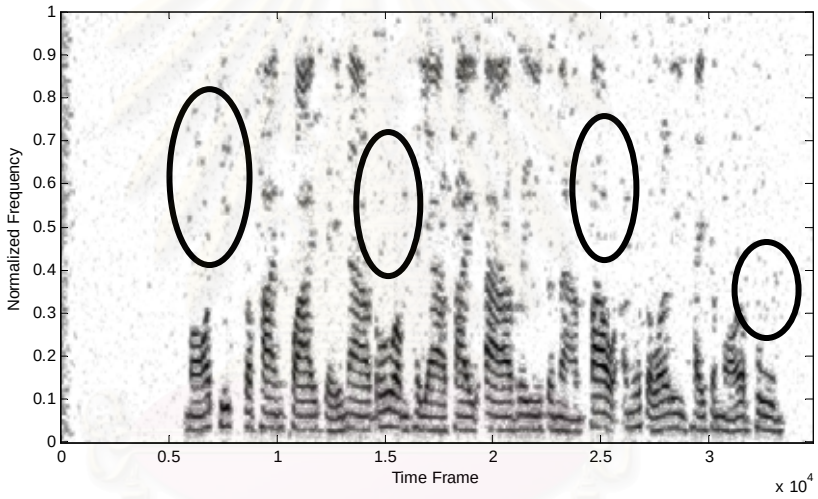
ความถี่ต่ำ ทั้งนี้เนื่องจาก ส่วนของเสียงพูด ณ องค์ประกอบทางความถี่ต่ำจะมีพลังงานที่ค่อนข้างสูงและไม่ถูกบดบังด้วยเสียงรบกวนจึงยังคงความต่อเนื่องอยู่ตลอด วิทยานิพนธ์ฉบับนี้เสนอให้ใช้ ค่า

$$\beta_{\mathrm{N}}(k) = 0.02 \times \eta_{\mathrm{N}}(k) \tag{3.24}$$

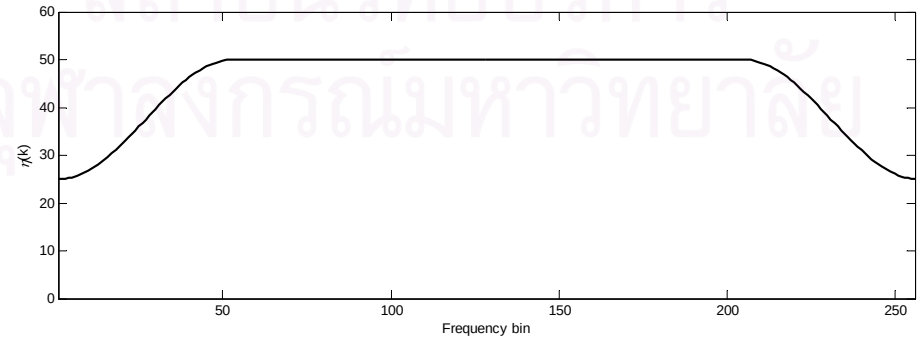
เมื่อ

$$\eta_{\mathrm{N}}(k) = \begin{cases} 50 - 25 \times \left(\cos(\pi \dfrac{k}{K_T}) + 1 \right), & k \le \left\lfloor 0.4 \times \dfrac{T}{2} \right\rfloor \\ 50, & \textit{otherwise} \\ 50 - 25 \times \left(\cos(\pi \dfrac{T-k}{K_T}) + 1 \right), & k > T - 1 - \left\lfloor 0.4 \times \dfrac{T}{2} \right\rfloor \end{cases} \tag{3.25}$$

ซึ่งสามารถแสดงค่า $\eta_{\mathrm{N}}(k)$ ได้ดังรูปที่ 3.14

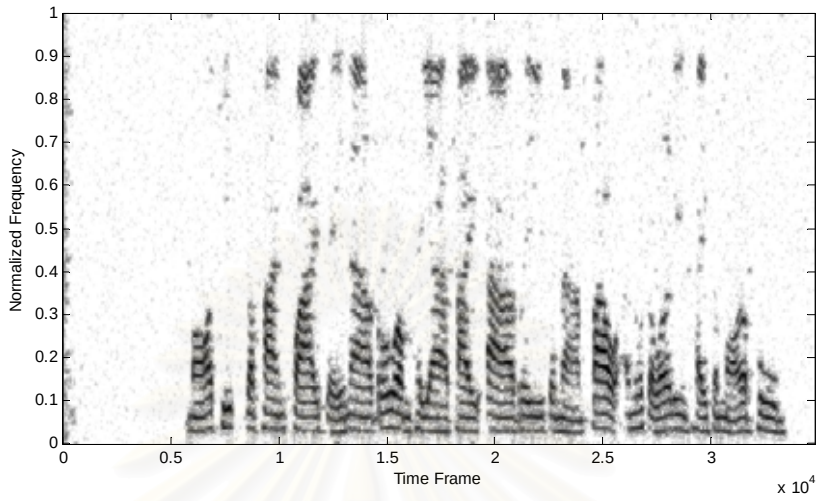


รูปที่ 3.13 สเปกโทรแกรมเสียงพูดที่ถูกปรับปรุง ที่อาศัย MTSW โดยเลือก $\beta = 0.02 \times 25$ หรือ $\eta_{\mathrm{N}}(k) = 25$



รูปที่ 3.14 ค่า $\eta_{\mathrm{N}}(k)$ สำหรับการประมาณ a prioir SNR ที่เสนอ

สเปกโทรแกรมของเสียงพูดที่ถูกปรับปรุงโดยอาศัยวิธี MTSW ที่ใช้ค่า $\beta_N(k)$ ในสมการที่ (3.24) กับสัญญาณเสียงพูดที่ถูกรบกวนที่ระดับ Global SNR 5 dB ถูกแสดงในรูปที่ 3.15

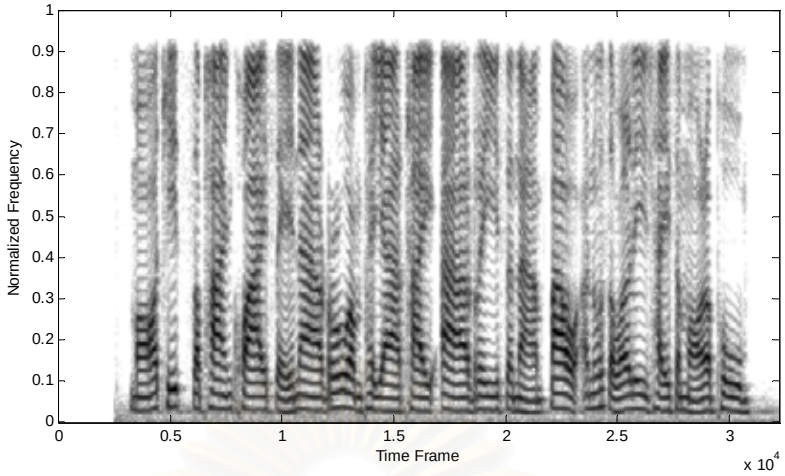


รูปที่ 3.15 สเปกโทรแกรมเสียงพูดที่ถูกปรับปรุง ที่อาศัย MTSW โดยเลือก $\beta_N(k)$ ตามสมการที่ (3.24)

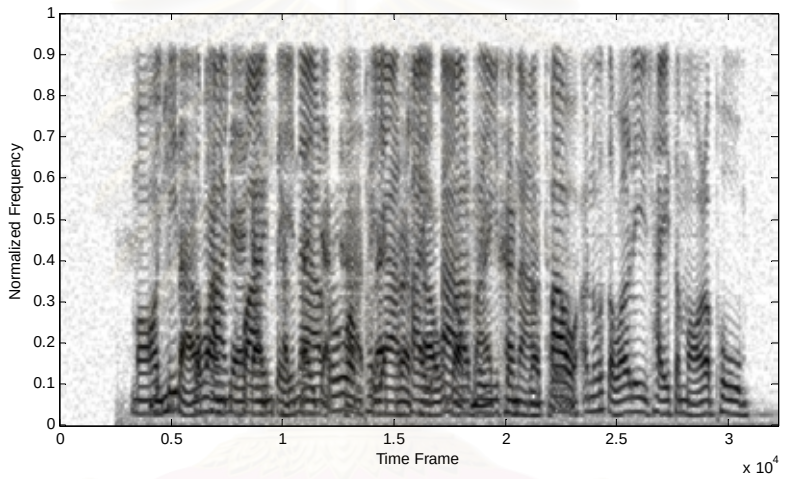
อย่างไรก็ตามเนื่องจากพลังงานของเสียงพูดที่องค์ประกอบทางความถี่ต่ำส่วนใหญ่มีค่าสูงเมื่อเทียบกับเสียงรบกวนอยู่แล้ว ดังนั้นไม่ว่าจะอาศัยวิธีการประมาณค่า a priori SNR แบบ TSW และ TSSP ธรรมดาหรือ วิธี MTSW และ MTSSP ที่ใช้ค่า $\beta_N(k)$ ตามสมการที่ (3.24) ก็จะให้ผลที่ไม่แตกต่างกันมากนักดังปรากฏในผลการทดลองในหัวข้อย่อยที่ 4.3 แต่อย่างไรก็ตามวิธี MTSW และ MTSSP ที่อาศัยค่า $\beta_N(k)$ ตามสมการที่ (3.24) จะมีประโยชน์เป็นอย่างยิ่งในการประมาณค่า a priori SER ดังจะกล่าวถึงต่อไป

3.2.3. การประมาณค่า a priori SER ที่นำเสนอ

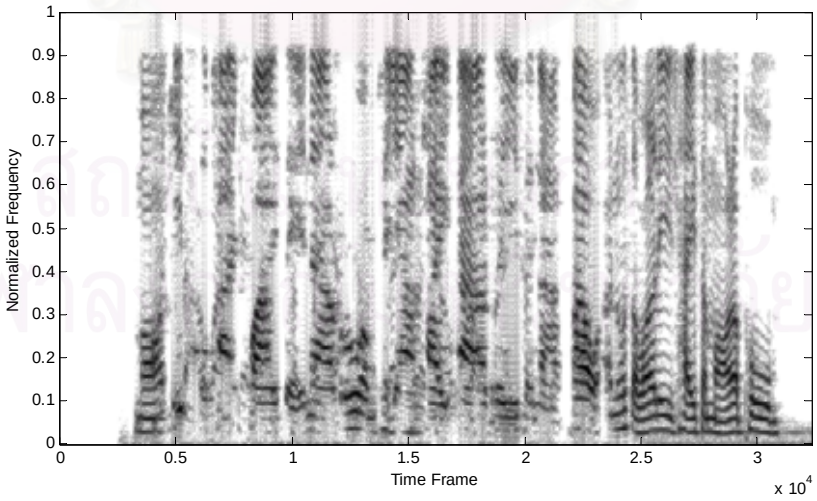
เนื่องจากเสียงสะท้อนมีลักษณะทางความถี่ที่คล้ายคลึงกับเสียงพูด ดังนั้นการกดเสียงสะท้อนจึงนำมาซึ่งการสูญเสียส่วนของเสียงพูดไปในลักษณะที่แตกต่างกับการกดเสียงรบกวน จากรูปที่ 3.16 จะเห็นว่าวิธีการกดเสียงสะท้อนจะนำมาซึ่งการสูญเสียส่วนของเสียงพูด ณ องค์ประกอบทางความถี่ต่ำเป็นส่วนใหญ่ ในขณะที่การกดเสียงรบกวนนำมาซึ่งการสูญเสียส่วนของเสียงพูด ณ องค์ประกอบทางความถี่สูง ดังรูปที่ 3.3 ทั้งนี้เป็นเพราะเสียงสะท้อนมีพลังงานสูงในช่วงองค์ประกอบความถี่ที่ต่ำ ทำให้ค่า Instantaneous SER ณ องค์ประกอบทางความถี่ที่ต่ำจึงมีค่าน้อย ดังนั้นการประมาณค่า a priori SER แบบเดิม ด้วยวิธี DD และ TSW จึงไม่สามารถติดตามค่า Instantaneous SER ดังกล่าวได้ และทำให้ไม่สามารถประมาณค่า a priori SER ได้อย่างเหมาะสม ส่งผลให้เกิดการกดมากเกินไปจนทำให้ส่วนของเสียงพูดในองค์ประกอบทางความถี่ต่ำนั้นสูญหายไป



(ก)



(ข)



(ค)

รูปที่ 3.16 สเปกโทรแกรมของ ก) เสียงพูดสะอาด ข) เสียงพูดที่ถูกก่อกวนด้วยเสียงสะท้อนที่ SNR 40 dB และ ค) เสียงพูดที่ถูกปรับปรุงจากวิธีการ AENS ใน [53]

SNR 40 dB จะทำให้ค่าประมาณ a priori SDR ได้รับผลมาจากส่วนของ a priori SER เพียงอย่างเดียว

เพราะฉะนั้นวิธีการประมาณค่า a priori SER จึงควรมีความไวในการติดตามค่า Instantaneous SNR โดยเฉพาะอย่างยิ่งในองค์ประกอบทางความถี่ต่ำ ดังนั้นวิทยานิพนธ์ฉบับนี้จึงเสนอให้ใช้วิธี MTSW และ MTSSP ดังนี้

$$\hat{\xi}^{E}_{\mathrm{MTSW}}(k,\ell) = M_g\,\gamma(k,\ell) \tag{3.26}$$

$$\hat{\xi}^{E}_{\mathrm{MTSSP}}(k,\ell) = \hat{\xi}^{E}_{\mathrm{DD}}(k,\ell) + M_g\left(\delta^{E}(k,\ell) - \hat{\xi}^{E}_{\mathrm{DD}}(k,\ell)\right) \tag{3.27}$$

เมื่อ

$$M_g = \left(\frac{\hat{\xi}^{E}_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^{E}_{\mathrm{DD}}(k,\ell) + \beta_{\mathrm{E}}}\right)^2 \tag{3.28}$$

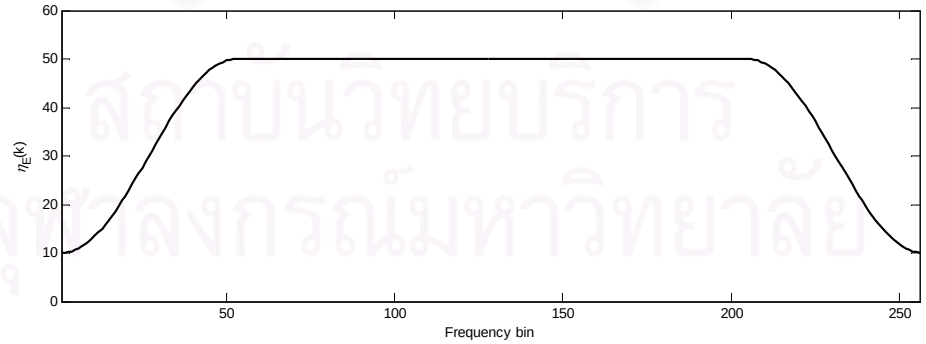
โดย

$$\beta_{\mathrm{E}}(k) = 0.02 \times \eta_{\mathrm{E}}(k) \tag{3.29}$$

เมื่อ

$$\eta_{\mathrm{E}}(k) = \begin{cases} 50 - 10\times\left(\cos(\pi\dfrac{k}{K_T}) + 1\right), & k \le \left\lfloor 0.4\times\dfrac{T}{2}\right\rfloor \\ 50, & otherwise \\ 50 - 10\times\left(\cos(\pi\dfrac{T-k}{K_T}) + 1\right), & k > T - 1 - \left\lfloor 0.4\times\dfrac{T}{2}\right\rfloor \end{cases} \tag{3.30}$$

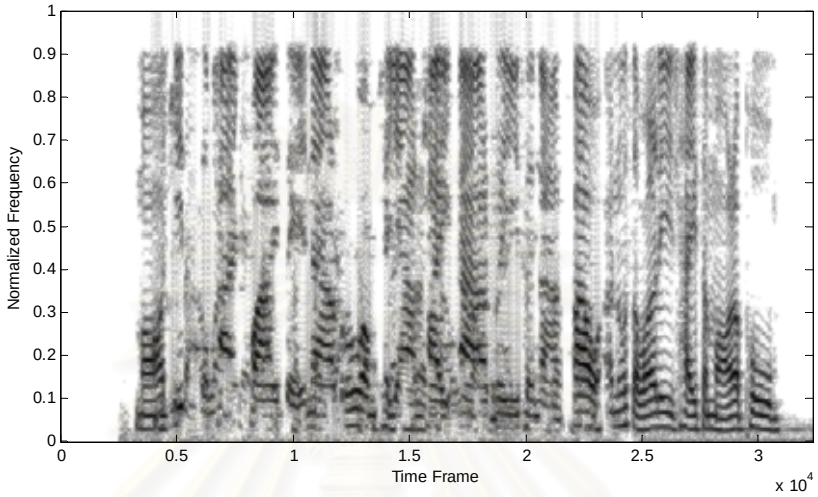
โดยค่า $\eta_{\mathrm{E}}(k)$ ตามสมการที่ (3.30) ถูกแสดงในรูปที่ 3.17



รูปที่ 3.17 ค่า $\eta_{\mathrm{E}}(k)$ สำหรับการประมาณ a prioir SER ที่เสนอ

โดยรูปที่ 3.18 ถูกใช้แสดงสเปกโทรแกรมของสัญญาณเสียงพูดที่ถูกปรับปรุงด้วย AENS ใน [53] โดยเปลี่ยนวิธีการประมาณค่า a priori SER เป็นวิธี MTSW ที่นำเสนอ ซึ่งสามารถสังเกตเห็นได้อย่างชัดเจนว่า วิธีการ

ประมาณค่า a priori SER ที่นำเสนอสามารถรักษาส่วนของเสียงพูดในบางองค์ประกอบทางความถี่ต่ำ ไว้ได้มากยิ่งขึ้น นำมาซึ่งการลดความผิดเพี้ยนของสัญญาณเสียงพูดที่ถูกปรับปรุงลงได้ ทั้งนี้เมื่อใช้วิธีการประมาณค่า a priori SER ที่นำเสนอกอปรกับการประมาณค่า EPSD ที่นำเสนอจะสามารถช่วยลดผลของความผิดเพี้ยนของเสียงพูดที่ถูกปรับปรุงลงไปได้อีก ดังเห็นได้จากผลการทดลองในบทที่ 4



รูปที่ 3.18 สเปกโทรแกรมของเสียงพูดที่ถูกปรับปรุงด้วยวิธีการ AENS ใน [53] โดยเปลี่ยนวิธีการประมาณค่า a priori SER เป็น MTSSP ที่นำเสนอ จากเสียงพูดที่ถูกก่อกวนด้วยเสียงสะท้อนที่ SNR 40 dB

3.2.4. การประมาณค่า a priori SDR จากค่าประมาณ a priori SNR และค่าประมาณ a priori SER ที่นำเสนอ

หลังจากได้ค่าประมาณ a priori SNR $\hat{\xi}^N(k,\ell)$ และค่าประมาณ a priori SER $\hat{\xi}^E(k,\ell)$ ตามกระบวนการที่นำเสนอแล้ว ค่าประมาณ a priori SDR $\hat{\xi}^D(k,\ell)$ สามารถถูกประมาณได้โดยอาศัยสมการที่ (2.115) ซึ่งถูกนำมาเขียนใหม่ดังนี้

$$\hat{\xi}^D(k,\ell) = \frac{1}{1/\hat{\xi}^E(k,\ell) + 1/\hat{\xi}^N(k,\ell)}$$

3.3. การเปรียบเทียบความซับซ้อนในการคำนวณ

ความซับซ้อนในการคำนวณสำหรับขั้นตอนวิธีการต่างๆ ตลอดทั้งวิทยานิพนธ์ฉบับนี้จะพิจารณาจากการคูณจำนวนจริงต่อหนึ่งรอบการชักตัวอย่าง (Real Multiplication per Sample, RMPS) โดยทำการประมาณว่าการหารจำนวนจริงหนึ่งครั้งจะใช้ความซับซ้อนในการคำนวณเท่ากับ 16 RMPS และการแปลง DFT ซึ่งกระทำโดยใช้กระบวนการ FFT แบบ Radix-2 ซึ่งมีความซับซ้อนในการคำนวณต่อการทำหนึ่งครั้งเท่ากับ $N\log_2 N$ RMPS [38] เมื่อ N คือจำนวนตัวอย่างที่ใช้ในการทำการแปลงหนึ่งครั้ง

