JAIST Repository
https://dspace.jaist.ac.jp/

Title	音環境バリアフリーのためのパワーエンベロープ処理体系
Author(s)	森田，翔太
Citation	
Issue Date	2017-03
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/14249
Rights	
Description	Supervisor:鵜木　祐史，情報科学研究科，博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

博士論文

音環境バリアフリーのための
パワーエンベロープ処理体系

森田 翔太

主指導教員　鵜木 祐史 教授

北陸先端科学技術大学院大学
情報科学研究科

2017 年 3 月

要 旨

雑音や残響といった音環境のバリアによって円滑に音声コミュニケーションができないという問題がある．従来の雑音残響除去では，音環境と人の調和が取れた状況となっていないために，人や機械が聴き取りやすい・理解しやすい音声に回復できていない．雑音残響環境における円滑な音声コミュニケーションの実現のためには，音環境と人の調和が取れる音環境バリアフリー（音環境バリアの相殺）が必要である．本研究の目的は，音環境バリアフリーのためのパワーエンベロープ処理体系を実現することである．パワーエンベロープ処理体系では，変調伝達関数（MTF）の概念に基づくことにより，信号対雑音比（SNR）と残響時間の二つのパラメータで雑音残響環境を一つの音環境として MTF で扱うことができる．さらに，原信号のパワーエンベロープの変調度が 1 であることに着目することで，音環境と人の調和が取れた処理を実現できる．音声信号処理の要素技術である音声区間検出（VAD）は，パワーエンベロープ処理体系において重要な技術である．そこで，雑音残響に頑健な VAD を提案する．頑健な VAD 法は，回復パワーエンベロープと最適化したパワー閾値を用いることで，変調度 1 のパワーエンベロープに対して音声/非音声判別を行った時と同等の検出性能が得られるようにする．パワーエンベロープ処理体系を実現するために，統合的音声信号処理を提案する．統合的音声信号処理は，頑健な VAD，パワーエンベロープ回復処理（パワーエンベロープ減算処理，MTF 逆フィルタ処理），SNR 推定，残響時間推定で構成される．これらの各処理は，変調度を 1 にするというコンセプトに基づく処理で実現した．統合的音声信号処理の評価として，VAD，帯域分割型パワーエンベロープ回復処理，SNR 推定，残響時間推定をそれぞれ評価する．評価結果より，各処理が概ね精度よく回復・推定できることが明らかとなった．統合的音声信号処理の応用として，帯域分割型パワーエンベロープ回復処理を前処理とした音声認識システムと STI 推定を示し，認識性能の向上と STI を推定できることを示した．総合的な結果，統合的音声信号処理により音環境バリアフリーのためのパワーエンベロープ処理体系を実現できた．

目 次

1 序論 1

1.1 はじめに 1

1.2 音環境バリア 3

1.3 音環境バリアフリー 6

1.3.1 これまでの音環境バリアフリーの試み 6

1.3.2 本研究で扱う音環境バリアフリーの範囲 7

1.4 音声環境バリアフリーのための音声信号処理技術 8

1.4.1 音環境に関する情報の測定・推定 8

1.4.2 雑音・残響除去 11

1.4.3 音声強調・音声回復 14

1.4.4 音声認識，音声区間検出 15

1.5 問題意識と問題設定 17

1.6 本研究の目的 19

1.7 本研究の構成 20

2 変調伝達関数に基づいたパワーエンベロープ処理体系 23

2.1 音声区間の検出と閾値最適化 24

2.1.1 信号検出理論 24

2.1.2 パワー閾値の最適化 26

2.1.3 音声区間検出における頑健な処理 27

2.2 変調伝達関数 29

2.2.1 変調伝達関数の概要 29

2.2.2 MTF と音環境 31

2.3 音環境バリアフリーと逆フィルタ処理 35

2.3.1 音環境と変調スペクトル 35

2.3.2 雑音残響環境での最適化問題 38
2.3.3 MTF の逆フィルタ処理による最適化 43

3 パワーエンベロープ処理体系に基づく統合的音声信号処理 45
3.1 統合的音声信号処理の概要 45
3.2 音声区間検出 46
3.2.1 雑音残響に頑健な音声区間検出法の概要 46
3.2.2 回復パワーエンベロープ 47
3.2.3 閾値最適化 48
3.3 パワーエンベロープ回復処理 49
3.3.1 パワーエンベロープ抽出 52
3.3.2 パワーエンベロープ減算処理 52
3.3.3 パワーエンベロープ逆フィルタ処理 53
3.3.4 帯域分割型パワーエンベロープ回復処理 53
3.4 パラメータ推定 54
3.4.1 SNR 推定 54
3.4.2 残響時間推定 59
3.5 性能評価 60
3.5.1 音声区間検出 60
3.5.2 パワーエンベロープの回復 63
3.5.3 パラメータ推定 65

4 応用 76
4.1 音声認識のフロントエンド 76
4.2 STI 推定 80

5 結論 88
5.1 本論文で明らかにしたこと 88
5.2 今後の展望 90

謝辞 93

本研究に関する研究業績 **113**

その他の研究業績 **116**

第 1章

序論

1.1　はじめに

音声コミュニケーションは，人にとって欠くことのできない情報伝達方法であることは周知の事実である．ユビキタス音声コミュニケーションは,「いつでも」,「どこでも」,「誰とでも」,「安心・安全」な音声会話の実現を目指したものである．近年，人と機械の音声コミュニケーションも音声認識（ASR: Automatic Speech Recognition）技術の発展とともに欠くことのできない技術になっている．そのため，人と人の音声コミュニケーションだけでなく人と機械の実現を目指して，様々なアプローチによる取り組みがある．

まず，ユビキタス音声コミュニケーションにおける「いつでも」に着目すると，即時性や可動性などが考えられる．即時性や可動性に関しては，短波を利用した無線通信が電話よりも優位に立った時代が長かったが，衛星通信による衛星電話の出現により地上・海上からいつでも世界中の誰とでも音声会話が可能となった．さらに，ICT（Information and Communication Technology）技術が発展し，携帯電話・スマートフォンの普及に伴い世界中の多くの地域において，いつでも電話をかけることができるようになった．また，4G 高速携帯電話網や光回線インターネット網の普及に伴い，ほとんど遅延なく世界中の相手との動画と高音質音声による相互通信コミュニケーション通話が可能となっている．

次に「誰とでも」に着目すると，人と人，人と機械（人から機械，機械から人），機械と機械といった 3 パターンが挙げられる．人と人との音声コミュニケーション

において，健常な聴力や発話能力を有する人による音声会話は，個人の持つ発話の明瞭性や聴取能力に若干の差異はあるが，高騒音環境を除いて音声会話によるコミュニケーションが成立する．しかし，聴覚や発話に障害を持つ人と健常な聴力や発話能力を有する人との間には，阻害要因（バリア）が存在する．人の聴力に関する問題として，聴覚に障害を持ち健常な視覚能力を有する人は，聴覚からの情報の取得が困難なため，手話の読み取りや口元からの読み取り，書き取りのようなコミュニケーションが必要となる．最近では，ASR やタブレット端末の普及により音声からテキストに変換する補助装置なども社会で利用されており，介護での利用などを考えた関連研究も行われている．また，加齢に伴う聴力低下によっても円滑な音声コミュニケーションが難しくなるために，補聴器が広く普及しており，音声強調を行うことで明瞭性改善の一助を担っている [1]．しかし，バッテリーに起因する稼働時間の問題や音環境の変化により雑音まで強調されて不快感があるなど，利用者の要求を満たすことができておらず様々な研究が継続されている．さらに，重度の難聴者に対しては，人工内耳を埋め込む方法 [1] や超音波補聴器 [2] の利用により音を取り戻す研究も進められている．人の発話に関する問題としては，発話に障害を持った人が音声を発話することは難しく，発話できた音声も健常な聴力を有する人にとっては聴き取りにくい．疾患によって声帯を摘出した人は，電気式人工咽頭（補声器）や手術によって埋め込む気管食道シャント法により音声発話が可能となる．しかし，これらの音声は子音の聴き取りが難しいなどの様々な問題が残されており，円滑な音声コミュニケーションのために様々な研究が進められている．また，人と人の音声コミュニケーションでは，主の情報である言語情報以外にも，非言語情報も重要な要素である．人は感情を声の大きさや高さ，テンポなどの非言語情報によって表現しており，話者の感情を理解することで，円滑にコミュニケーションを行っている．

人と機械の音声コミュニケーションにおいては，機械の理解として ASR，機械の発話として音声合成技術が広く利用されている．ASR は，音声情報の文字化や機械の操作に利用されている．機械の理解に着目すると収音系やコンピュータ性能，学習や辞書，言語といった問題がある。例えば，現状の技術では，英語を 100 % 人間と同じ能力で理解して文字に変換することはできていない上に，全世界の言葉に自動翻訳することもできていない．そのため，ASR や自動翻訳などの様々

な研究が今なお数多く行われている．一方，機械の発話に関しては，文字情報からの音声合成技術が公共の場での案内として多く利用されている．例えば，駅構内などでは事前に収音した単語を組み合わせて連続音声生成する波形接続型音声合成が利用されているが，人が発話した音声と合成音声を比較すると自然性が低く，イントネーションを中心に不自然さが残るという課題がある．