3.3.1. ความซับซ้อนในการคำนวณของ NS

จากตารางที่ 3.2 จะเห็นว่าความซับซ้อนในการคำนวณของวิธี NS นั้นขึ้นอยู่กับขั้นตอนวิธีที่เลือกใช้ด้วยโดยหากเลือกใช้ จำนวนตัวอย่างที่ใช้ในการแปลง $T = 256$ และช่วงก้าวระหว่างเฟรม $M = 128$ และเลือกวิธีการต่างๆ ดังตารางที่ 3.2 แล้วจะได้ว่าความซับซ้อนในกาคำนวณของวิธี NS จะมีค่าประมาณ 130 RMPS

ตารางที่ 3.2 ความซับซ้อนในการคำนวณของ NS

ขั้นตอน	วิธีที่ใช้	ความซับซ้อนในการคำนวณ
STFT	-	$\frac{2T\left(1+\log_2(T)\right)}{M}$
Power & Phase Computation	-	$\frac{9T}{M}$
NPSD Estimation	เฉลี่ยในช่วงที่ไม่มีเสียงพูด	$\frac{T}{M}$
Posterior SNR Estimation	-	$\frac{8T}{M}$
A priori SNR Estimation	DD	$\frac{20T}{M}$
Gain Computation & Multiplication	Wiener Gain	$\frac{18T}{M}$
Total		$\frac{2T\left(1+\log_2(T)\right)+56T}{M}$

ทั้งนี้การคำนวณดังกล่าวในข้างต้นไม่ได้รวมส่วนความซับซ้อนในการคำนวณของ VAD ซึ่งใช้ในการตรวจหาช่วงที่มีเสียงพูดในขั้นตอนการประมาณค่า NPSD

3.3.2. ความซับซ้อนในการคำนวณของ AEC

ในหัวข้อย่อยนี้จะทำการพิจารณาความซับซ้อนในการคำนวณของระบบ AER ที่อาศัยเทคนิค 2 ประเภทเปรียบเทียบกัน ได้แก่ AEC และ AES

ตารางที่ 3.3 ความซับซ้อนในการคำนวณของ AEC และ AES

ขั้นตอนวิธี		ความซับซ้อนในการคำนวณ	ความซับซ้อนในการคำนวณเมื่อ $L = 256$ (RMPS)
AEC	NLMS	$2L + 19$	531
	VSNLMS	$2L + 19 + 4N$	931 (เลือก $N = 100$)
AES	STSES (NLMS)	$\text{STFT} + \frac{4TP}{M} + \frac{4T}{M}$	60
	PAES	$\text{STFT} + \frac{2IP}{M} + \frac{4T}{M}$	45

เมื่อ I คือจำนวนซองสเปกตรัมที่เลือกใช้ และ

P คือจำนวนสัมประสิทธิ์ที่ใช้ในแต่ละองค์ประกอบทางความถี่

โดยการคำนวณด้านบนเลือกใช้ $I = 17$ และ $P = 2$

AEC เป็นวิธีการที่การทำในโดเมนเวลาดังนั้นจึงไม่มีการแปลง STFT แต่สำหรับวิธีการกดทางสเปกตรัมแล้วต้องมีการรวมความซับซ้อนในการคำนวณของ STFT เข้าไว้ด้วย ในตารางที่ 3.3 ได้นำความซับซ้อนในการคำนวณของ AES 2 เทคนิคมาเปรียบเทียบกับ AEC เพื่อให้เห็นถึงความซับซ้อนในการคำนวณที่แตกต่างกันมากอีกด้วย

3.3.3. ความซับซ้อนในการคำนวณของระบบร่วม AECNS และ AENS

ความซับซ้อนในการคำนวณของระบบร่วม AECNS สามารถหาได้จากการนำความซับซ้อนในการคำนวณของ AEC และ NS ที่เลือกใช้มาบวกกัน

ความซับซ้อนในการคำนวณของ AENS ใน [53] สามารถหาได้โดยพิจารณาคล้ายกับในกรณีของ NS กล่าวคือจะประกอบด้วยขั้นตอนต่างๆ ดังนี้

1. STFT
2. การแปลงจากจำนวนเชิงซ้อนให้เป็นกำลังและเฟส
3. การประมาณค่าความหนาแน่นสเปกตรัมกำลังเสียงก่อกวนแต่ละชนิด
4. การประมาณค่า a priori SDR, posterior SER และ posterior SNR
5. การคำนวณค่า Spectral Gain และการคูณ Spectral Gain ในแต่ละองค์ประกอบทางความถี่

ตารางที่ 3.4 ความซับซ้อนในการคำนวณของ AENS ใน [53]

ขั้นตอน	วิธีที่ใช้	ความซับซ้อนในการคำนวณ
STFT	-	$\frac{3T\left(1+\log_2(T)\right)}{M}$
Power & Phase Computation	-	$\frac{18T}{M}$
NPSD Estimation	เฉลี่ยในช่วงที่ไม่มีเสียงพูด	$\frac{T}{M}$
EPSD Estimation	CFM	$\frac{11T}{M}$
Posterior SNR & Posterior SER Estimation	-	$\frac{16T}{M}$
A priori SNR Estimation	DD	$\frac{64T}{M}$
Gain Computation & Multiplication	Wiener Gain	$\frac{18T}{M}$
Total		$\frac{2T\left(1+\log_2(T)\right)+128T}{M}$

ความซับซ้อนในการคำนวณของขั้นตอน STFT และการแปลงจากจำนวนเชิงซ้อนไปเป็นกำลังและสเปกตรัมเพิ่มมากขึ้นเนื่องจาก AENS มีการกระทำขั้นตอนดังกล่าวกับสัญญาณทางห้องไกล $f(t)$ ด้วย (ในขณะที่ NS ไม่มีการพิจารณาถึงสัญญาณทางห้องไกล) เมื่อใช้จำนวนตัวอย่างในการแปลง $T = 256$ และช่วงก้าวระหว่างเฟรม $M = 128$ แล้ว AENS ใน [53] จะมีความซับซ้อนในการคำนวณอยู่ที่ 310 RMPS

AENS ที่นำเสนอสามารถคำนวณความซับซ้อนได้ในทำนองเดียวกันกับ AENS ใน [53] ดังตารางที่ 3.5

ตารางที่ 3.5 ความซับซ้อนในการคำนวณของ AENS ที่นำเสนอ

ขั้นตอน	วิธีที่ใช้	ความซับซ้อนในการคำนวณ
STFT	-	$\frac{3T\left(1+\log_2(T)\right)}{M}$
Power & Phase Computation	-	$\frac{18T}{M}$
NPSD Estimation	เฉลี่ยในช่วงที่ไม่มีเสียงพูด	$\frac{T}{M}$
EPSD Estimation	NLMS	$\frac{(P+9)T}{M}$
Posterior SNR & Posterior SER Estimation	-	$\frac{16T}{M}$
A priori SNR Estimation	TSSP, MTSW และ MTSSP	$\frac{(64+9)T}{M}$
Gain Computation & Multiplication	Wiener Gain	$\frac{18T}{M}$
Total		$\frac{2T\left(1+\log_2(T)\right)+(135+P)T}{M}$

เมื่อใช้จำนวนตัวอย่างในการแปลง $T=256$ และช่วงก้าวระหว่างเฟรม $M=128$ แล้ว AENS ที่นำเสนอ จะมีความซับซ้อนในการคำนวณอยู่ที่ 328 RMPS ซึ่งจะเห็นกว่ามากกว่า AENS ใน [53] อยู่เพียงเล็กน้อย

ตารางที่ 3.6 ถูกใช้เปรียบเทียบความซับซ้อนในการคำนวณระหว่างระบบร่วม AECNS ที่ถูกเลือกมาทำการทดลองเปรียบเทียบในบทที่ 4 ซึ่งได้แก่ AEC ที่ใช้ VSNLMS ต่อด้วย NS ที่ใช้การประมาณค่า a priori SNR ที่นำเสนอ โดยในหลักสุดท้ายของตารางเป็นค่าความซับซ้อนในการคำนวณเมื่อให้ $T=256$, $M=128$, $L=256$ และ $P=\left\lfloor \frac{L}{M} \right\rfloor = 2$

ตารางที่ 3.6 ความซับซ้อนในการคำนวณของ AECNS และ AENS

	วิธีที่ใช้	ความซับซ้อนในการคำนวณ	RMPS
AECNS	VSNLMS + NS	$2L+19+4N+\frac{2T\left(1+\log_2(T)\right)+56T}{M}$	1060
AENS	ใน [53]	$\frac{2T\left(1+\log_2(T)\right)+128T}{M}$	310
	ที่นำเสนอ	$\frac{2T\left(1+\log_2(T)\right)+(135+P)T}{M}$	328

บทที่ 4

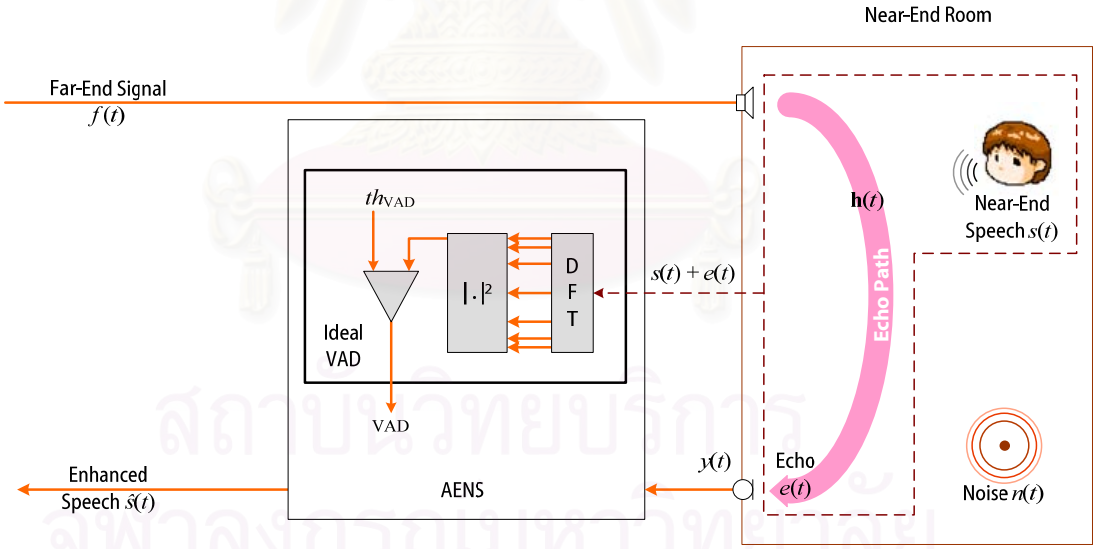
ผลการทดลองและการวิเคราะห์ผล

ในบทนี้จะกล่าวถึงขั้นตอนต่างๆ ที่ถูกใช้ในทางปฏิบัติ ได้แก่ ตัวตรวจหาเสียงพูด (Voice Activity Detector, VAD) ตัวตรวจหาสถานการณ์ดับเบิทอล์ก (Double-Talk Detector, DTD) รวมถึงตัวชี้วัดประสิทธิภาพที่ถูกใช้ในการประเมินผลการทดลอง ตามด้วยการทดลอง NS และ AENS ผลการทดลอง และวิเคราะห์การทดลอง

4.1. การตรวจหาเสียงพูดและการตรวจหาสถานการณ์ดับเบิลทอล์ก

4.1.1. VAD ในอุดมคติ

หน้าที่หลักของ VAD คือทำการตัวตรวจหาว่าเสียงพูดมีอยู่หรือไม่ ณ ขณะเวลา หรือเฟรมเวลานั้นๆ เนื่องจากหน้าที่ของ VAD มีความสำคัญและสามารถนำไปประยุกต์ใช้กับงานประยุกต์ประเภทต่างๆ ได้อย่างกว้างขวาง (ไม่เฉพาะการลดเสียงรบกวน) VAD จึงได้รับการพัฒนาจากนักวิจัยหลายกลุ่มอาทิเช่น [6] และ [24] เป็นต้น อย่างไรก็ตามในวิทยานิพนธ์ฉบับนี้ไม่ได้ให้ความสนใจเป็นพิเศษในเรื่องของการพัฒนา VAD หากแต่อาศัย VAD ในอุดมคติซึ่งมีวิธีการทำงานดังรูปที่ 4.1 ในทุกๆ การทดลอง



รูปที่ 4.1 การทำงานของ VAD ในอุดมคติ

สังเกตในการจำลองสถานการณ์โดยคอมพิวเตอร์ สเปกตรัมเสียงทางห้องใกล้ $S(k,\ell)$ และสเปกตรัมเสียงสะท้อน $E(k,\ell)$ สามารถหามาได้ ทำให้การเปรียบเทียบผลรวมพลังงานสเปกตรัมของเสียงทั้งสองเพื่อตัดสินว่าเฟรมเวลา ℓ นั้นๆ มีเสียงพูดอยู่หรือไม่จึงเป็นไปได้ดังนี้

$$\text{VAD} = \begin{cases} 1 & ,\varsigma(\ell) \geq th_{\text{VAD}} \\ 0 & ,\varsigma(\ell) < th_{\text{VAD}} \end{cases} \tag{4.1}$$

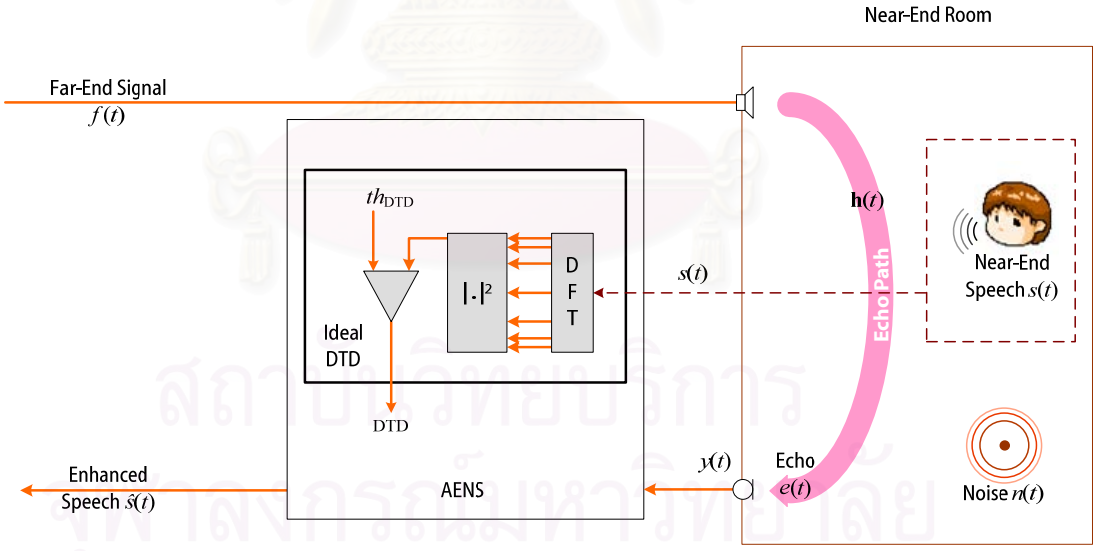
เมื่อ

$$\varsigma(\ell) = \sum_{k} \left(|S(k,\ell)|^2 + |E(k,\ell)|^2 \right) \tag{4.2}$$

ค่า th_{VAD} ได้มาจากการทดลองซึ่งในวิทยานิพนธ์ฉบับนี้ใช้ $th_{\text{VAD}} = 2 \times 10^{-4}$ โดยที่ $\text{VAD} = 1$ หมายถึงเฟรมเวลานั้นๆ ถูกตัดสินว่าเป็นช่วงที่มีเสียงพูด (Speech activity period) และ $\text{VAD} = 0$ หมายถึงเฟรมเวลานั้นๆ ถูกตัดสินว่าไม่มีเสียงพูดอยู่ (Non-speech activity period)

4.1.2. DTD ในอุดมคติ

DTD มีหน้าที่ในการตรวจหาสถานการณ์ดับเบิลทอล์ก (DTS) ซึ่งคือสถานการณ์ที่ ผู้พูดทั้งทางห้องไกลและห้องใกล้พูดขึ้นพร้อมๆ กัน ทั้งนี้เนื่องมาจากว่าใน DTS วงจรกรองแบบปรับตัวของระบบ AEC จะลู่ออกจากสถานะอยู่ตัว ดังนั้นเมื่อตรวจหา DTS ได้แล้ว โดยส่วนมากจะทำการยับยั้งการทำงานของระบบ AEC ชั่วคราว เมื่อผ่านช่วง DTS แล้ว (ส่วนมากน้อยกว่า 1 วินาที) จึงอนุญาตให้ระบบ AEC ทำงานต่อ วิทยานิพนธ์ฉบับนี้จะใช้ DTD ในอุดมคติซึ่งมีวิธีการทำงานดังรูปที่ 4.2



รูปที่ 4.2 การทำงานของ DTD ในอุดมคติ

เช่นเดียวกับในกรณีของ VAD การตัดสิน DTS สามารถทำได้โดยการเปรียบเทียบพลังงานเสียงพูดทางห้องใกล้กับค่าขีดเริ่มค่าหนึ่งดังนี้

$$\mathrm{DTD} = \begin{cases} 1 & , \Delta(\ell) \geq th_{\mathrm{DTD}} \\ 0 & , \Delta(\ell) < th_{\mathrm{DTD}} \end{cases} \tag{4.3}$$

เมื่อ

$$\Delta(\ell) = \sum_{k} \left(|S(k,\ell)|^2 \right) \tag{4.4}$$

ในวิทยานิพนธ์ฉบับนี้เลือก $th_{\mathrm{DTD}} = 2 \times 10^{-4}$ โดย $\mathrm{DTD} = 1$ หมายถึงเวลาดังกล่าวเป็นช่วง DTS และ $\mathrm{DTD} = 0$ หมายถึงเวลาดังกล่าวเป็นช่วง STS

4.2. ตัวชี้วัดประสิทธิภาพ

4.2.1. Echo Attenuation (EA)

สามารถหาได้จากอัตราส่วนระหว่างสัญญาณเสียงรบกวนและสัญญาณเสียงรบกวนตกค้างซึ่งดัดแปลงมาจาก [49] ดังนี้

$$\mathrm{EA\ (dB)} = \frac{1}{L_E} \sum_{\ell \in L_E} F_{\mathrm{EA}} \left\{ 10 \times \log_{10} \left(\frac{\sum_{t=0}^{T-1} e^2(t + \ell T/2)}{\sum_{t=0}^{T-1} \tilde{e}^2(t + \ell T/2)} \right) \right\} \tag{4.5}$$

เมื่อ $\tilde{e}(t)$ คือ สัญญาณเสียงสะท้อนตกค้าง

$F_{\mathrm{EA}}\{\cdot\}$ คือ ฟังก์ชันที่เลือกเอาเฉพาะที่มีค่าอยู่ระหว่าง -80 dB และ 50 dB

L_E คือ จำนวนเฟรมเวลาที่ อัตราส่วนภายในวงเล็บในสมการที่ (4.5) มีค่าอยู่ระหว่าง -80 dB และ 50 dB

4.2.2. การลดของเสียงรบกวน (Noise Attenuation, NA)

สามารถหาได้จากอัตราส่วนระหว่างสัญญาณเสียงรบกวนและสัญญาณเสียงรบกวนตกค้างซึ่งดัดแปลงมาจาก [49] ดังนี้

$$\mathrm{NA\ (dB)} = \frac{1}{L_N} \sum_{\ell \in L_N} F_{\mathrm{NA}} \left\{ 10 \times \log_{10} \left(\frac{\sum_{t=0}^{T-1} n^2(t + \ell T/2)}{\sum_{t=0}^{T-1} \tilde{n}^2(t + \ell T/2)} \right) \right\} \tag{4.6}$$

เมื่อ $\tilde{n}(t)$ คือ สัญญาณเสียงสะท้อนตกค้าง

$F_{\mathrm{NA}}\{\cdot\}$ คือ ฟังก์ชันที่เลือกเอาเฉพาะเฟรมเวลาที่ไม่มีเสียงพูด และ มีค่าอยู่ระหว่าง -80 dB และ 50 dB

L_N คือ จำนวนเฟรมเวลาเฉพาะที่ไม่มีเสียงพูด และ มีค่าอยู่ระหว่าง -80 dB และ 50 dB

4.2.3. ค่าความผิดเพี้ยนแยกส่วน (Segmental SNR, SegSNR)

SegSNR ถูกใช้เป็นตัวประเมินผลในเรื่องความผิดเพี้ยนของเสียงพูดตัวหนึ่งในวิทยานิพนธ์ฉบับนี้ เนื่องจากให้ผลค่อนข้างที่จะสอดคล้องกับการฟังจริง โดย SegSNR สามารถหาได้ดังนี้ [18]

$$\mathrm{SegSNR}_{\hat{S}}^{S}\ (\mathrm{dB}) = \frac{1}{L_S}\sum_{\ell\in L_S} F_{\mathrm{S}}\left\{10\times\log_{10}\left(\frac{\sum_{t=0}^{T-1} s^2(t+\ell T/2)}{\sum_{t=0}^{T-1}\left(s(t+\ell T/2)-\hat{s}(t+\ell T/2)\right)^2}\right)\right\} \tag{4.7}$$

เมื่อ $\hat{s}(t)$ คือ ค่าประมาณสัญญาณเสียงพูด

$F_{\mathrm{S}}\{\cdot\}$ คือ ฟังก์ชันที่เลือกเอาเฉพาะเฟรมเวลาที่มีเสียงพูด

L_S คือ จำนวนเฟรมเวลาเฉพาะที่มีเสียงพูด

ในทางปฏิบัติจะใช้ SegSNR Improvement $\Delta\mathrm{SegSNR}$ ในการแสดงผลแทนโดย SegSNR Improvement หาได้จาก

$$\Delta\mathrm{SegSNR} = \mathrm{SegSNR}_{\hat{S}}^{S} - \mathrm{SegSNR}_{Y}^{S} \tag{4.8}$$

เมื่อ SegSNR_{Y}^{S} สามารถหาได้จากสมการที่ (4.7) โดยแทนที่ $\hat{s}(t)$ ด้วยสัญญาณไมโครโฟน $y(t)$

4.2.4. ระยะสเปกตรัมลอการิทึม (Log-Spectral Distance, LSD)

นอกจาก SegSNR แล้วปริมาณอีกชนิดหนึ่งซึ่งถูกเลือกใช้ในการประเมินผลความผิดเพี้ยนของเสียงพูดในวิทยานิพนธ์ฉบับนี้ได้แก่ LSD ซึ่งสามารถหาได้ดังนี้ [18]

$$\mathrm{LSD}_{S}^{S}\ (\mathrm{dB}) = \frac{1}{L_S}\sum_{\ell\in L_S} F_{\mathrm{S}}\left\{\left(\frac{1}{K/2+1}\sum_{k=0}^{K/2}\left[20\times\log_{10}\left(\frac{|S(k,\ell)|}{|\hat{S}(k,\ell)|}\right)\right]\right)^2\right\} \tag{4.9}$$

เมื่อ $\hat{S}(k,\ell)$ คือ ค่าประมาณสเปกตรัมเสียงพูดที่ได้จากการประมวล

เช่นเดียวกันในทางปฏิบัติจะทำการแสดงผลในรูปของ LSD Improvement $\Delta\mathrm{LSD}$ ดังนี้

$$\Delta\mathrm{LSD} = \mathrm{LSD}_{Y}^{S} - \mathrm{LSD}_{\hat{S}}^{S} \tag{4.10}$$

เมื่อ LSD_{Y}^{S} สามารถหาได้จากสมการที่ (4.9) โดยแทนที่ $\hat{S}(k,\ell)$ ด้วยสเปกตรัมสัญญาณไมโครโฟน $Y(k,\ell)$

4.2.5. สเปกโทรแกรม

เครื่องมือที่ใช้ประเมินค่าความผิดเพี้ยนของเสียงพูดตัวสุดท้ายในวิทยานิพนธ์ได้แก่ สเปกโทรแกรมซึ่งได้กล่าวถึงไปบ้างแล้วในหัวข้อย่อยที่ผ่านๆ มา ในหัวข้อย่อยนี้จะกล่าวถึงรายละเอียดของค่าตัวแปรต่างๆ ที่ใช้สร้างสเปกโทรแกรม

สเปกโทรแกรมเป็นเครื่องมือที่ใช้แสดงค่าพลังงานในแต่ละองค์ประกอบทางความถี่ ณ เวลาเฟรมเวลาต่างๆ กล่าวคือทำการแสดง $|X(k,\ell)|^2$ ที่ k และ ℓ ใดๆ ซึ่งหลักการแปลงจากโดเมนเวลามายังโดเมน STFT นั้นอาศัยหลักการเดียวกับที่ได้กล่าวถึงในหัวข้อที่ 2.1.1 โดยในวิทยานิพนธ์ฉบับนี้เลือกตัวแปรต่างๆ ดังนี้

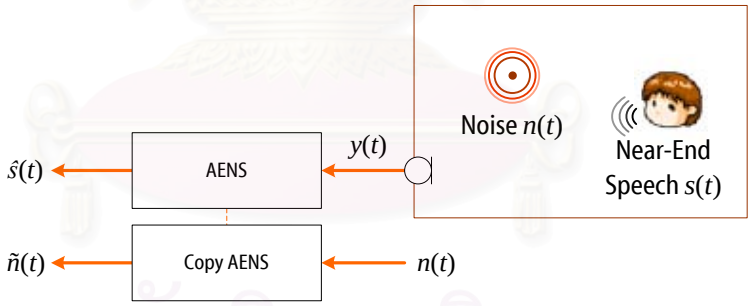
จำนวนตัวอย่างที่ใช้แปลง STFT $T = 256$ หน้าต่างที่ใช้ในการแปลงได้แก่ หน้าต่าง Hamming ค่าช่วงก้าวระหว่างเฟรม $M = 128$ (50% Overlap) แสดงผลด้วย ระดับสีเทา โดยสีที่เข้มกว่าแสดงถึงระดับพลังงานที่สูงกว่า

ตารางที่ 4.1 คุณลักษณะของสเปกโทรแกรมที่เลือกใช้

	ปริมาณ/วิธีการที่เลือกใช้
จำนวนตัวอย่างที่ใช้แปลง STFT T	512
ค่าช่วงก้างระหว่างเฟรม M	400 (78% Overlap)
หน้าต่างที่ใช้ในการแปลง	Hamming Window
การแสดงผล	Gray Scale

4.3. การทดลอง NS

ในหัวข้อย่อยนี้เป็นการวัดประสิทธิภาพวิธีการประมาณค่า a priori SNR แบบ TSSP, MTSSP และ MTSW เปรียบเทียบกับ DD, TSW และ SAAF ตัวชี้วัดประสิทธิภาพที่เลือกใช้ได้แก่ NA, SegSNR Improvement, LSD Improvement และ สเปกโทรแกรม ระบบการทดลองถูกบรรยายไว้ในรูปที่ 4.3



รูปที่ 4.3 การทดลองระบบ NS

เสียงพูดที่ใช้เป็นเสียงผู้ชายจำนวน 3 คน โดยพูดคนละ 1 ประโยค ที่ความถี่ชักตัวอย่าง 8 kHz แต่ละสัญญาณเสียงพูดถูกก่อกวนด้วยเสียงรบกวนสีขาว ที่ระดับ Global SNR ในช่วง 5 ถึง 20 dB

เสียงพูดที่ถูกรบกวนถูกแปลงสู่โดเมน STFT ด้วย Fast Fourier Transform (FFT) และใช้หน้าต่างแบบ Hanning จำนวน 256 ตัวอย่าง ทำการแปลงแบบเหลื่อมกัน (Overlap) 50%

Spectral Gain G_η ที่เลือกใช้ได้แก่ G_{SE} โดยอาศัยขั้นตอนการปรับระดับ Gain ดังนี้ [17]

$$G_\eta = \begin{cases} G_\eta, & G_\eta > G_{\mathrm{floor}} \\ G_{\mathrm{floor}}, & \text{otherwise} \end{cases} \tag{4.11}$$

เมื่อ G_{floor} เป็นค่า Gain ที่ต่ำที่สุด ในวิทยานิพนธ์ฉบับนี้เลือก $G_{\mathrm{floor}} = 0.05$

ค่าประมาณ NPSD $\lambda_N(k,\ell)$ อาศัยการถ่วงน้ำหนักตามสมการที่ (2.34) โดยทำงานร่วมกับ VAD ในอุดมคติ

ตารางที่ 4.2 การทดลอง NS

	ปริมาณ/วิธีการที่เลือกใช้
Sampling rate	8 kHz
Quantization bit rate	16 bit
Length of frame T	256 ตัวอย่าง
Frame step M	128 ตัวอย่าง
Window function	Hanning
Speech	ผู้ชาย
Noise	เสียงรบกวนสีขาว
NPSD estimation	ค่าเฉลี่ยในช่วงที่ไม่มีเสียงพูด อาศัย VAD ในอุดมคติ
A priori SNR estimation	DD [7], TSW [18], MTSW, TSSP, MTSSP และ SAAF [20]
Spectral Gain	G_{SE} โดยอาศัยการปรับระดับ
Objective evaluation	SegSNR Improvement, LSD Improvement และ สเปกโทรแกรม

ตารางที่ 4.3 SegSNR Improvement ในการทดสอบการประมาณ a priori SNR
เมื่อเสียงรบกวนเป็นเสียงรบกวนสีขาว ในระดับต่างๆ และ $G_\eta = G_{\mathrm{SE}}$

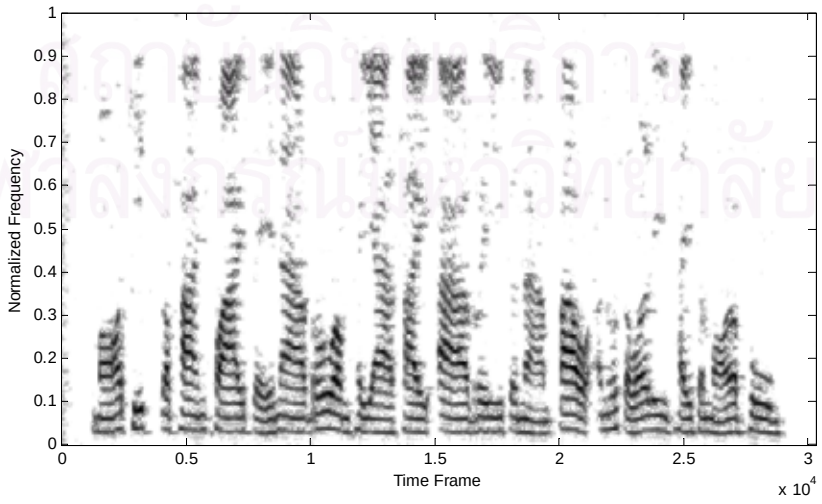
SegSNR Improvement (dB)				
Input SNR	5 dB	10 dB	15 dB	20 dB
DD	5.1580	3.2892	1.3689	-0.2866
TSW	5.8136	4.0510	2.2077	0.6351
MTSW	6.0180	4.2153	4.2153	0.7775
TSSP	6.1334	4.3719	2.5895	1.0312
MTSSP	6.2253	4.4568	2.6669	1.0973
SAAF	5.3523	4.2488	2.9924	1.6830

ตารางที่ 4.4 LSD Improvement ในการทดสอบการประมาณ a priori SNR
เมื่อเสียงรบกวนเป็นเสียงรบกวนสีขาว ในระดับต่างๆ และ $G_\eta = G_{SE}$

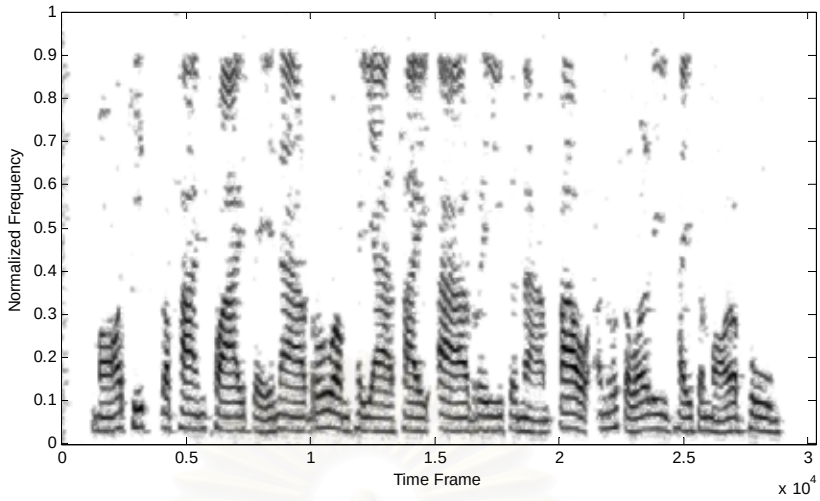
LSD Improvement (dB)				
Input SNR	5 dB	10 dB	15 dB	20 dB
DD	8.4001	6.0722	4.2216	3.1602
TSW	8.4778	6.1392	4.2819	3.2386
MTSW	8.5323	6.1694	4.3172	3.2740
TSSP	8.7036	6.5754	4.8866	3.9082
MTSSP	8.7314	6.5918	4.9025	3.9226
SAAF	8.7314	5.8184	4.8328	4.0333

ตารางที่ 4.5 NA ในการทดสอบการประมาณ a priori SNR
เมื่อเสียงรบกวนเป็นเสียงรบกวนสีขาว ในระดับต่างๆ และ $G_\eta = G_{SE}$

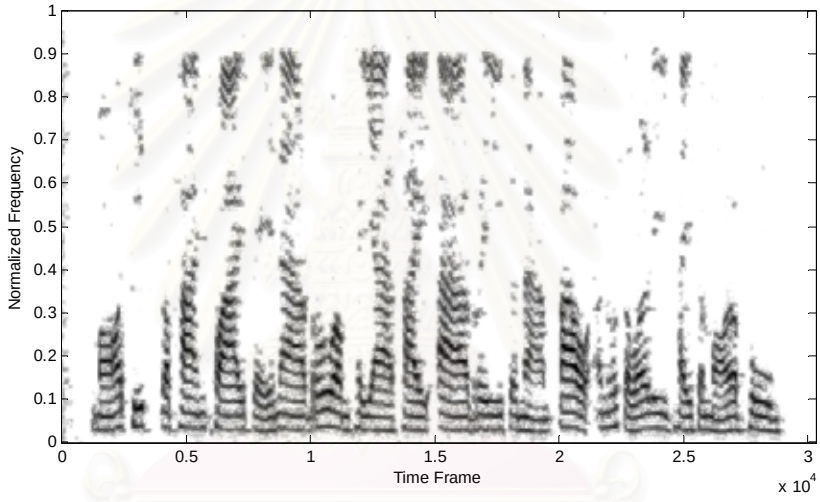
NA (dB)				
Input SNR	5 dB	10 dB	15 dB	20 dB
DD	18.4745	16.8598	15.4477	14.1195
TSW	19.3156	17.7150	16.3618	14.9985
MTSW	18.6039	17.2167	15.9494	14.6691
TSSP	17.7717	16.4255	15.1285	13.8209
MTSSP	17.5657	16.3170	15.0654	13.7555
SAAF	16.5145	15.2515	14.0265	13.2354



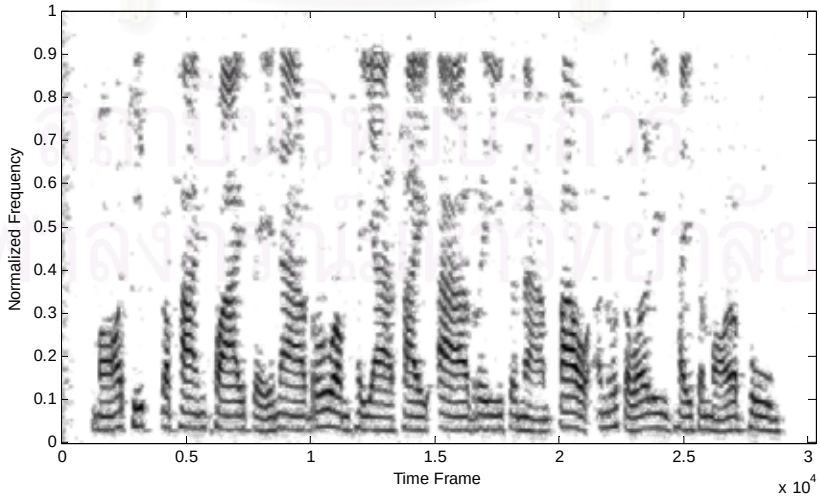
(ก)


Normalized Frequency
Time Frame
x 10^4

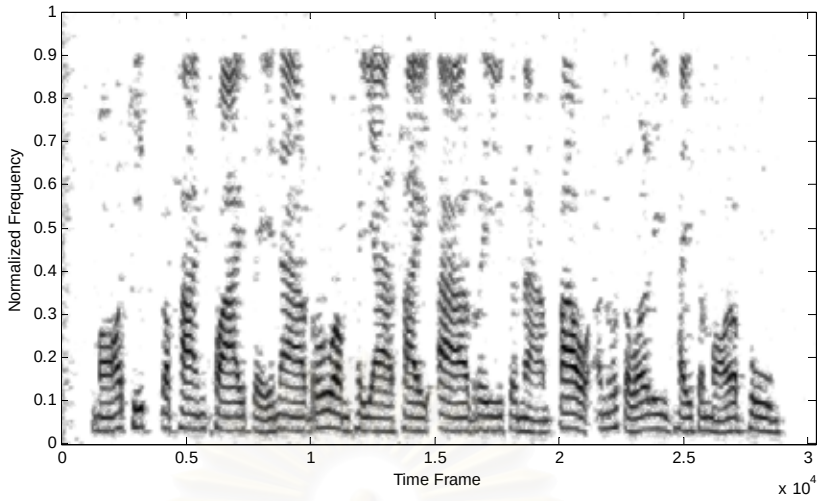
(ข)


Normalized Frequency
Time Frame
x 10^4

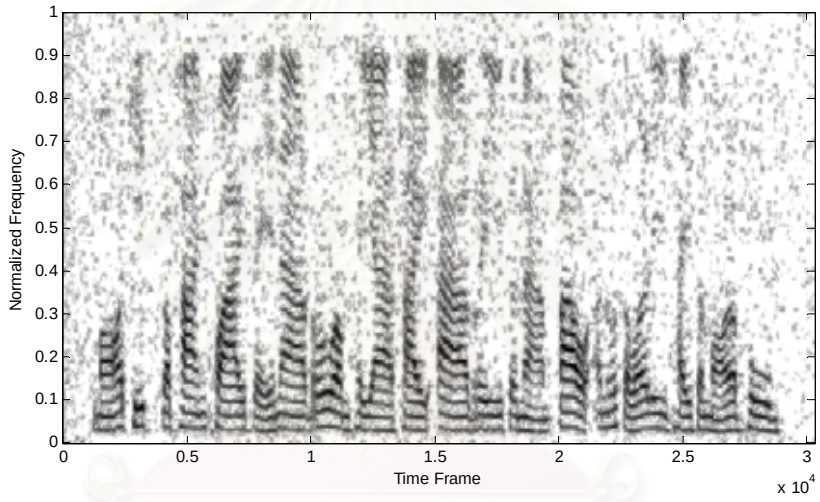
(ค)


Normalized Frequency
Time Frame
x 10^4

(ง)



(จ)



(ช)

รูปที่ 4.4 สเปกโทรแกรมสัญญาณเสียงพูดที่ถูกปรับปรุงจากเสียงพูดที่ถูกรบกวนที่ระดับ SNR 10 dB โดยอาศัยการประมาณ a priori SNR แบบ
(ก) DD (ข) TSW (ค) MTSW (ง) TSSP (จ) MTSSP และ (ช) SAAF

จากผลการทดลองเชิงปรวิสัย (Objective test) เห็นว่าการประมาณค่า a priori SNR ที่นำเสนอได้แก่ MTSW, TSSP และ MTSSP ให้ผลความผิดเพี้ยนของเสียงพูดที่ดีขึ้นกว่าวิธีการประมาณที่มีอยู่เดิม โดยการประมาณในตระกูลของ TSSP นั้นให้ผลการผิดเพี้ยนของเสียงพูดที่ดีขึ้นกว่าตระกูล TSW นอกจากนี้หากวิเคราะห์จากสเปกโทรแกรมจะเห็นว่า SAAF สามารถรักษาส่วนของเสียงพูดในแต่ละองค์ประกอบทางความถี่ไว้ได้มากที่สุด แต่อย่างไรก็ตาม SAAF ให้ผลการลดเสียงรบกวนที่แย่กว่าทุกๆ การประมาณ ซึ่งสังเกตได้จากทั้ง ค่า NA และสเปกโทรแกรม