最後に,「どこでも」に着目すると，どんな場所でも快適に音声コミュニケーションを行うことが求められる．前述の通り，電話やスマートフォン，タブレット端末の普及に伴い，駅や空港，工場，学校，ショッピングセンターなどで音声会話やASRを利用する機会が増えている．これらの音環境の多くは，静音環境ではなく音環境の悪い音環境である．このような環境では，雑音や残響といった音環境の阻害要因（音環境バリア）によって音声コミュニケーションが妨げられるという問題がある．この音環境バリアの問題は，ユビキタス音声コミュニケーションの実現に向けた，音バリアの外的要因における最も重要な問題である．音環境バリアの問題解決の重要性は，騒音問題や雑音残響除去，音声強調，音源分離，頑健な音声信号処理などの音環境バリアフリーに向けた研究分野で非常に多くの研究者や企業が取り組んでいることからも明らかである．また，音環境バリアの問題は，音信号処理全般への波及効果としても大きく貢献できる問題であり，自動車業界をはじめとした産業界からも問題の解決が熱望されている．そのため，本研究においても「どこでも」の音環境バリアの問題に着目して研究を進める．次節以降，音環境の問題に取り組む背景とそのアプローチについて述べる．

1.2　音環境バリア

バリアとは，阻害要因であり，円滑に物事を進めるのを妨げる障壁を意味するが，その捉え方は様々である．ここで，雨天のキャッチボールを例にバリアの捉え方について考える．二人組が，天気が良い日にキャッチボールをしたときには，落球というミスをしないとする．しかし，雨天にこの二人がキャッチボールしたときには，ミスを連発するとする．この時，雨が降っていなければミスをしないのだから，この時の一次的なバリアは，天候である雨である．目に雨が入って見えにくい，指が雨で濡れてボールが滑りやすいなどの影響は，雨が降っていなければ

起こらない二次的な問題であることから一次的なバリアではなく二次的なバリアである．ここで，音声コミュニケーション（音声のキャッチボール）で同様に考える．健常な二人が静かな会議室で会話をしていると，円滑に音声コミュニケーションができる．しかし，駅構内などの雑音と残響の影響を受ける環境では，子音がマスクされるなどして円滑な音声コミュニケーションができなくなる．この時の音環境の一次的なバリアは，雑音や残響であり，雑音や残響がなければバリアは存在せず，円滑に音声コミュニケーションができる．そして，雑音や残響の影響による同時マスキングや継時マスキングは二次的なバリアとなる．本研究における一次的な音環境バリアは，雑音と残響である．

まず，一次的なバリアである雑音について述べる．鉄道路線や空港，工場の周辺では，雑音の影響が騒音問題としてよく取り上げられている．騒音は，人的なストレスになるだけでなく，睡眠の妨げや聴力損失を引き起こす原因となる．音声信号に対する雑音の影響には，加法性雑音と乗法性雑音がある．加法性雑音は，信号に雑音が加法的に影響する，室における雑音である．乗法性雑音は，信号に雑音が乗法的に影響する，通信路における雑音である．音場に自由音場を仮定すると雑音の影響は加法性であることから，音環境における雑音は加法性雑音である．加法性雑音では，目的信号に雑音が加法されるため，二次的なバリアである同時マスキングが起こり，明瞭性や了解性，認識性能の低下を引き起こす．

人と人との音声コミュニケーションにおける雑音の影響を考えると，子音などのパワーの小さい音声は，音声が雑音にマスクされやすいという同時マスキングの影響が大きい．そのため，音声が聴き取りにくくなり，明瞭性が低下してしまうだけでなく，異聴を起こすことで了解性の低下にもつながる．植松らによる雑音環境での明瞭度・了解度の実験においても，信号対雑音比（SNR: Signal to Noise Ratio）が低下するにつれて明瞭度・了解度が低下する結果が示されており [3, 4]，Morimoto et al の実験でも SNR が 15 dB 以下の条件て明瞭度が低下し，聴き取りにくくなる結果が示されている [5]．高騒音環境下では，母音・子音に関係なく発話内容の理解が難しくなる．人と機械の音声コミュニケーションにおいても同様で，音声特有の特徴が雑音に埋もれることにより，機械での認識性能が低下する [6]．雑音の種類に関しては，機械音のように定常的な雑音もあれば，大勢の人が話すことで生じるバブル雑音のように非定常な雑音も多く存在する．また，ドア