4.4. การทดลอง AENS

4.4.1. การทดสอบเปรียบเทียบระหว่าง AENS ใน X[53]X และ AENS ที่นำเสนอ

การทดลองจะแบ่งออกเป็น 2 ช่วงได้แก่ STS และ DTS โดยมีรูปแบบการทำงานดังต่อไปนี้

4.4.1.1. STS

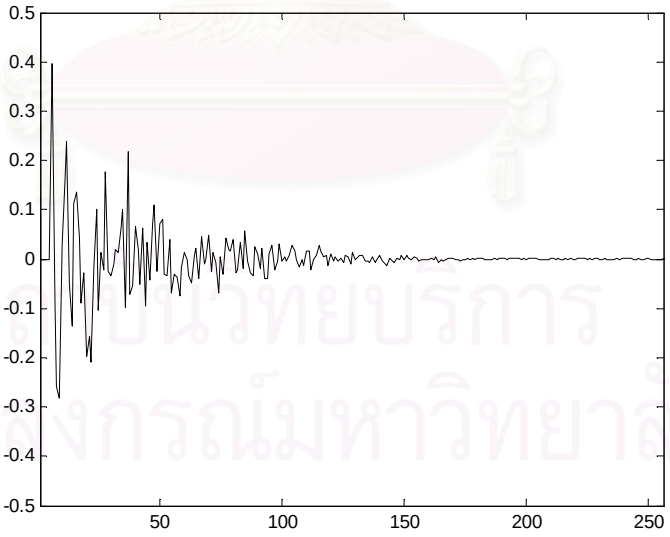
การทดลองนี้อาศัยตัวชี้วัดประสิทธิภาพคือ EA และ NA เสียงพูดทางห้องไกลที่ใช้เป็นเสียงผู้ชาย 3 คนโดยพูดคนละ 1 ประโยคความถี่ชักตัวอย่างอยู่ที่ 8 kHz วิถีสะท้อนทางเสียงที่ใช้ได้มาจากการจำลองดังนี้

$$h(t) = r(t)\cos(8\pi t)\exp(-0.0034\frac{2048}{L}t) \tag{4.12}$$

เมื่อ $r(t)$ คือสัญญาณสุ่มสีขาวที่มีค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนเป็น 1 และระดับความดังของเสียงสะท้อนสามารถควบคุมได้ดังนี้

$$h(t) = \alpha_h \frac{h(t)}{\|h(t)\|^2} \tag{4.13}$$

โดย α_h เป็นค่าลดทอนความดังของเสียงสะท้อน รูปที่ 4.5 ถูกใช้แสดงตัวอย่างวิถีสะท้อนทางเสียง เมื่อใช้ $L = 256$ และ $\alpha_h = 0.9$



รูปที่ 4.5 ตัวอย่างวิถีสะท้อนทางเสียง เมื่อใช้ $L = 256$ และ $\alpha_h = 0.9$

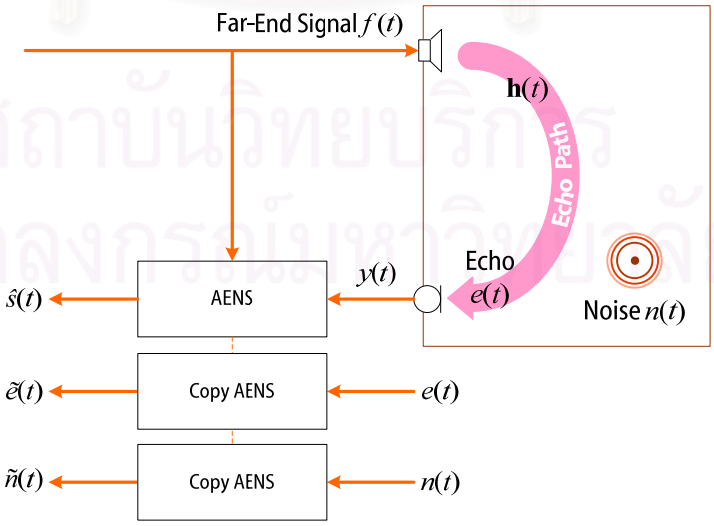
เสียงรบกวนที่ใช้เป็นเสียงรบกวนสีขาวที่ระดับ SNR 5 และ 10 dB เนื่องจากเสียงสะท้อนมีคุณสมบัติที่แตกต่างจากเสียงรบกวนพื้นหลัง ดังนั้นการปรับระดับ Gain ในสมการที่ (4.11) จึงเปลี่ยนเป็น [54]

$$G_\eta(k,\ell)=\begin{cases} G_\eta(k,\ell), & G_\eta(k,\ell) > G_{\text{floor}} \\ \dfrac{\lambda_N(k,\ell)}{\lambda_E(k,\ell)+\lambda_N(k,\ell)} G_{\text{floor}}, & \text{otherwise} \end{cases} \tag{4.14}$$

การทดลองถูกสรุปไว้ในตารางที่ 4.6 และรูปที่ 4.6

ตารางที่ 4.6 การทดลอง AENS ในช่วง STS

	ปริมาณ/วิธีการที่เลือกใช้	
Sampling rate	8 kHz	
Quantization bit rate	16 bit	
Length of frame T	256 ตัวอย่าง	
Frame step M	128 ตัวอย่าง	
Window function	Hanning	
Far-End Speech	ผู้ชาย	
Noise	เสียงรบกวนสีขาว	
Echo path	32 ms (256 ตัวอย่าง) และ $\alpha_h = 0.9$	
NPSD estimation	เฉลี่ยในช่วงที่ไม่มีเสียงพูด อาศัย VAD ในอุดมคติ	
	AENS [53]	Proposed AENS
EPSD estimation	CFM	Proposed
A priori SNR estimation	DD	TSW, MTSW, TSSP และ MTSSP
Spectral Gain	G_{SE} โดยอาศัยการปรับระดับ	
Objective Evaluation	EA และ NA	



รูปที่ 4.6 การทดลองระบบ AENS (STS)

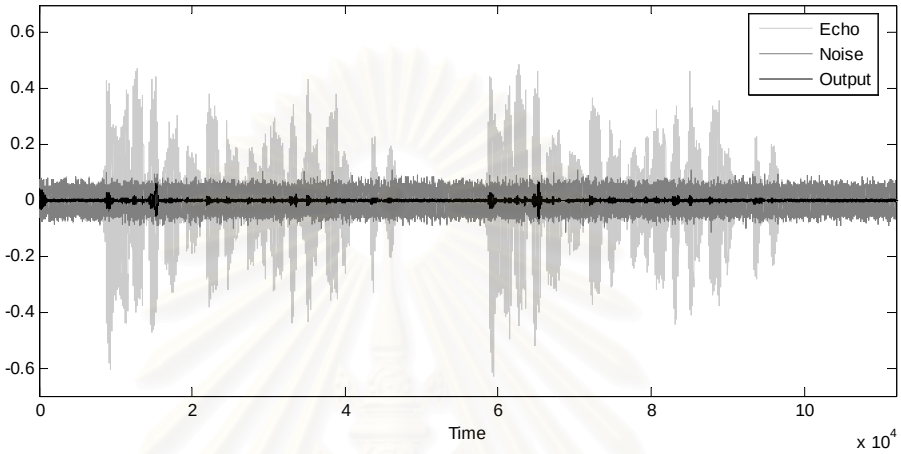
ตารางที่ 4.7 EA ของการทดลอง AENS ในช่วง STS

EA (dB)			
Algorithm		Input SNR	
EPSD Estimation	A priori SDR Estimation	5 dB	10 dB
CFM [53]	DD	29.2497	28.7182
Proposed	TSW	40.8044	39.3651
Proposed	MTSW	38.6209	37.4222
Proposed	TSSP	36.3504	33.8319
Proposed	MTSSP	35.2764	32.9677

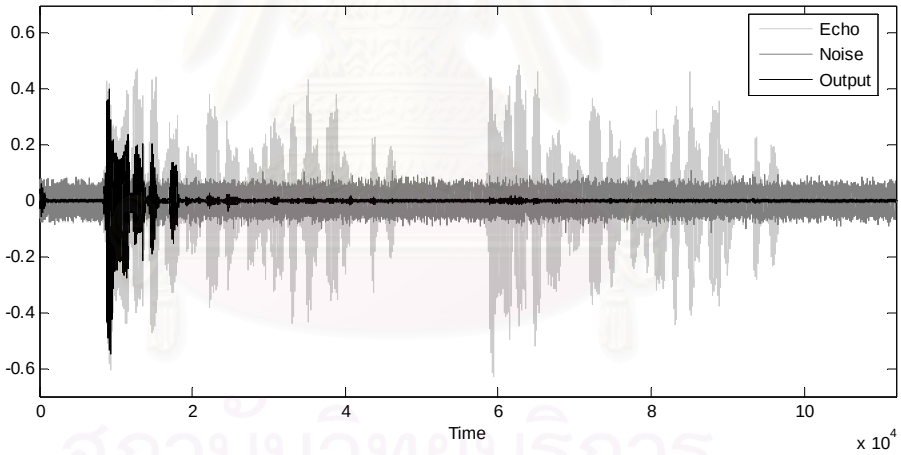
ตารางที่ 4.8 NA ของการทดลอง AENS ในช่วง STS

NA (dB)			
Algorithm		Input SNR	
EPSD Estimation	A priori SDR Estimation	5 dB	10 dB
CFM [53]	DD	26.7742	27.1341
Proposed	TSW	26.2165	25.8117
Proposed	MTSW	26.2165	25.1362
Proposed	TSSP	25.3345	24.6894
Proposed	MTSSP	24.9014	24.1813

จากตารางที่ 4.7 และตารางที่ 4.8 จะเห็นว่า AENS ที่นำเสนอให้ผลการกดเสียงสะท้อนลงได้มากกว่า AENS ใน [53] ในขณะที่ปนิมาณเสียงรบกวนที่ลดลงยังอยู่ในระดับที่ยอมรับได้ ทั้งนี้แลกมาซึ่งความซับซ้อนในการคำนวณที่เพิ่มมากยิ่งขึ้น ดังที่ได้แสดงไว้ในหัวข้อที่ 3.3 อย่างไรก็ตามการประมาณ EPSD ที่นำเสนอสามารถรองรับกับวิถีสะท้อนทางเสียงที่มีระยะเวลาประวิงนานขึ้นได้ โดยเพิ่มจำนวนสัมประสิทธิ์ของวงจรกรองแบบปรับตัว ในขณะที่ CFM ใน [53] ไม่สามารถรองรับวิถีสะท้อนที่มีระยะเวลาประวิงนานมากได้



(ก)



(ข)

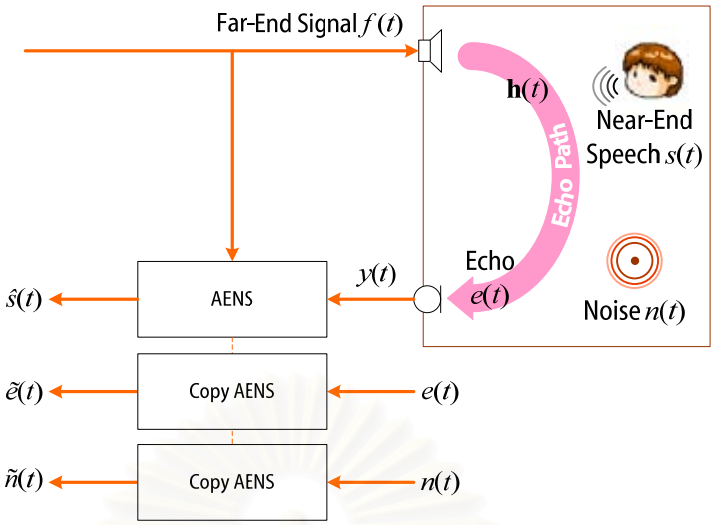
รูปที่ 4.7 เสียงสะท้อนและเสียงรบกวนตกค้างในทางเวลาของ
(ก) AENS [53] และ (ข) Proposed AENS ที่ใช้ MTSSP

4.4.1.2. DTS

หัวข้อนี้มุ่งเน้นเปรียบเทียบ AENS ใน [53] กับ AENS ที่นำเสนอในด้านความผิดเพี้ยนของเสียงพูด ในช่วง DTS ตัวประเมินค่าที่ใช้ได้แก่ SegSNR Improvement, LSD Improvement, EA, NA และ สเปกโทรแกรมของเสียงพูดที่ถูกปรับปรุง เสียงสะท้อนถูกจำลองจากเสียงพูดทางห้องไกลและวิถีสะท้อนทางเสียงเช่นเดียวกับในช่วง STS โดยใช้ $L = 256$ และ $\alpha_h = 0.9$ เสียงพูดทางห้องใกล้เป็นเสียงผู้ชาย 3 คนพูดคนละหนึ่งประโยค เสียงรบกวนสีขาวที่ระดับ SNR 5 และ 10 dB ถูกใช้ในการทดลอง และกำหนดให้ DTS เกิดขึ้นในช่วงที่ วงจรกรองแบบปรับตัวอยู่ในสถานะอยู่ตัวแล้ว การทดลองถูกสรุปดังตารางที่ 4.9 และรูปที่ 4.8

ตารางที่ 4.9 การทดลอง AENS ในช่วง DTS

	ปริมาณ/วิธีการที่เลือกใช้	
Sampling rate	8 kHz	
Quantization bit rate	16 bit	
Length of frame T	256 ตัวอย่าง	
Frame step M	128 ตัวอย่าง	
Window function	Hanning	
Near-End Speech	ผู้ชาย	
Far-End Speech	ผู้ชาย	
Noise	เสียงรบกวนสีขาว	
Echo path	32 ms (256 ตัวอย่าง) และ $\alpha_h = 0.9$	
NPSD estimation	ค่าเฉลี่ยในช่วงที่ไม่มีเสียงพูด อาศัย VAD ในอุดมคติ	
	AENS [53]	Proposed AENS
EPSD estimation	CFM	Proposed ($\mu = 0.02$)
A priori SNR estimation	DD	TSW, MTSW, TSSP และ MTSSP
Spectral Gain	G_{SE} โดยอาศัยการปรับระดับ	
Objective Evaluation	SegSNR Improvement, LSD Improvement, EA, NA และ สเปกโทรแกรม	
Subjective evaluation	Mean Opinion Score (MOS)	



รูปที่ 4.8 การทดลองระบบ AENS (DTS)

ตารางที่ 4.10 SegSNR Improvement ของการทดลอง AENS ในช่วง DTS

SegSNR Improvement (dB)			
Algorithm		Input SNR	
EPSD Estimation	A priori SDR Estimation	5 dB	10 dB
CFM [53]	DD	3.9399	2.7063
Proposed Method	TSW	6.2575	5.8682
Proposed Method	MTSW	6.9259	6.4843
Proposed Method	TSSP	6.8001	6.3873
Proposed Method	MTSSP	7.1113	6.6995

ตารางที่ 4.11 LSD Improvement ของการทดลอง AENS ในช่วง DTS

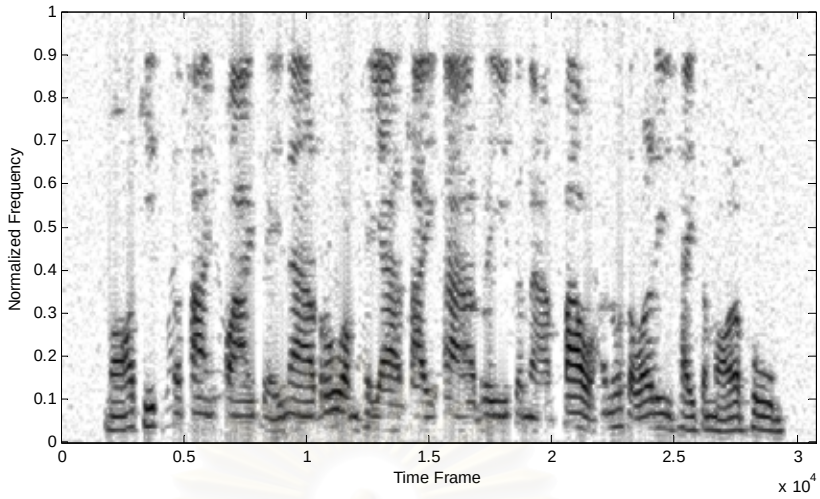
LSD Improvement (dB)			
Algorithm		Input SNR	
EPSD Estimation	A priori SDR Estimation	5 dB	10 dB
CFM [53]	DD	6.6701	4.5211
Proposed Method	TSW	7.1370	5.4258
Proposed Method	MTSW	7.3035	5.5215
Proposed Method	TSSP	7.6030	6.0864
Proposed Method	MTSSP	7.6615	6.1477

ตารางที่ 4.12 EA ของการทดลอง AENS ในช่วง DTS

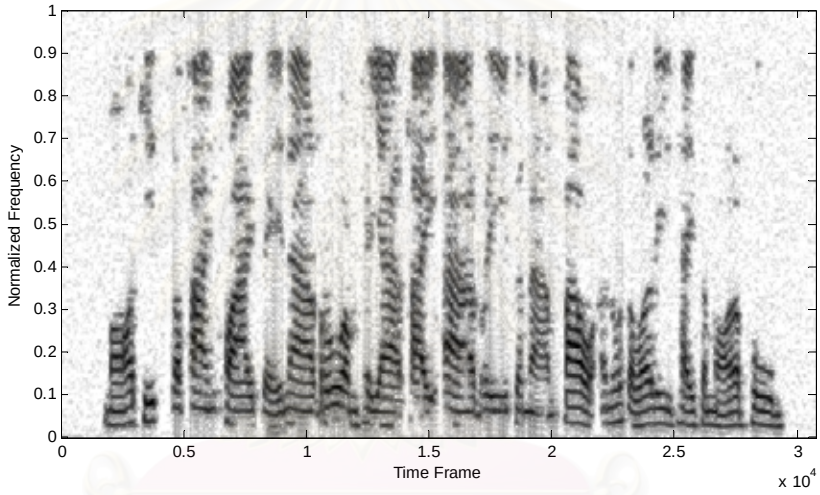
EA (dB)			
Algorithm		Input SNR	
EPSD Estimation	A priori SDR Estimation	5 dB	10 dB
CFM [53]	DD	25.8282	25.2079
Proposed Method	TSW	32.8353	30.4970
Proposed Method	MTSW	29.7149	27.1554
Proposed Method	TSSP	27.9823	25.5123
Proposed Method	MTSSP	26.7844	24.2142

ตารางที่ 4.13 NA ของการทดลอง AENS ในช่วง DTS

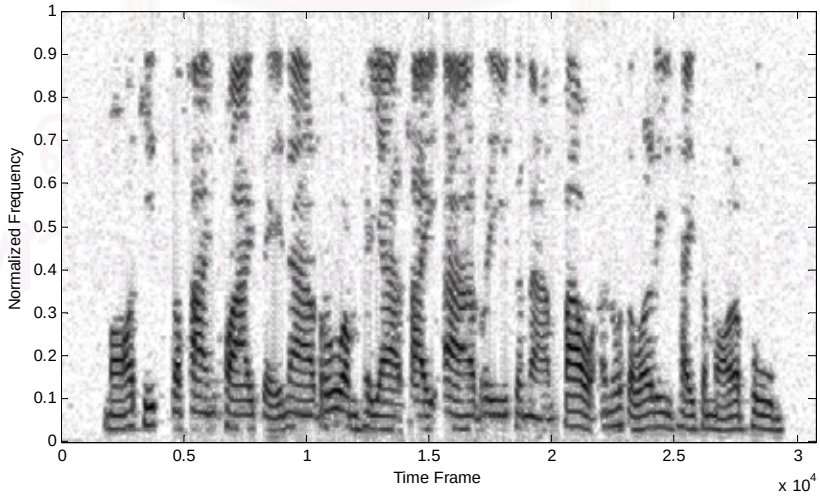
NA (dB)			
Algorithm		Input SNR	
EPSD Estimation	A priori SDR Estimation	5 dB	10 dB
CFM [53]	DD	24.7637	24.7597
Proposed Method	TSW	23.5918	22.6536
Proposed Method	MTSW	22.7598	21.7667
Proposed Method	TSSP	21.7667	20.4663
Proposed Method	MTSSP	21.0520	20.0251


Normalized Frequency
Time Frame
x 10^4

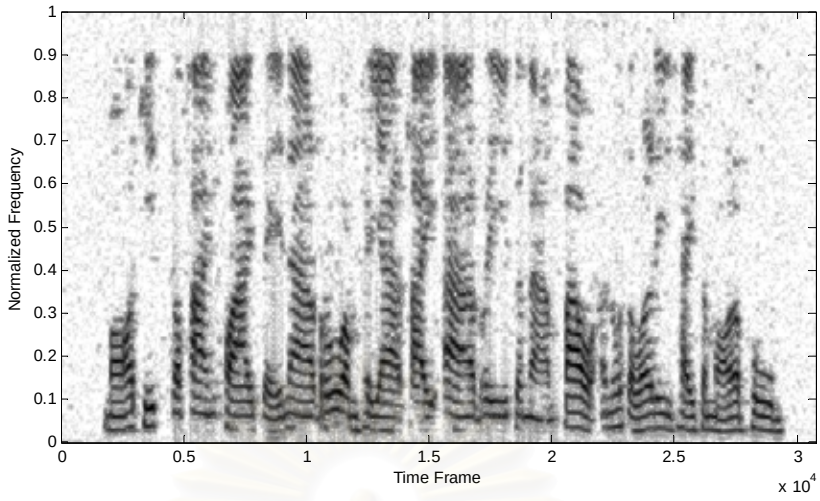
(ก)


Normalized Frequency
Time Frame
x 10^4

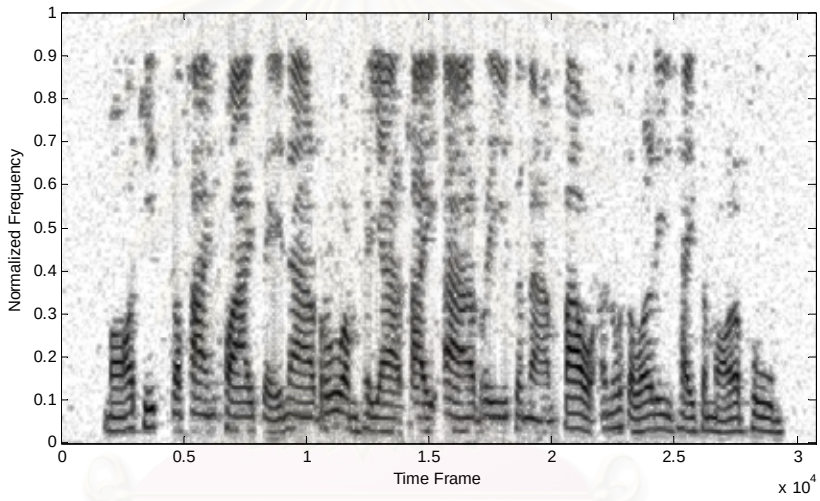
(ข)


Normalized Frequency
Time Frame
x 10^4

(ค)

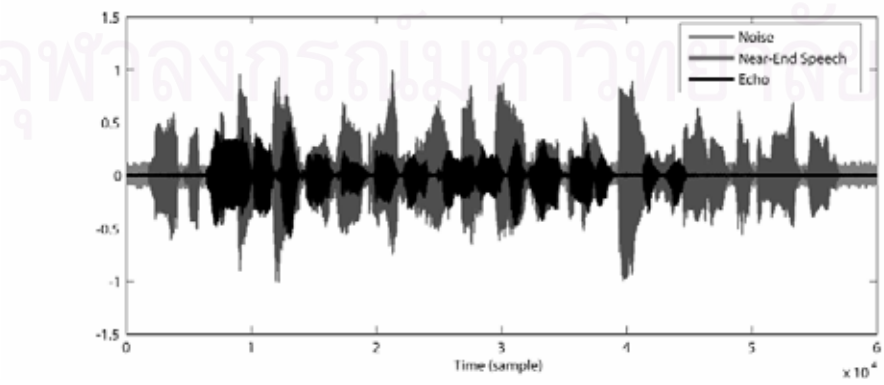


(ง)



(จ)

รูปที่ 4.9 สเปกโทรแกรมเสียงพูดที่ถูกปรับปรุงจากเสียงพูดที่ถูกรบกวนที่ระดับ SNR 10 dB โดย (ก) AENS [53] (ข) Proposed AENS ที่ใช้ TSW (ค) MTSW (ง) TSSP และ (จ) MTSSP



รูปที่ 4.10 สัญญาณเสียงในทางเวลาเพื่อใช้บ่งบอกช่วง DTS สำหรับสเปกโทรแกรมในรูปที่ 4.9

การทดลองชี้ให้เห็นว่า AENS ที่นำเสนอให้ผลความผิดเพี้ยนของเสียงพูดที่น้อยกว่า AENS ที่ถูกเสนอใน [53] โดยเมื่อเทียบระหว่างวิธีที่นำเสนอกันเองแล้ว AENS ที่ใช้คู่กับ MTSSP ให้ค่าความผิดเพี้ยนของเสียงพูดที่น้อยที่สุด ในขณะที่ถึงแม้ค่า NA และ EA ของ MTSW และ MTSSP ลดลงจากของ TSW และ TSSP ตามลำดับ ก็ตาม แต่ค่า EA และ NA ดังกล่าวนี้ยังคงอยู่ในระดับที่รับได้

การทดลองยังชี้ให้เห็นว่าถึงแม้ MTSW และ MTSSP จะไม่ได้ให้ความแตกต่างกับ TSW และ TSSP มากนักใน NS แต่สำหรับ AENS แล้ว MTSW และ MTSW ให้ผลการรักษาส่วนของเสียงพูดในองค์ประกอบทางความถี่ไว้ได้ดีกว่า TSW และ TSSP มาก ทั้งนี้เนื่องจากเมื่อเสียงก่อกวนเป็นเสียงสะท้อนซึ่งมีลักษณะคล้ายคลึงกับเสียงพูดแล้ว จะส่งผลให้แม้กระทั่งเสียงพูด ณ องค์ประกอบทางความถี่ต่ำ ไม่ได้มีพลังงานงานมากเมื่อเทียบกับเสียงก่อกวนอีกต่อไป ทำให้การประมาณค่า a priori SER ที่ไวต่อการติดตามค่า Instantaneous SNR มีผลได้เปรียบกว่า วิธีการประมาณที่เชื่องช้าเป็นอย่างมาก โดยในสถานการณ์ที่มีเฉพาะเสียงรบกวนดังเช่นใน NS นั้น เสียงพูด ณ องค์ประกอบทางความถี่ต่ำมีพลังงานมากเมื่อเทียบกับเสียงรบกวนอยู่แล้ว ดังนั้นไม่ว่าการประมาณ a priori SNR เช่นไร (เชื่องช้าต่อการเปลี่ยนแปลง หรือรวดเร็วต่อการเปลี่ยนแปลง) ก็สามารถติดตามการเปลี่ยนแปลงได้ทัน

นอกจากผลการทดลองเชิงปรวิสัยแล้ว ผลการทดลองเชิงอัตวิสัย (Subjective test) กับผู้ฟังจำนวน 20 คน โดยอาศัยหลักการคิดคะแนนเฉลี่ย (The Mean Opinion Score, MOS) ตามมาตรฐานของ ITU-R [61] ยังถูกนนำเสนอในการทดลองนี้อีกด้วย ขั้นตอนการทดสอบเชิงอัตวิสัยแบบ MOS คือ ให้ผู้ทำการทดสอบรับฟังเสียงพูดที่ถูกปรับปรุงแล้วแต่ละชุดที่ต้องการทดสอบ และจากนั้นผู้ทดสอบจึงให้คะแนนแก่คุณภาพของเสียงพูดที่ถูกปรับปรุงแล้วในแต่ละชุดนั้นๆ อย่างอิสระต่อกัน (หรือที่เรียกว่าให้เกรด) โดยสามารถให้คะแนนได้ตั้งแต่ 1 คะแนนถึง 5 คะแนน เมื่อความสัมพันธ์ของคุณภาพเสียงกับคะแนนถูกอธิบบายดังตารางที่ 4.14

วิธีการทดลองเชิงอัตวิสัยกระทำโดยให้ผู้ฟังจำนวน 20 คนฟังผลของเสียงพูดที่ถูกปรับปรุงจากขั้นตอนวิธีการต่างๆ แล้วให้คะแนนในด้านความเป็นธรรมชาติของเสียงพูด โดยแต่ละคนมีสามารถให้คะแนนในแต่ละวิธีได้สูงสุด 5 คะแนน ผลการทดลองได้จากการนำคะแนนที่ทดสอบได้มาเฉลี่ย

ตารางที่ 4.14 คุณภาพของเสียงพูดที่ถูกปรับปรุงตามคะแนนต่างๆ ที่ให้กับการทดสอบเชิงอัตวิสัยแบบ MOS

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Unsatisfactory (Bad)	Very annoying and objectionable

โดยให้ผู้ทำการทดสอบได้ฟังเสียงพูดที่ถูกก่อกวนด้วยเสียงรบกวนที่ระดับ SNR 10 dB และ เสียงสะท้อนที่ระดับ $\alpha_h = 0.9$ ก่อนเพื่อใช้เป็นบรรทัดฐานในการตัดสินใจคุณภาพของเสียงพูดที่ถูกปรับปรุงจากเสียงพูดที่ถูกก่อกวนดังกล่าวต่อไป หลังจากผู้ทดสอบได้ฟังเสียงพูดที่ถูกก่อกวนแล้ว ต่อจากนั้นจะให้ผู้ทดสอบได้ฟังเสียงพูด

ที่ถูกปรับปรุงด้วยวิธีการ AENS ใน [53] วิธีการ AENS ที่นำเสนอ และ เสียงพูดสะอาด เพื่อทำการให้คะแนนกับเสียงพูดที่ถูกปรับปรุงจากแต่ละวิธีการ โดยผลการทดสอบเชิงอัตวิสัยเป็นดังตารางที่ 4.15

ตารางที่ 4.15 Mean Opinion Score (MOS)

MOS Test	
Clean Speech	4.8750
Disturbed Speech	1.2083
AENS [53]	2.6667
Proposed AENS (MTSSP)	3.5833

จากตารางที่ 4.15 จะสามารถเป็นได้อย่างชัดเจนว่า เสียงพูดที่ถูกปรับปรุงด้วยวิธี AENS ที่นำเสนอมีคุณภาพและให้ความรู้สึกที่ดีกว่า AENS ใน [53] สอดคล้องกับผลการทดลองที่ได้จากการทดลองแบบปริวิสัย

4.4.2. การทดลองปรียบเทียบระหว่าง AECNS และ AENS ที่นำเสนอ

เนื่องจาก AENS ถูกบ่งชี้ว่ามีข้อด้อยกว่า AECNS เดิมอย่างยิ่ง เนื่องจากทำให้เกิด ความผิดเพี้ยนของเสียงพูดทางห้องใกล้อย่างมาก โดยเฉพาะอย่างยิ่งใน DTS ดังนั้นในการเปรียบเทียบในหัวข้อนี้จะทำการเปรียบเทียบผลดังกล่าวเฉพาะในช่วง DTS เท่านั้น และเกิดขึ้นในช่วงที่ วงจรกรองแบบปรับตัวของทั้ง AEC และ AENS อยู่ในสภาวะอยู่ตัวแล้วทั้งสิ้น โดยขั้นตอนวิธีการปรับตัวของ AEC เลือกใช้ VSNLMS [43] และขั้นตอน NS สำหรับ AECNS เลือกใช้การประมาณค่าต่างๆ เช่นเดียวกับ AENS ทั้งสิ้น โดย AENS ที่เลือกมาเปรียบเทียบได้แก่ AENS ที่นำเสนอโดยอาศัย MTSSP ในการประมาณ a priori SDR

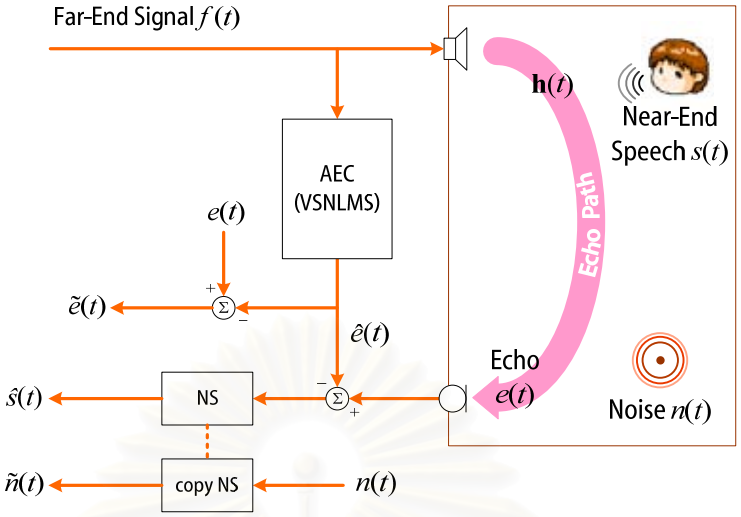
ค่าชี้วัดที่ใช้ในการทดลองนี้ได้แก่ SegSNR Improvement, LSD Improvement, EA, NA และ สเปกโทรแกรมของเสียงพูดที่ถูกปรับปรุง เนื่องจากปริมาณ $\tilde{e}(t)$ และ $\tilde{n}(t)$ ของวิธี AECNS ไม่สามารถหาได้ในลักษณะเดียวกับของวิธี AENS ที่นำเสนอ ดังนั้นการวัดค่าปริมาณทั้งสองสำหรับวิธี AECNS จึงแตกต่างไปจากขั้นตอนในรูปที่ 4.8 โดยจะทำการวัดปริมาณทั้งสองดังรูปที่ 4.11

ตารางที่ 4.16 วิธีการ VSNLMS ใน AECNS

ตัวแปรและค่าที่เลือกใช้ในวิธีการ VSNLMS	
Step Size	3
δ	10^{-6}

ตารางที่ 4.17 วิธี NS ใน AECNS

ตัวแปรและค่าที่เลือกใช้ในวิธี NS ใน AECNS	
A priori SNR estimation	MTSSP
Spectral Gain	G_{SE} โดยอาศัยการปรับระดับ



รูปที่ 4.11 การหาปริมาณ (ก) $\tilde{e}(t)$ (ข) $\tilde{n}(t)$

ตารางที่ 4.18 SegSNR Improvement ของการทดลองเปรียบเทียบ AECNS และ AENS ในช่วง DTS

SegSNR Improvement (dB)		
Input SNR	5 dB	10 dB
AECNS	9.2358	10.2479
Proposed AENS	6.7158	6.2089

ตารางที่ 4.19 LSD Improvement ของการทดลองเปรียบเทียบ AECNS และ AENS ในช่วง DTS

LSD Improvement (dB)		
Input SNR	5 dB	10 dB
AECNS	8.0387	6.5856
Proposed AENS	7.6251	6.1601

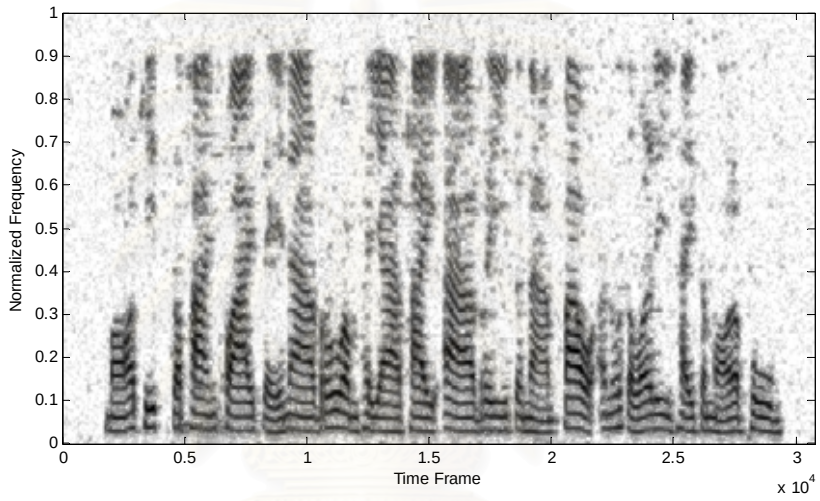
ตารางที่ 4.20 EA ของการทดลองเปรียบเทียบ AECNS และ AENS ในช่วง DTS

EA (dB)		
Input SNR	5 dB	10 dB
AECNS	19.7369	23.4155
Proposed AENS	26.4089	24.4856

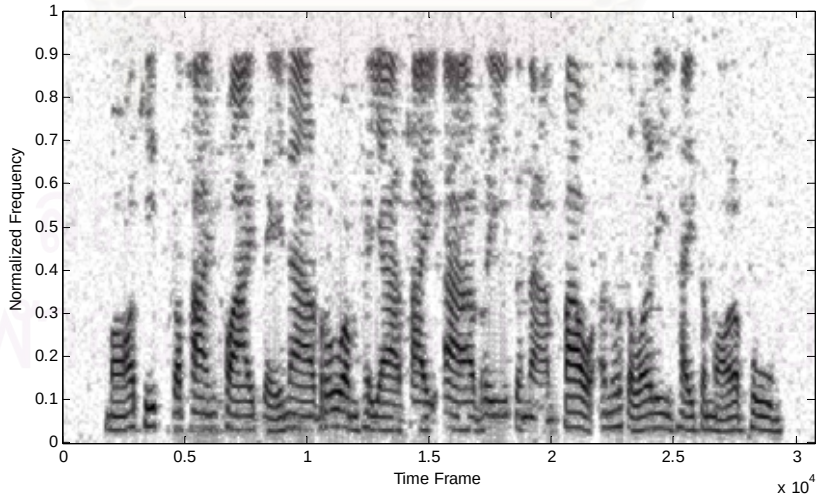
ตารางที่ 4.21 NA ของการทดลองเปรียบเทียบ AECNS และ AENS ในช่วง DTS

NA (dB)		
Input SNR	5 dB	10 dB
AECNS	17.2833	15.3579
Proposed AENS	21.0343	19.8737

จะเห็นว่าวิธีการ AENS ที่นำเสนอยังคงให้ความผิดเพี้ยนของเสียงพูดที่มากกว่า AECNS อยู่ แต่ทั้งนี้ความซับซ้อนในการคำนวณของ AENS ที่นำเสนอก็ยังน้อยกว่า AECNS อยู่มาก (หัวข้อย่อยที่ 3.3)



(ก)



(ข)

รูปที่ 4.12 สเปกโทรแกรมเสียงพูดที่ถูกปรับปรุงจากเสียงพูดที่ถูกรบกวนที่ระดับ SNR 10 dB โดย (ก) AECNS (ข) Proposed AENS ที่ใช้ MTSSP ช่วง DTS เป็นเช่นเดียวกับใน รูปที่ 4.10 ของการทดลองในหัวข้อย่อยที่ 4.4.1.2

บทที่ 5

สรุปการวิจัยและข้อเสนอแนะ

สรุปการวิจัย

การเพิ่มสมรรถนะให้แก่สัญญาณเสียงพูดสามารถกระทำได้โดยลดเสียงก่อกวนที่ปะปนมากับสัญญาณเสียงที่รับมาได้ลง วิทยานิพนธ์ฉบับนี้ทำการวิจัยการลดเสียงก่อกวนโดยอาศัยเทคนิคการกดทางสเปกตรัม และมุ่งเน้นลงไปที่เสียงก่อกวนสองชนิดได้แก่ เสียงสะท้อน และเสียงรบกวนพื้นหลัง ที่มักเกิดขึ้นในการสื่อสารแบบแฮนด์ฟรี หลักการขอการกดทางสเปกตรัมอาศัยการทำงานในโดเมน STFT ซึ่งมีข้อสมมุติฐานคือ สเปกตรัมสัญญาณ ณ แต่ละองค์ประกอบทางความถี่ในโดเมน STFT มีความอิสระจากกัน ทำให้สามารถ “กด” สเปกตรัมสัญญาณในแต่ละองค์ประกอบทางความถี่ได้อย่างอิสระ โดยขึ้นกับเฉพาะค่า SDR ในแต่ละองค์ประกอบทางความถี่นั้นๆ เท่านั้น หลักการกดดังกล่าวนี้สามารถนำไปประยุกต์ใช้กับ การแปลงแบบอื่นได้เช่นกัน อาทิเช่น Wavelet Transform และ Karhunen-Loève Transform เป็นต้น อย่างไรก็ตามเนื่องจากสปกตรัมสัญญาณที่ได้จากการแปลง STFT อยู่ในรูปของจำนวนเชิงซ้อนหรือปริมาณเวกเตอร์ที่มีทั้งขนาดและเฟสทำให้ SDR สำหรับหลักการกดทางสเปกตรัม ถูกนิยามเป็น 2 ปริมาณได้แก่ a priori SDR และ posterior SDR โดย a priori SDR เป็นปริมาณหลักที่บ่งชี้ถึงอัตราส่วนระหว่างกำลังเสียงพูดและกำลังเสียงก่อกวน ส่วน posterior SDR เป็นปริมาณที่สะท้อนถึงผลของการรวมกันแบบเวกเตอร์ของสเปกตรัมเสียงพูดและสเปกตรัมเสียงก่อกวน

การออกแบบตัวกดทางสเปกตรัมเริ่มจากการพยายามหาค่าผิดเพี้ยนระหว่างสเปกตรัมเสียงพูดและค่าประมาณสเปกตรัมเสียงพูดที่ต่ำที่สุด โดยสามารถเลือกใช้ค่าผิดเพี้ยนได้หลากหลายรูปแบบขึ้นกับผู้ออกแบบ อาทิเช่น d_{SE} , d_{LSA} และ d_{SP} เป็นต้น การวิเคราะห์หาค่าผิดเพี้ยนดังกล่าวสามารถกระทำอย่างอิสระจากกันในแต่ละองค์ประกอบทางความถี่เนื่องจากคุณสมบัติของ STFT ที่กล่าวไว้ข้างต้น ผลจากการวิเคราะห์ค่าผิดเพี้ยนในรูปแบบต่างๆ ทำให้ได้มาซึ่งค่าประมาณสเปกตรัมเสียงพูด $\hat{S}(k,\ell) = G_{\eta}(k,\ell)Y(k,\ell)$ โดยที่ Spectral Gain $G_{\eta} \in [0,1]$ คือปริมาณที่ใช้ “กด” สเปกตรัมสัญญาณไมโครโฟน $Y(k,\ell)$ และเป็นฟังก์ชันของ a priori SDR และ posterior SDR ในแต่ละองค์ประกอบทางความถี่นั้นๆ การทำงานในลักษณะนี้สามารถตีความได้ว่าส่วนที่ถูกกดจะมากหรือน้อยขึ้นอยู่กับ SDR ทั้งสองเช่นเดียวกับที่กล่าวไว้ในหลักการของการกดทางสเปกตรัมในย่อหน้าแรกของหัวข้อนี้นั่นเอง

ในทางปฏิบัติแล้ว a priori SDR และ posterior SDR ไม่สามารถทำการวัดได้โดยตรง ดังนั้นจึงจำเป็นต้องทำการประมาณค่าปริมาณทั้งสอง ปริมาณที่จำเป็นอย่างยิ่งต่อการประมาณค่า SDR ทั้งสองได้แก่ ค่าความหนาแน่นสเปกตรัมกำลัง (Power Spectral Density, PSD) ของเสียงก่อกวนแต่ละชนิด หลังจากทราบค่า PSD ของเสียงก่อกวนแต่ละชนิดแล้ว posterior SDR สามารถถูกประมาณออกมาได้โดยตรง แต่ a priori SDR ยังคงต้องได้รับการประมาณด้วยวิธีการที่เหมาะสมต่อไปอีก วิธีการประมาณ a priori SDR ที่มีชื่อเสียงในอดีตได้แก่ Decision Direct (DD) ซึ่งสามารถลดผลของเสียงรบกวนตกค้างที่เรียกว่า Musical Noise ได้เป็นอย่างดี

การกดทางสเปกตรัมจะเป็นไปอย่างแม่นยำตามที่ออกแบบไว้หรือไม่นั้น ขึ้นอยู่กับว่าสามารถประมาณค่า a priori SDR และ posterior SDR ได้อย่างถูกต้องมากน้อยเพียงใด ดังนั้นการประมาณค่าปริมาณทั้งสองจึงเป็นสิ่งซึ่งสำคัญไม่แพ้การออกแบบตัวกดทางสเปกตรัมหรือ Spectral Gain

วิทยานิพนธ์ฉบับนี้ทำการพัฒนากระบวนการ 2 กระบวนการได้แก่ การประมาณหาค่าความหนาแน่นสเปกตรัมกำลังเสียงสะท้อน (Echo Power Spectral Density, EPSD) และการประมาณค่า a priori SDR

การประมาณค่า EPSD ที่นำเสนอเริ่มจากการมองว่าสามารถทำการประมาณ EPSD ได้จากค่าสเปกตรัมกำลังเสียงสะท้อน ณ ขณะนั้น (Instantaneous Echo Power Spectrum, IEPS $\varepsilon(k,\ell)$) โดยมีสุมมติฐานว่า IEPS มีความสัมพันธ์แบบเชิงเส้นกับค่าสเปกตรัมกำลังเสียงทางห้องไกล ณ ขณะนั้น (Instantaneous Far-End Speech Power Spectrum, IFPS $\varphi(k,\ell)$) โดยสามารถเขียนได้ว่า $\varepsilon(k,\ell)=h(k,\ell)*\varphi(k,\ell)$ ดังนั้นเมื่อสมมุติให้ผลตอบสนองเชิงเส้น $h(k,\ell)$ เป็นผลตอบสนองแบบจำกัดแล้ว จะสามารถทำการหาค่าประมาณของผลตอบสนอง $\hat{h}(k,\ell)$ ได้โดยอาศัยหลักการทำงานของวงจรกรองแบบปรับตัว และสามารถหาค่าประมาณ IEPS $\hat{\varepsilon}(k,\ell)$ ได้จาก $\hat{\varepsilon}(k,\ell)=\hat{h}(k,\ell)*\varphi(k,\ell)$ ก่อนจะทำการประมาณอีกครั้งว่า $\mathrm{EPSD}(k,\ell)\approx\hat{\varepsilon}(k,\ell)$ ผลจากการประมาณ EPSD ดังกล่าวทำให้ได้ค่าประมาณ EPSD ที่ใกล้เคียงกับความเป็นจริงมากยิ่งขึ้น โดยแลกมาด้วยความซับซ้อนในการคำนวณที่เพิ่มมากขึ้นจากวิธีที่นำเสนอใน [53] เพียงเล็กน้อย ความซับซ้อนในการคำนวณของการประมาณค่า EPSD ที่นำเสนอนี้ยังคงน้อยมากเมื่อเทียบกับความซับซ้อนในการคำนวณที่ใช้ในการประมาณสัญญาณเสียงสะท้อนในระบบ AEC ทั้งนี้เนื่องจาก การประมาณค่า EPSD ไม่ได้อาศัยข้อมูลทางเฟสในการประมาณ และไม่ต้องประมาณค่าเฟสสเปกตรัมเสียงสะท้อนเหมือนดังเช่น AEC อีกด้วย

การประมาณค่า a priori SDR ที่นำเสนอเกิดจากความพยายามที่จะรักษาส่วนของเสียงพูดในแต่ละองค์ประกอบทางความถี่เอาไว้ให้ได้มากยิ่งขึ้น ในขณะที่ไม่ก่อให้เกิดผลของ Musical Noise สมการช่วงเปลี่ยน (Transition Equation, TE) ถูกตั้งขึ้นเพื่อใช้ในการพัฒนาการประมาณค่า a priori SDR ที่นำเสนอ การพัฒนาเริ่มจากเสนอให้ใช้ Spectral Gain $G_{\psi}=G_{\mathrm{SP}}$ ในขั้นที่สองของการประมาณค่า a priori SNR แบบ Two-Step แทน $G_{\psi}=G_{\mathrm{SE}}$ (Two-Step Wiener, TSW) โดยเรียกวิธีการประมาณดังกล่าวนี้ว่า Two-Step Spectral Power (TSSP) นอกจากนี้ยังเสนอให้ใช้รูปแบบปรับปรุงของทั้งสองวิธีการเพื่อเพิ่มความสามารถในการรักษาส่วนของเสียงพูดอีกด้วย จากการทดลองรูปแบบปรับปรุงของวิธีการประมาณ a priori SDR พบว่าให้ผลการติดตามค่า Instantaneous SNR และค่า Instantaneous SER ที่เหมาะสม และสามารถรักษาส่วนของเสียงพูดในแต่ละองค์ประกอบทางความถี่ไว้ได้มากยิ่งขึ้น แต่อย่างไรก็ตาม เนื่องจากเสียพูดที่ถูกปรับปรุง ณ องค์ประกอบทางความถี่สูง มีความต่อเนื่องค่อนข้างน้อย โดยเฉพาะอย่างยิ่งในสภาพแวดล้อมที่ Global SNR มีค่าต่ำ ทำให้ส่วนของเสียงพูดที่รักษาไว้ได้เกิดความไม่ต่อเนื่องและทำให้เกิดเสียงรบกวนตกค้างแบบ Musical Noise ที่เกิดจากส่วนของเสียงพูดนั้นเอง ดังนั้นจึงเสนอให้ใช้รูปแบบปรับปรุงนี้กับเฉพาะ องค์ประกอบทางความถี่ต่ำเท่านั้นเพื่อลดผลของเสียงรบกวนตกค้างแบบ Musical Noise ดังกล่าว การประมาณค่า a priori SDR โดยใช้รูปแบบการปรับปรุงเฉพาะองค์ประกอบทางความถี่ต่ำนี้ให้ผลการทดลอง NS ที่ไม่ต่างกับรูปแบบปกติมากนัก เนื่องจากสำหรับสถานการณ์ที่ เสียงก่อกวนเป็นเสียงรบกวนสีขาวแล้ว เสียงพูดมีพลังงานที่สูงกว่าเสียงรบกวนค่อนข้างมากในองค์ประกอบทางความถี่ต่ำ ทำให้ไม่ว่าการประมาณแบบใดก็สามารถติดตามกค่า Instantaneous SNR ได้ใกล้เคียงกัน แต่เมื่อทดลองในสถานการณ์ที่มีทั้งเสียงสะท้อนและเสียงรบกวนแล้ว รูปแบบปรับปรุงดังกล่าวให้ผลที่ดีกว่ารูปแบบปกติค่อนข้างมาก ทั้งนี้เนื่องจากสำหรับเสียงก่อกวนที่เป็นเสียงสะท้อนแล้ว เสียงพูด

ทางห้องใกล้จะมีพลังงานไม่มากเทียบกับเสียงสะท้อนไม่ว่าในองค์ประกอบทางความถี่ต่ำ หรือสูง ดังนั้นจึงจำเป็นต้องใช้การประมาณที่มีความไวในการติดตามค่า Instantaneous SER จึงจะสามารถติดตามการเปลี่ยนแปลงค่า Instantaneous SER ในสถานการณ์ดังกล่าวได้ทัน

ผลการทดลองด้วยเสียงพูดจริง กับสถานการณ์ห้องใกล้จำลอง โดยกำหนดให้ค่าระยะเวลาสะท้อนกลับ อยู่ที่ 32 ms หรือ 256 ตัวอย่าง ณ ความถี่ชักตัวอย่าง 8 kHz ที่ระดับ Global SNR ต่างๆ ชี้ให้เห็นว่าสำหรับในช่วง STS นั้น วิธี AENS ที่นำเสนอสามารถกดเสียงสะท้อนลงได้มากกว่าวิธี AENS ใน [53] โดยแลกมาด้วยความซับซ้อนในการคำนวณที่เพิ่มมากขึ้นเล็กน้อย และระยะเวลาช่วงหนึ่งก่อนการลู่เข้า โดยระยะเวลาดังกล่าวถือว่าน้อยมากเมื่อเทียบกับระยะเวลาที่วงจรกรองแบบปรับตัวในระบบ AEC ใช้ ดังนั้นจึงไม่เป็นปัญหาสำหรับระบบการสื่อสารแบบแฮนด์ฟรีที่พิจารณา สำหรับในช่วง DTS วิธี AENS ที่นำเสนอให้ผลความผิดเพี้ยนของเสียงพูดที่ถูกปรับปรุงน้อยลงกว่าวิธี AENS ใน [53] และจากการทดสอบเชิงอัตวิสัยพบว่าเสียงพูดที่ถูกปรับปรุงจาก AENS ที่นำเสนอมีความเป็นธรรมชาติมากกว่าเสียงพูดที่ถูกปรับปรุงจาก AENS ใน [53] แต่อย่างไรก็ตามเสียงพูดที่ถูกปรับปรุงจาก AENS ที่นำเสนอยังคงมีความผิดเพี้ยนจากเสียงพูดสะอาดมากกว่าเสียงพูดที่ถูกปรับปรุงจากระบบร่วม AECNS อยู่ แต่ทั้งนี้ความซับซ้อนในการคำนวณของ AECNS นั้นมากกว่า AENS ที่นำเสนออยู่มาก

ข้อเสนอแนะ

ดังที่ได้กล่าวมา ระบบการลดเสียงสะท้อนและเสียงรบกวน (AENR) มีบทบาท และความสำคัญอย่างยิ่งในระบบการสื่อสารแบบแฮนด์ฟรี โดยวิธีการที่ถูกนำเสนอในช่วงแรกของการพัฒนาระบบดังกล่าวเป็นการหยิบเอาการหักล้างเสียงสะท้อน (AEC) และการกดเสียงรบกวน (NS) มาออกแบบเพื่อให้ทั้งสองวิธีสามารถทำงานร่วมกันได้อย่างมีประสิทธิภาพ (NSAEC, AECNS และ AECP) ต่อมามีผู้เสนอให้ใช้ AENR ที่อาศัยโครงสร้างของเทคนิคการกดทางสเปกตรัม (SS) ในการลดทั้งเสียงสะท้อนและเสียงรบกวน โดยเรียกว่า วิธีการกดเสียงสะท้อนและเสียงรบกวน (AENS) สิ่งที่นำเสนอในวิทยานิพนธ์ฉบับนี้คือการเพิ่มประสิทธิภาพให้กับโครงสร้าง AENS เนื่องจากเล็งเห็นว่าความซับซ้อนในการคำนวณของโครงสร้างดังกล่าวอยู่ในระดับที่ต่ำเมื่อเทียบกับระบบร่วมระหว่าง AEC และ NS ทำให้สามารถนำไปใช้ในงานประยุกต์ที่ไม่สามารถรองรับความซับซ้อนที่สูงได้

จุดประสงค์ของ AENR ในระบบการสื่อสารแบบแฮนด์ฟรี คือสามารถลดเสียงสะท้อนและเสียงรบกวนลงให้ได้มากที่สุด โดยยังคงความเป็นธรรมชาติของเสียงพูดเอาไว้ให้ได้มากที่สุด เทคนิคที่ใช้ใน AENR จะต้องเป็นเทคนิคที่เหมาะสมกับจุดประสงค์ดังกล่าว ทั้งนี้จะเห็นว่า AENR ที่ถูกบรรยายถึงในวิทยานิพนธ์ฉบับนี้อาศัยโครงสร้างที่ตั้งอยู่บนพื้นฐานของเทคนิคเพียง 2 เทคนิคหลัก ได้แก่ การหักล้าง (Cancellation) และการกดทางสเปกตรัม (Spectral Suppression) เท่านั้น นั่นแปลว่ายังคงมีเทคนิคในการลดเสียงก่อกวนอยู่อีกเป็นจำนวนมาก (ดังเช่นที่ได้บรรยายถึงในหัวข้อย่อยที่ 2.1 และ 2.2 รวมทั้งที่ไม่ได้กล่าวถึงในวิทยานิพนธ์ฉบับนี้) ที่สามารถนำมาออกแบบและพัฒนา AENR ได้ การพัฒนาที่อาจเป็นไปได้จึงอาจเป็นได้ 3 แนวทางหลักๆ ได้แก่ 1) เพิ่มประสิทธิภาพให้กับเทคนิคที่มีอยู่เดิม 2) นำเทคนิคใหม่ในการลดเสียงก่อกวนมาออกแบบโครงสร้างใหม่ให้กับ AENR และ 3) คิดค้นเทคนิคในการลดเสียงสะท้อนและ/หรือเสียงรบกวนขึ้นใหม่ อย่างไรก็ตามเนื่องจากวิทยานิพนธ์ฉบับนี้มุ่งเน้นการปรับปรุงประสิทธภาพโดยยังคงรักษาความซับซ้อนในการคำนวณไว้ในระดับที่ไม่สูงมากนัก ดังนั้นจึงขอนำเสนอวิธีการที่อาศัยเทคนิค SS เป็นหลัก

เทคนิค SS เป็นเทคนิคที่ให้ผลของการลดลงของเสียงก่อกวนที่ดี และมีความซับซ้อนในการคำนวณต่ำ อย่างไรก็ตามเทตนิตดังกล่าวนี้ยังคงมีข้อบกพร่องอันได้แก่ ความผิดเพี้ยนของเสียงพูดที่ถูกปรับปรุง ซึ่งทั้งนี้เป็นเพราะส่วนของเสียงพูดในบางองค์ประกอบทางความถี่มีพลังงานค่อนข้างต่ำเมื่อเปรียบเทียบกับพลังงานของเสียงก่อกวน ทำให้ไม่สามารถทำการประมาณค่า a priori SDR ณ องค์ประกอบทางความถี่ดังกล่าวได้อย่างเหมาะสม เมื่อข้อมูลที่รับมาได้มีเพียงสัญญาณเสียงพูดที่ถูกก่อกวน ส่งผลให้ค่า Spectral Gain ที่ได้มีค่าน้อยเกินความเป็นจริงไป และทำให้เกิดการกดมากเกินไปนั่นเอง การกดมากเกินไปนี้จะทำให้ส่วนของเสียงพูดที่มีพลังงานน้อยในองค์ประกอบทางความถี่ดังกล่าว ถูกกดทิ้งไปหมด สัญญาณขาออกที่ได้จึงปราศจากส่วนของเสียงพูดดังกล่าวนั้น และเป็นเหตุให้เกิดการผิดเพี้ยนไปจากเสียงพูดสะอาดนั่นเอง อย่างไรก็ตามสถานการณ์ข้างต้นนี้จะเกิดขึ้นในเฉพาะบางองค์ประกอบทางความถี่เท่านั้น ยังมีองค์ประกอบทางความถี่อื่นๆ อีกเป็นจำนวนมากที่สามารถประมาณค่า a priori SDR ได้อย่างเหมาะสม และไม่ประสบกับการกดมากเกินไป ดังนั้นสัญญาณขาออกของเทคนิค SS จึงยังคงสามารถรับฟังได้อย่างรู้เรื่อง ความผิดเพี้ยนที่เกิดขึ้น อาจปรากฏให้เห็นในรูปของความไม่เป็นธรรมชาติของเสียงพูดที่ถูกปรับปรุง จากการวิจัยเมื่อไม่นานมานี้ (พ.ศ. 2550) พบว่า ส่วนสำคัญที่ทำให้เกิดความเป็นธรรมชาติของเสียงพูดขึ้น คือ ส่วนของเสียงพูดชั่วครู่ (Transient Speech Components) ซึ่งถูกนิยามอย่างระเอียดใน [59] ซึ่งส่วนของเสียงพูดชั่วครู่นี้เป็นส่วนจะมีพลังงานไม่มาก ดังนั้นเมื่อถูกก่อกวนด้วยสัญญาณก่อกวนที่มีค่าสูงจะทำให้ค่าประมาณ a priori SDR ในองค์ประกอบทางความถี่ที่มีส่วนของเสียงพูดชั่วครู่อยู่นี้เกิดการผิดพลาดขึ้น ส่งผลให้เกิดเกิดการกดมากเกินไปจนทำให้ส่วนของเสียงพูดชั่วครู่ดังกล่าวนี้สูญหายไป และเกิดความไม่เป็นธรรมชาติของสัญญาณเสียงขาออกขึ้น อย่างไรก็ตามสำหรับสถานการณ์ที่เสียงก่อกวนมีพลังงานต่ำกว่าส่วนของเสียงพูดชั่วครู่ในแต่ละองค์ประกอบทางความถี่แล้ว วิธี SS จะยังคงรักษาส่วนของเสียงชั่วครู่เอาไว้ได้ทำให้ไม่ประสบกับปัญหาความไม่เป็นธรรมชาติของเสียงพูดเมื่อใช้วิธี SS กับสถานการณ์ที่เสียงก่อกวนมีพลังงานต่ำ

จากที่กล่าวมาจะเห็นได้ว่าหากสามารถพัฒนาวิธี SS ให้สามารถรักษาส่วนของเสียงพูดชั่วครู่เอาไว้ได้อย่างดียิ่งขึ้นแล้ว ก็จะสามารถรักษาความเป็นธรรมชาติของเสียงพูดเอาไว้ได้อย่างดียิ่งขึ้นได้ ดังนั้นการพัฒนาการเพิ่มสมรรถนะเสียงพูดโดยอาศัยวิธีการ SS ให้คงความเป็นธรรมชาติของเสียงพูดเอาไว้ได้มากยิ่งขึ้น จึงควรมีการศึกษาเกี่ยวกับส่วนของเสียงพูดชั่วครู่เพิ่มเติมต่อไป และนอกจากนี้ยังควรศึกษาปัจจัยอื่นๆ ที่บ่งบอกถึงความผิดเพี้ยนของเสียงพูดที่แท้จริงเพิ่มขึ้นอีกด้วย

วิธีการที่ถูกกล่าวถึงในวิทยานิพนธ์ฉบับนี้เป็นเพียงเทคนิคการลดเสียงก่อกวนที่อาศัยไมโครโฟนเพียงตัวเดียว ในส่วนของเทคนิคที่อาศัยไมโครโฟนหลายตัวยังคงมีสิ่งที่น่าสนใจและไม่สามารถพบได้ในเทคนิคที่อาศัยไมโครโฟนตัวเดียว อยู่อีกมาก อาทิเช่น เรื่องของทิศทาง (Direction) และเรื่องของความอิสระของข้อมูลที่รับมาได้ (Independent) เป็นต้น ดังนั้นจึงเป็นการน่าสนใจเป็นอย่างยิ่งที่จะทำการศึกษาเกี่ยวกับเทคนิคที่อาศัยไมโครโฟนหลายตัวเพื่อเพิ่มสมรรถภาพเสียงพูดให้กับระบบการสื่อสารทางเสียงแบบแฮนด์ฟรี

รายการอ้างอิง

[1] Benesty, J., Makino, S., and Chen, J., eds. Speech Enhancement. Springer, 2005.

[2] Haykin, S. Unsupervised Adaptive Filtering. 2vols. New York: John Wiley & Sons, 2000.

[3] Haykin, S. Adaptive Filter Theory. 4th ed. New Jersey: Prentice-Hall, 1996.

[4] Shewchuk, J. R. An introduction to the conjugate gradient method without the agonizing pain [Computer file]. 2005. Available from: http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient-figs.pdf [2006, April 12]

[5] Boll, S. F. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech and Signal Processing. ASSP-26 (April 1979): 113-120.

[6] McAulay, R. J., and Malpass, M. L. Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech and Signal Processing. ASSP-28 (April 1979): 137-145.

[7] Ephraim, Y., and Malah, D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech and Signal Processing. ASSP-32 (December 1984): 1109-1121.

[8] Ephraim, Y., and Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech and Signal Processing. ASSP-33, 2 (April 1985): 443-445.

[9] Cappe, O. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech and Audio Processing. 2, 2 (April 1994): 345-349.

[10] Patrick, J., Wolfe, J. G., and Simon, J. G. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. EURASIP Journal on Applied Signal Processing. 2003, 10 (2003): 1043-1051.

[11] Cohen, I., and Berdugo, B. Speech enhancement for non-stationary noise environments. Signal Processing. 81 (November 2001): 2403-2418.

[12] Cohen, I. Relaxed Statistical Model for Speech Enhancement and a Priori SNR Estimation. IEEE Trans. Speech and Audio Processing. 13, 5 (September 2005): 870-881.

[13] Ephraim, Y., and Cohen, I. Recent advancements in speech enhancement. The Electrical Engineering Handbook. CRC Press, 2006.

[14] Martin, R. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans. Speech and Audio Processing. 13, 5, 2 (September 2005): 845-856.

[15] Martin, R. Noise power spectral density estimation based on optimal and minimum statistics. IEEE Trans. Speech and Audio Processing. 9, 5 (July 2001): 504-512.

[16] Lotter, T., and Vary, P. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. EURASIP Journal on Applied Signal Processing. 2005, 7(2005): 1110-1126.

[17] Tsukamoto, Y., Kawwamura, A., and Iiguni, Y. Speech enhancement based on MAP estimation using a variable speech distribution. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. E90-A, 8 (July 2007): 1587-1593.

[18] Plapous, C., Marro, C., and Scalart, P. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans. Speech and Audio Processing. 14, 6 (September 2006): 2098-2108.

[19] Scalart, P., and Filho, J. V. Speech enhancement based on a priori signal to noise estimation. in Proceedings ICASSP-96, Acoust., Speech and Signal Processing. (May 1996): 629-632.

[20] Hasan, M. K., Salahuddin, S., and Khan, M. R. A modified a priori SNR for speech enhancement using spectral subtraction rules. IEEE, Signal Processing Letters. 11, 4 (April 2004): 450-453.

[21] Kato, M., Serizawa, M., Toki, N., Mori, U., Morishita, Y., and Hayashi, K. Noise suppression with high speech quality based on weighted noise estimation for 3G handsets. NEC Res. & Develop. 44 (October 2003): 66-73.

[22] Thoonsaengngam, R. and Tangsangiumvisai, N. On improvement of the a priori SNR estimation via gain modification for speech enhancement. will be published in Proceedings EECON-30, DS (October 2007)

[23] Lin, L., Holmes, W. H., and Ambikairajah, E. Adaptive noise estimation algorithm for speech enhancement. IEEE, Electronic Letters. 39, 9 (April 2003): 754-755.

[24] Wu, B-F., and Wang, K-C. Noise spectrum estimation with entropy-based VAD in non-stationary environment. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. E89-A, 2 (February 2006): 479-485.

[25] Wang, D., and Lim, J. The unimportance of phase in speech enhancement. IEEE Trans. Acoust., Speech and Signal Processing. 30, 4 (August 1982): 679-681.

[26] Whitmal, N. A., Rutledge, J. C., and Cohen, J. Wavelet-baesd noise reduction. in Proceedings ICASSP-95, Acoust., Speech, and Signal Processing. (May 1995): 3003-3006.

[27] Whitmal, N. A., Rutledge, J. C., and Cohen, J. Reducing correlated noise in digital hearing aids. IEEE. Engineering in Medicine and Biology Magazine. 15, 5 (October 1996): 88-96.

[28] Ephraim, Y., and Van Trees, H. L. A signal subspace approach for speech enhancement. IEEE Trans. Speech and Audio Processing. 3, 4 (July 1995): 251-266.

[29] Widrow, B. U.S. Patent 2005075866, 28 (April. 2005)

[30] Sasaoka, N., Sumi, K., Itoh, Y., and Fujji, K. A study on noise reduction system based on ALE and noise reconstruction filter in Proc. ISPACS 2004, Intelligent Signal Processing and Communication Systems. (November 2004): 305-308.

[31] Paliwal, K., and Basu, A. A speech enhancement method based on Kalman filtering. in Proceedings ICASSP-87, Acoust., Speech, and Signal Processing.(May 1987): 177-180.

[32] Gannot, S., Burshtein, D., and Weinstein, E. Iterative and sequential Kalman filter-based speech enhancement algorithms. IEEE Trans. Speech and Audio Processing. 6, 4 (July 2006): 373-385.

[33] Hansler, E. Adaptive echo compensation applied to the hands-free telephone problem. in Proc IEEE int. Circuits and Systems. 1 (May 1990): 279-282.

[34] Glentis, G.-O., Berberidis, K. and Theodoridis, S. Efficient least squares adaptive algorithms for FIR transversal filtering. IEEE Signal Processing Magazine. 16, 4 (July 1999): 13-41.

[35] Gay, S. L. A fast converging, low complexity adaptive filtering algorithm. IEEE Workshop, Application of Signal Processing to Audio and Acoustics. (October 1993): 4-7.

[36] Shynk, J. J., and Roy, S. The LMS algorithm with momentum updating. IEEE International Symposium. Circuits and Systems. 3 (July 1998): 2651-2654.

[37] Grant, S., and Gay, S. L. A multiple principle components based adaptive filter. in Proc. Conference Record of The Thirty-Eighth Asilomar Conference, Signals, Systems and Computers. (November 2004): 945-949.

[38] Shynk, J. J. Frequency-domain and multirate adaptive filtering. IEEE Signal Processing Magazine. 9, 1 (January 1992): 14-37.

[39] Soo, J-S., and Pang, K. K. Multidelay block frequency domain adaptive filter. IEEE Trans. Speech and Audio Processing. 38, 2 (February 1990): 373-376.

[40] Bendel, Y., Burshtein, D., Shalvi, O., and Weinstein, E. Delayless frequency domain acoustic echo cancellation. IEEE Trans. Speech and Audio Processing. 9, 5 (July 2001): 589-597.

[41] Jin, Q., Luo, Z-Q., and Wong, K. M. Optimum filter banks for signal decomposition and its application in adaptive eco cancellation. IEEE Trans. Speech and Audio Processing. 44, 7 (July 1996): 1669-1680.

[42] Chhetri, A. S., Stokes, J. W. and Florencio, D. A. Acoustic echo cancellation for high noise environments. in Proc. 2006 IEEE International, Multimedia and Expo. (July 2006): 905-908.

[43] Hua, Y., Xiu-Yin, C., and Bo-Xiu, W. A robust adaptive filtering algorithm with variable step size. in Proc. ICC-90, Communications. (April 1990): 636-640.

[44] Ochiai, K., Araseki, T. and Ogihara, T. Echo canceller with two echo path models. IEEE Trans. Communication. 25, 6 (June 1977): 589-595.

[45] Duttweiler, D. L. Proportionate normalized least-mean-squares adaptation in echo cancellers. IEEE Trans. Speech and Audio Processing. 8 (September 2000): 508-518.

[46] Avendano, C. Acoustic echo suppression in the STFT domain. IEEE Workshop, Application of Signal Processing to Audio and Acoustics. (October 2001): 175-178.

[47] Faller, C., and Chen, J. Suppression acoustic echo in a spectral envelope space. IEEE Trans. Speech and Audio Processing. 13, 5 (September 2005): 1048-1062.

[48] Martin, R., and Vary, P. Combined Acoustic Echo Control and noise reduction for hands-free telephony state of the art and perspectives. in Proc. EUSIPCO '96. (September 1996): 1107-1110.

[49] Gustafsson, S., Martin, R., and Vary, P. Combined acoustic echo control and noise reduction for hands-free telephony. Signal Processing. 64 (January 1998): 21-32.

[50] Gustafsson, S., Martin, R., Jax, P., and Vary, P. A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. IEEE Trans. Speech and Audio Processing. 10, 5 (July 2002): 245-256.

[51] Jeannes, R. L. B., Scalart, P., Faucon, G., and Beaugeant, C. Combined noise and echo reduction in hands-free systems: A survey. IEEE Trans. Speech and Audio Processing. 9, 8 (November 2001): 808-820.

[52] Guelou, Y., Benamar, A., and Scalart, P. Analysis of two structures for combined acoustic echo cancellation and noise reduction. in Proceedings ICASSP-96, Acoust., Speech and Signal Processing. (May 1996): 637-340.

[53] Beaugeant, C., Turbin, V., Scalart, P., and Gilloire, A. New optimal filtering approaches for hands-free telecommunication terminals. Signal Processing. 64 (January 1998): 33-47.

[54] Habets, E. A. P., Cohen, I., and Gannot, S. MMSE log-spectral amplitude estimator for multiple interferences. in Proc. IWAENC 2006. (September 2006)

[55] Ichikawam, O., and Nishimura, M. Simultaneous Adaptation of echo cancellation and spectral subtraction for in-car speech recognition. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. E88-A, 7 (July 2005): 1732-1738.

[56] ITU Telecommunication Standardization Sector. ITU-T Recommendation G.167 [Computer file]. 1993. Available from: http://www.itu.int/rec/T-REC-G.167-199303-W/en [2005, November 12]

[57] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data viw the EM algorithm. Journal of Royal Statistical Society. Series B (Methodological). 39, 1 (1997): 1-38.

[58] Kalman, R. E. A new approach to linear filtering and prediction problems. Transactions of the ASME 82(Series D). (1960): 35-45.

[59] Tantibundhit, C., Boston, J. R., Li, C. C., Durratn, D., Shaiman, S., Kovacyk, K., and El-Jaroudi, A. Speech Enhancement using transient speech components. in Proc. ICASSP 2006 Acoustic, Speech, and Signal Processing. 1 (May 2006): 833-836.

[60] Rice University Digital Signal Processing (DSP) Group NOISE DATA [Computer file]. 1995. Available from: http://spib.rice.edu/spib/select_noise.html [2007, May 15]

[61] CCITT Recommendation E.432

ภาคผนวก

ภาคผนวก ก

วิธีพิสูจน์การประมาณค่า a priori SNR แบบ TSSP

แทนค่า Spectral Gain แบบ Spectral Power $G_{\mathrm{SP}}(k,\ell)$ ในสมการที่ (2.28) ลงในสมการที่ (3.14) จะได้

$$\hat{\xi}^{N}_{\mathrm{TSSP}}(k,\ell)=\left(\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1}\left(\frac{1}{\gamma^{N}(k,\ell)}+\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1}\right)\right)\gamma^{N}(k,\ell) \tag{ก.1}$$

เมื่อกระจาย $\gamma^{N}(k,\ell)$ เข้าไปจะได้

$$\hat{\xi}^{N}_{\mathrm{TSSP}}(k,\ell)=\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1}\left(1+\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1}\gamma^{N}(k,\ell)\right) \tag{ก.2}$$

$$\hat{\xi}^{N}_{\mathrm{TSSP}}(k,\ell)=\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1}+\left(\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1}\right)^{2}\gamma^{N}(k,\ell) \tag{ก.3}$$

บวกและลบด้วย $\xi^{N}_{\mathrm{DD}}(k,\ell)$ จะได้

$$\hat{\xi}^{N}_{\mathrm{TSSP}}(k,\ell)=\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1}+\left(\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1}\right)^{2}\gamma^{N}(k,\ell)-\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell) \tag{ก.4}$$

$$\hat{\xi}^{N}_{\mathrm{TSSP}}(k,\ell)=\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1\right)+\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)\right)^{2}\gamma^{N}(k,\ell)-\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1\right)^{2}}{\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1\right)^{2}} \tag{ก.5}$$

$$\hat{\xi}^{N}_{\mathrm{TSSP}}(k,\ell)=\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1\right)^{2}}\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1+\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)\gamma^{N}(k,\ell)-\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)\right)^{2}-2\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)-1\right) \tag{ก.6}$$

$$\hat{\xi}^{N}_{\mathrm{TSSP}}(k,\ell)=\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1\right)^{2}}\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)\left(1+\gamma^{N}(k,\ell)\right)-\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)\right)^{2}-2\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)-1\right) \tag{ก.7}$$

$$\hat{\xi}^{N}_{\mathrm{TSSP}}(k,\ell)=\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+\frac{\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)}{\left(\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)+1\right)^{2}}\left(1+\gamma^{N}(k,\ell)-\hat{\xi}^{N}_{\mathrm{DD}}(k,\ell)-2\right) \tag{ก.8}$$

$$\hat{\xi}^N_{\mathrm{TSSP}}(k,\ell) = \hat{\xi}^N_{\mathrm{DD}}(k,\ell) + \left(\frac{\hat{\xi}^N_{\mathrm{DD}}(k,\ell)}{\hat{\xi}^N_{\mathrm{DD}}(k,\ell)+1}\right)^2 \left(\left(\gamma^N(k,\ell)-1\right) - \hat{\xi}^N_{\mathrm{DD}}(k,\ell)\right) \qquad \text{(ก.9)}$$

$$\hat{\xi}^N_{\mathrm{TSSP}}(k,\ell) = \hat{\xi}^N_{\mathrm{DD}}(k,\ell) + K_g \left(\left(\gamma^N(k,\ell)-1\right) - \hat{\xi}^N_{\mathrm{DD}}(k,\ell)\right) \qquad \text{(ก.10)}$$

เนื่องจาก $\left(\gamma^N(k,\ell)-1\right)$ เปรียบได้กับค่า Instantaneous SNR เมื่อมีค่าเป็นบวก และจะเห็นว่าพจน์หลังทางขวามือจะมีความสำคัญก็ต่อเมื่อ K_g มีค่าไม่น้อยมาก ซึ่งจะเกิดขึ้นในช่วงที่มีเสียงพูดหรือช่วงที่ค่า Instantaneous SNR มีค่ามากเท่านั้น ดังนั้นจึงสามารถแทนค่าดังกล่าวด้วยค่า Instantaneous SNR $\delta^N(k,\ell)$ ได้ทำให้สมการที่ (ก.10) กลายเป็น

$$\hat{\xi}^N_{\mathrm{TSSP}}(k,\ell) = \hat{\xi}^N_{\mathrm{DD}}(k,\ell) + K_g \left(\delta^N(k,\ell) - \hat{\xi}^N_{\mathrm{DD}}(k,\ell)\right) \qquad \text{(ก.11)}$$

ดังสมการที่ (3.15)

ภาคผนวก ข

บทความที่ได้รับการเผยแพร่

1. Thoonsaengngam, R., and Tangsangiumvisai, N. On improvement of the a priori SNR estimation via gain modification for speech enhancement. will be published in Proceedings EECON-30, DS (October 2007)

On improvement of the a priori SNR estimation via gain modification for speech enhancement

R. Thoonsaengngam and N. Tangsangiumvisai
Digital Signal Processing Research Laboratory, Department of Electrical Engineering,
Faculty of Engineering, Chulalongkorn University, Phyathai Road, Pathumwan, Bangkok, 10330
E-mail: Nisachon.T@chula.ac.th

Abstract

This paper presents an improved approach for speech enhancement in single-microphone applications. The proposed Noise Suppression (NS) technique is based upon a modified a priori Signal-to-Noise Ratio (SNR) estimation. The proposed a priori SNR estimator aims at tracking very fast the rise and fall of the speech spectrum, while reducing the musical noise effect. This yields an efficient NS technique that provides minimum speech distortion, especially for the low-level speech spectrum. Simulation results demonstrate improved performance of the proposed technique in terms of the ability to preserve the enhanced speech quality.

Keywords: a priori SNR estimation, noise suppression, noise reduction, speech enhancement.

1. Introduction

In a voice communication system, Noise Suppression (NS) techniques play a very important role for speech enhancement. Most of the NS techniques are based on the estimation of the short-time spectral gain, which is a function of the *a priori* Signal-to-Noise Ratio (SNR) [1] – [4]. However, in single-microphone applications, only the noisy speech signal is available. It is therefore necessary to estimate accurately the short-time spectral gain in order to enable efficient elimination of the additive ambient noise.

An efficient method for estimating the a priori SNR is the Decision-Directed (DD) approach [1]. This method is able to reduce a disturbing artifact, known as *musical noise*, which is characterized by tones at random frequencies [6]. It was, however, investigated in [3] that the DD approach can lead to inexact short-time spectral gain due to its dependence on the speech spectrum estimation in the previous frame. As a consequence, this limits the noise suppression performance. In fact, the frame delay bias has a potential to cause the unattractive reverberation effect [5]. To alleviate these problems while maintaining the advantages of the DD approach, a Two-Step Noise Reduction (TSNR) algorithm has been introduced in [5].

When the speech spectral components are considered to be at low level at some frequencies, even during the speech activity period, the a priori SNR estimate using the DD approach can only slowly track sudden rises and falls of the speech spectrum. Although the TSNR algorithm aims to improve the tracking performance for these abrupt changes, it suffers from speech degradation of the enhanced speech spectrum due to inappropriate spectral gains at some frequencies. On the other hand, the Self-Adaptive Averaging Factor (SAAF) in [6] is developed, based on the DD approach, to track the rapid changes of these rises and falls of the speech spectrum. It, however, brings about the undesirable musical noise.

It is the purpose of this paper to introduce a modified priori SNR estimator which is essentially based upon the advantages of both TSNR [5] and SAAF [6] algorithms. The proposed estimator features significant distortion reduction of enhanced speech spectrum, while keeping the musical noise effect at minimum. In addition, it potentially offers an improvement in the noise suppression performance.

This paper is organized as follows. Section II describes the NS techniques in details. In Section III, the proposed a priori SNR estimator is presented. Simulation results based on speech signals in several noisy environments are given in Section IV, followed by the conclusions in Section V.

2. The NS techniques

Let the observed signal be described as

$$y(t) = s(t) + n(t) \tag{1}$$

where $s(t)$ and $n(t)$ denote the speech and uncorrelated additive noise signals, respectively. Basically, three main components of the NS techniques are described as follows. First, the noisy speech signal is analyzed via the Short-Time Fourier Transform (STFT), usually with short frame size to preserve the stationary of the speech signal. By applying the STFT to the noisy signal $y(t)$, we obtained its time-frequency domain as

$$Y(k,l) = S(k,l) + N(k,l) \tag{2}$$

where $k = 0,1,\ldots,N-1$ is the frequency bin index and $l = 0,1,\ldots$ is the time frame index. Next, the noise power spectrum $\lambda_N(k,l) = E\{|N(k,l)|^2\}$, where $E\{\cdot\}$ denotes the expectation operator, is estimated. By assuming that the noise signal is slowly changing with time, $\lambda_N(k,l)$ within a short frame is considered to be fairly stationary. The estimate of the noise power spectrum can be obtained recursively as

$$\hat{\lambda}_N(k,l) = \rho\,\hat{\lambda}_N(k,l-1) + (1-\rho)\,|Y(k,l)|^2 \tag{3}$$

where $\rho \in (0,1]$ is a forgetting factor. Then, a spectral gain is to be computed. Usually, the spectral gain depends on two main parameters; a priori SNR, which is defined as

$$\xi(k,l) = \frac{\lambda_S(k,l)}{\lambda_N(k,l)} = \frac{E\{|S(k,l)|^2\}}{E\{|N(k,l)|^2\}} \tag{4}$$

and the posteriori SNR, which is given by

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\lambda_N(k,l)} \tag{5}$$

Several methods have been introduced to derive the spectral gains by minimizing some distortion measures, for example, Square Error Distortion [1], Log Spectral Amplitude (LSA) Distortion [2], Spectral Power Distortion [4], etc. The corresponding spectral gains obtained by these methods; which are known as Wiener gain, LSA gain, and SP gain, respectively, are given as follows.

$$G_{Wiener}(k,l) = \frac{\xi(k,l)}{\xi(k,l)+1} \tag{6}$$

$$G_{LSA}(k,l) = \frac{\xi(k,l)}{\xi(k,l)+1}\exp\left(\frac{1}{2}\int_{v(k,l)}^{\infty}\frac{e^{-t}}{t}dt\right) \quad (7)$$

$$G_{SP}(k,l) = \sqrt{\frac{\xi(k,l)}{\xi(k,l)+1}\left(\frac{1}{\gamma(k,l)}+\frac{\xi(k,l)}{\xi(k,l)+1}\right)} \quad (8)$$

where

$$v(k,l)=\gamma(k,l)\frac{\xi(k,l)}{\xi(k,l)+1} \quad (9)$$

Consequently, the magnitude spectrum of the clean speech can be obtained as

$$|\hat{S}(k,l)| = G_{\eta}(k,l)|Y(k,l)| \quad (10)$$

where $|\cdot|$ denotes the magnitude response of any spectrum and $G_{\eta}(k,l)$ is a chosen spectral gain. The enhanced speech signal, $\hat{s}(t)$, is obtained from the inverse STFT (ISTFT) of the combination between the magnitude response $|\hat{S}(k,l)|$ in Eq.(10) and the phase of the noisy spectrum, $\angle Y(k,l)$.

2.1 A priori SNR estimation

The power spectrum of the clean speech, $\lambda_S(k,l)$, and the noise power spectrum, $\lambda_N(k,l)$, are usually unknown, it is therefore necessary to estimate the a priori SNR, $\xi(k,l)$, as given in Eq.(4). In [1], the DD approach is used to estimate $\xi(k,l)$ as

$$\hat{\xi}_{DD}(k,l) = \alpha_{DD}\tilde{\xi}(k,l-1)+(1-\alpha_{DD})\delta(k,l) \quad (11)$$

where $\tilde{\xi}(k,l-1)=|\hat{S}(k,l-1)|^2/\hat{\lambda}_N(k,l-1)$ represents the a priori SNR in the previous frame, according to the enhanced speech spectrum, and $\alpha_{DD}\in[0,1]$ is a forgetting factor. The instantaneous SNR is defined as

$$\delta(k,l)=\max(\gamma(k,l)-1,0) \quad (12)$$

2.2 Analysis of the A priori SNR estimator using the DD approach

The ability to track the rises and falls of the speech spectrum of $\hat{\xi}_{DD}(k,l)$ is analyzed through three cases [3]. In the case of low instantaneous SNR, $\delta(k,l)$, i.e. when the speech spectrum are considered to be at low level or absent, such as during speech pauses, $\hat{\xi}_{DD}(k,l)$ can be seen as a smooth version of $\delta(k,l)$, given by

$$\hat{\xi}_{DD}(k,l)=(1-\alpha_{DD})\delta(k,l) \quad (13)$$

Conversely, when $\delta(k,l)$ is high, $\hat{\xi}_{DD}(k,l)$ becomes a delay version of the instantaneous SNR, i.e.

$$\hat{\xi}_{DD}(k,l)\approx\delta(k,l-1) \quad (14)$$

When there is a change from speech pause to speech activity duration, the a priori SNR in the previous frame is approximated to be $\tilde{\xi}(k,l-1)\approx 0$, the same expression of $\hat{\xi}_{DD}(k,l)$ as given in Eq.(11) is obtained. This particular case will be called the *transition stage* and will be mainly investigated in this paper. Consequently, the expression of $\hat{\xi}_{DD}(k,l)$ for this case will be referred to as the *transition equation.* It was shown in [3] that $\hat{\xi}_{DD}(k,l)$ can track the abrupt change of the speech spectrum in the transition stage only if $\delta(k,l) >> 1/(1-\alpha_{DD})$.

2.3 The TSNR algorithm

The estimated a priori SNR is obtained in two steps [6]. First, the spectral gain, G_{ψ}, exploits the a priori estimate, $\hat{\xi}_{DD}(k,l)$, in [1]. Next, a refined version of the a priori SNR estimate is given by

$$\hat{\xi}_{TSNR}(k,l)=G_{\psi}^2(k,l)\gamma(k,l) \quad (15)$$

To obtain the enhanced speech spectrum, the spectral gain $G_{\eta}(k,l)$ in Eq.(10) can be chosen differently from $G_{\psi}(k,l)$. The authors claimed that this algorithm can further reduce the musical noise, as compared to that in [1], while providing a better tracking performance for sudden changes of the speech spectrum. The transition equation using the TSNR algorithm, when using the Wiener gain, is therefore given by

$$\hat{\xi}_{TSNR}(k,l)=\left(\frac{(1-\alpha_{DD})\delta(k,l)}{1+(1-\alpha_{DD})\delta(k,l)}\right)^2\gamma(k,l) \quad (16)$$

2.4 The SAAF technique

It is suggested in [6] that the forgetting factor of the DD approach, α_{DD}, should be self-adaptive to deal with the sudden changes of the speech spectrum, as given by

$$\alpha_{SAAF}(k,l)=\frac{1}{1+\left(\frac{\delta(k,l)-\tilde{\xi}(k,l-1)}{\delta(k,l)+1}\right)^2} \quad (17)$$

Therefore, the a priori SNR is estimated to be

$$\hat{\xi}_{SAAF}(k,l)=\alpha_{SAAF}(k,l)\tilde{\xi}(k,l-1)+(1-\alpha_{SAAF}(k,l))\delta(k,l) \quad (18)$$

and the transition equation becomes

$$\hat{\xi}_{SAAF}(k,l) = \frac{\delta(k,l)+1}{1+\left(\frac{\delta(k,l)+1}{\delta(k,l)}\right)^2} \quad (19)$$

3. The proposed a priori SNR estimation

The transition equations of the investigated a priori SNR estimates will be compared in this section to observe their tracking behavior for sudden changes of the speech spectrum. Since the SP gain minimizes spectral power distortion, its transition equation is also compared as a reference. By using SP gain in Eq.(8), the estimated a priori SNR of the TSNR algorithm is given by

$$\hat{\xi}_{TSSP}(k,l) = \hat{\xi}_{DD}(k,l) + K(k,l)\left(\delta(k,l) - \hat{\xi}_{DD}(k,l)\right) \quad (20)$$

where

$$K(k,l) = \left(\frac{\hat{\xi}_{DD}(k,l)}{\hat{\xi}_{DD}(k,l)+1}\right)^2 \quad (21)$$

The transition equation is therefore given by

$$\hat{\xi}_{TSSP}(k,l) = (1-\alpha_{DD})\delta(k,l) + \left(\frac{(1-\alpha_{DD})\delta(k,l)}{(1-\alpha_{DD})\delta(k,l)+1}\right)^2 \left(\alpha_{DD}\delta(k,l)\right) \quad (22)$$

For a speech-activity frame, i.e. there exists speech spectrum components more frequently than the noise spectrum components within that frame, the ideal a priori SNR estimate, $\hat{\xi}(k,l)$, should converge to the instantaneous SNR, $\delta(k,l)$. The spectral gain with the suggested choice of a priori SNR estimate will preserve the speech spectrum components and thus, minimum speech distortion is obtained. However, there is no perfect parameter to indicate such a frame. From the literature, the instantaneous SNR, $\delta(k,l)$, was inherently used for this purpose. On the other hand, the a priori SNR estimate, $\hat{\xi}(k,l)$, should be very small for a non-speech activity frame in order to suppress the background noise. For the transition stage from speech pause to speech activity duration, the a priori SNR estimate, $\hat{\xi}(k,l)$, should be able to track the sudden change of the instantaneous SNR $\delta(k,l)$, i.e. rapidly increasing.

The tracking behavior for those a priori SNR estimates can be observed by plotting their transition equations, with numerous values of $\delta(k,l)$. It can be seen in Fig. 1 that $\hat{\xi}_{SAAF}(k,l)$ of the SAAF technique gives the fastest convergence performance to $\delta(k,l)$. However, the a priori SNR estimate with fast rate of convergence should be obtained only at the transition stage, but not elsewhere, as mentioned above. As a result, the enhanced speech spectrum when using the SAAF technique will include the musical noise.

In order to further improve the tracking performance for sudden changes of the speech spectrum, the a priori SNR estimate is modified as follows. The parameter $K(k,l)$ in Eq.(21) should be replaced with

$$K_{proposed}(k,l) = \left(\frac{\hat{\xi}_{DD}(k,l)}{\hat{\xi}_{DD}(k,l)+\beta}\right)^{\theta} \quad (23)$$

for some constants, θ and β. Eq.(22) becomes

$$\hat{\xi}_{proposed}(k,l) = (1-\alpha_{DD})\delta(k,l) + \left(\frac{(1-\alpha_{DD})\delta(k,l)}{(1-\alpha_{DD})\delta(k,l)+\beta}\right)^{\theta} \left(\alpha_{DD}\delta(k,l)\right) \quad (24)$$

The effect of increasing θ makes $\hat{\xi}(k,l)$ converge to $\delta(k,l)$ faster with steeper slope. By reducing the values of β, $\hat{\xi}(k,l)$ approaches to $\delta(k,l)$ at lower level of $\delta(k,l)$. With the choices of $\theta = 8$ and $\beta = 0.027$, it is demonstrated in Fig. 2 that the proposed a priori SNR estimate yields faster tracking ability than that of the TSNR algorithm, but slower than that of the SAAF one.

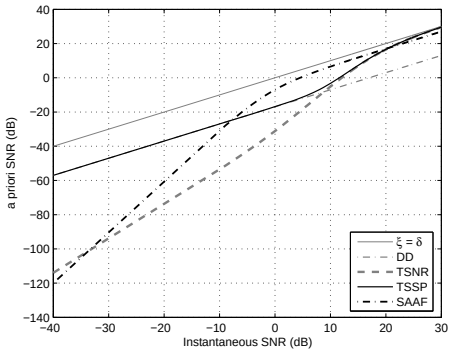


Figure 1: Tracking performance of various a priori SNR estimates.

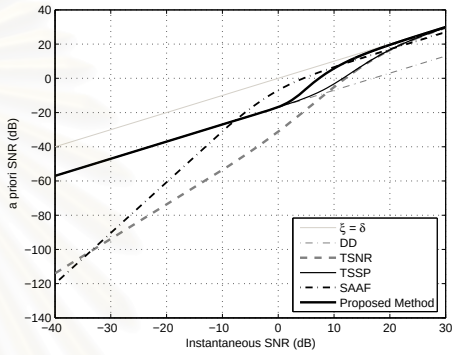


Figure 2: Comparison of the tracking performance of the proposed a priori SNR estimate with the other estimators.

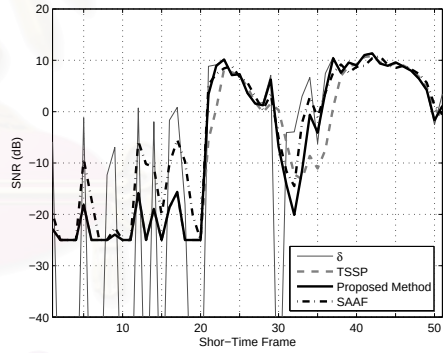


Figure 3: Tracking performance of various a priori SNR estimates against short-time frames, as compared to the instantaneous SNR.

The tracking performance is also investigated when the pure tone signal at 2 kHz is employed as a representative of speech signal and is corrupted by additive white Gaussian noise at -18dB global SNR. Fig. 3 shows that the parameters θ and β have a great impact on the tracking performance of the proposed method. Its tracking speed is much faster than the modified version of the TSNR algorithm (TSSP), and is approaching to that of the SAAF technique. In addtion, the proposed method introduces less amount of annoying musical noise due to less fluctuation of the $\hat{\xi}(k,l)$ during speech-pause duration (frame 1 to frame 20) than the SAAF technique.

4. Simulation Results

The NS techniques employing 4 other a priori SNR estimators; the decision-directed method (DD), the two-step noise reduction

algorithm using the DD approach in its first step (TSNR), the two-step noise reduction algorithm using the SP gain in its first step (TSSP), and the self-adaptive averaging factor technique (SAAF), were compared with the proposed a priori SNR estimator. The input speech signals of 3 male speakers were used with sampling rate of 8 kHz, at different input SNR levels. For speech analysis (STFT), the frame size of 32 ms was used with 50% overlap for all cases. The spectral gain, $G_\eta(k,l)$, was chosen to be the LSA gain in Eq.(7). The performance of the investigated a priori SNR estimators was evaluated in terms of speech distortion measures via the Segmental SNR (SEGSNR) [7] and the Log Spectral Distance (LSD) [7].

Table 1: SEGSNR improvement of various a priori SNR estimators (in dB)

	Input SNR			
	5 dB	*10 dB*	*15 dB*	*20 dB*
DD	5.3116	3.5590	2.0350	0.8030
TSNR	6.0202	4.2342	2.7522	1.5582
TSSP	6.3524	4.6758	3.2823	2.1580
Proposed	6.7843	5.1916	3.8306	2.7010
SAAF	5.7310	4.5607	3.4902	2.5135

Table 2: LSD improvement of various a priori SNR estimators (in dB)

	Input SNR			
	5 dB	*10 dB*	*15 dB*	*20 dB*
DD	8.6301	6.7511	5.2949	4.4185
TSNR	8.4506	6.0074	4.2516	3.3974
TSSP	8.8465	7.1159	5.7848	4.9947
Proposed	8.9676	7.4565	6.2510	5.4553
SAAF	6.5739	5.5521	4.6820	4.0264

From Table 1 and 2, it shows that the proposed algorithm gives the best performance in terms of speech distortion. This is also clearly seen from the spectrogram of the enhanced speech spectrum, as depicted in Fig. 4, that the proposed methods can preserve the low-level speech spectral components especially at high frequencies. Moreover, from informal listening tests, the proposed algorithm gives satisfied speech naturalness. It is, however, suggested here that the proposed a priori SNR estimate should be further improved in such a way that the parameters θ and β should be adaptive.

4. Conclusions

An improved NS technique has been presented in this paper, based upon a modified a priori SNR estimator. It has been demonstrated that the proposed a priori SNR estimator allows for fast tracking of the rise and fall of the speech spectrum. In addition, the musical noise effect can be reduced, while enhancing the perceptual quality of the speech spectrum, particularly for the low-level speech spectrum.

5. Acknowledgment

This work has been supported by the Cooperation Project between the department of Electrical Engineering and Private Sector for Research and Development, Chulalongkorn University.

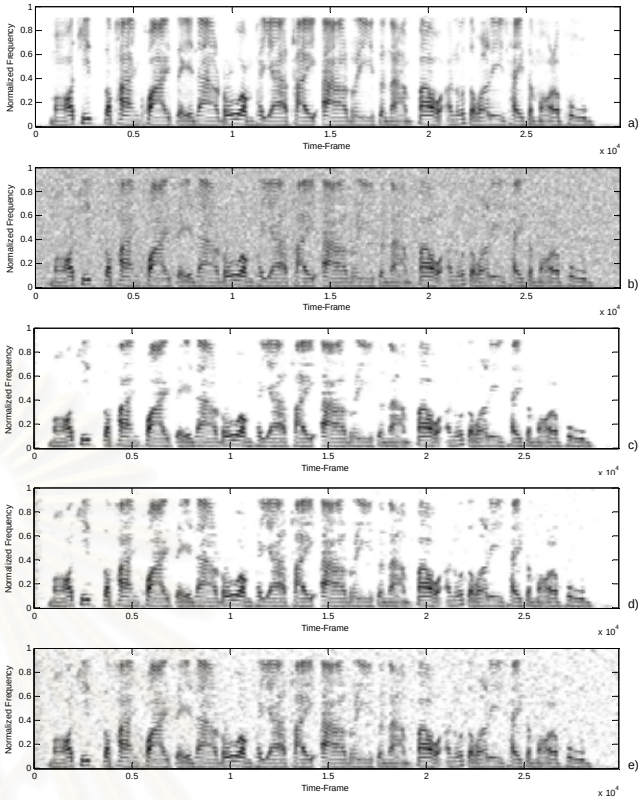


Figure 4: Spectrogram of (a) the clean speech (b) noisy speech (5 dB input SNR) and the enhanced speech spectrum using the (c) TSSP, (d) proposed, and (e) SAAF approaches.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-32, pp. 1109-1121, December 1984.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-33, No. 2, pp. 443-445, April 1985.

[3] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Processing, Vol.2, No.2, pp. 345-349, April 1994.

[4] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," EURASIP Journal on Applied Signal Processing, Vol. 2003, No. 10, pp. 1043-1051, 2003.

[5] C. Plapous, et. al., "A two-step noise reduction technique," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing (ICASSP), Montreal, Canana, Vol. 1, pp. 289-292, May 2004.

[6] Md. K. Hasan, et. al., "A modified a priori SNR for speech enhancement using spectral subtraction rules," IEEE Signal Processing Letters, Vol. 11, No. 4, pp. 450-453, April 2004.

[7] I. Cohen, "Relax Statistical Model for Speech Enhancement and a priori SNR estimation," IEEE Trans. Speech Audio Processing, Vol.13, No.5, pp. 870-881, September 2005.

ประวัติผู้เขียนวิทยานิพนธ์

นายรัฐพล ทูลแสงงาม เกิดวันที่ 4 ธันวาคม พ.ศ. 2525 ที่จังหวัดกรุงเทพมหานคร เข้าศึกษาในหลักสูตรวิศวกรรมศาสตร์บัณฑิต คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยในปีการศึกษา 2544 สำเร็จการศึกษาวิศวกรรมศาสตร์บัณฑิต สาขาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยในปีการศึกษา 2547 เข้าศึกษาต่อในหลักสูตรวิศวกรรมศาสตร์มหาบัณฑิต คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยในปีการศึกษา 2548