の開閉音などの突発性雑音は，非定常雑音であり，雑音が生じる時刻が未知であるとともに瞬時的なパワーが非常に大きいため，音声の明瞭性や認識性能に大きな影響を与える雑音としてよく知られている．雑音の影響は，雑音の種類や SNR によって明瞭性や認識性能に大きな差が生じることが知られている．

次に，残響について述べる．残響は，室での壁や床などからの複数の反射によっておこる現象である．残響は，室で音を発した後に連続的に徐々に減衰して響きが残る．残響の特性は，インパルス音やチャープ信号の一種である TSP (Time Stretched Pulse) 信号を用いて測定される室内インパルス応答（RIR: Room Impulse Response）によって知ることができる．残響音声は，音源から呈示された原音声が室による複数かつ複雑な反射を繰り返すことで生じることから，原音声に RIR が重畳される形で表現され，継時マスキングが起こる．そのため，音声の特徴が歪み，音声信号の尾（音声の終点）が原音声よりも長くなる．残響音声の聴取という点では，重畳成分が影響を及ぼすことで聴き取りにくくなり，結果的に残響音声の後部（音声の開始時点から時間が経過した時点）ほど聴き取りにくくなる傾向がある．残響は，初期反射（Early Reflection）と後期残響（Late Reverberation）に分けて表現される．初期反射は，反射の回数が少なく複雑ではないために音声伝達に対する影響は小さい．一方，後期反射は，直接音から 150 ms 以上経って聴こえる反射音であり，反射による音声信号への歪みが大きい．そのため，後期反射は音声伝達に与える影響が大きい．人と人の音声コミュニケーションでは，残響の影響により，明瞭度の低下 [7, 8] や了解度の低下 [9]，聴き取りにくさの上昇 [10, 11] が確認されている．人と機械の音声コミュニケーションでは，音声認識率の低下 [12] が確認されている．ここまで，音声コミュニケーションにおける残響はバリアであると述べたが．コンサートホールなどの音楽を楽しむ音環境における残響は，バリアではなく響きとして必要不可欠なものであることを補足しておく．

最後に，実環境では，雑音と残響が同時に存在している．雑音残響音声は，雑音と残響の両方の影響を受けるため，目的音声に対する雑音による同時マスキングと残響による継時マスキングが起こる．雑音残響環境での認識性能や明瞭度の低下，聴き取りにくさの上昇は，雑音のみ，残響のみの音環境に比べて大きくなる．そのため，雑音残響環境において音環境バリアを取り除く，音環境バリアフリーという考え方が重要となる．

1.3　音環境バリアフリー

先の雨でのキャッチボールの問題を再び例に出してバリアフリーを考える．雨天時にキャップ帽を被ることで二次的なバリアである雨が目に入る問題を軽減させることは，バリアフリーの一つの方法である．しかし，この方法では，一次的なバリアである雨の影響を完全には排除できない．このように，二次的なバリアを取り除いただけではバリアを完全に取り除けず，バリアフリーとしては不完全である．そのため，一次的なバリアを取り除く必要がある．この問題での究極のバリアフリーは，降っている雨を降らないようにすることであり，これが一次的なバリアを取り除くことである．実際には自然界の問題を解決することは難しいため，最大限出来るバリアフリーの方法としては，キャッチボールを行う場所の上に屋根を設けることである．その結果，雨の影響を全く受けなくなる．同様に雑音や残響の影響を受ける状況での音声コミュニケーションについて考える．話者が意図的に声を張り上げることで結果的にSNRが改善することは音環境バリアフリーの一つの方法であるが，これは二次的なバリアに対するバリアフリーであって，雑音がさらに大きくなった場合には対応できない．一次的なバリアである雑音と残響を空間上で取り除くことが究極のバリアフリーであるが，現実的に不可能である．そこで，最大限実現可能なバリアフリーの方法としては，人の耳やマイクロフォンの入力部を完全に覆い，雑音と残響を相殺することで目的の音声のみを呈示する装置を実現させることである．

1.3.1　これまでの音環境バリアフリーの試み

音環境バリアフリーについて研究テーマとして取り上げられるようになったのは，2006年ごろからである．音環境バリアフリーに関してまとめられたものが．上羽らによる音バリアフリーの解説論文「音バリアフリーの現状と課題」である[13]．ここでは，バリアフリーの利用対象者は，障害者や高齢者に限定されてはおらず，健常者に，さらに機械を加えたものである．従来の音バリアフリーでは，人のみを対象としていたが，近年の音バリアフリーでは，ASR技術の向上や会話型ロボットの発展と共に，利用者としての機械も対象となっている．音バリアフリーでは，聞こえに関するもの（聴力に関するバリアフリー），音声発話に関するもの

（発話のバリアフリー），その他「音」で保証できるもの（音でバリアフリー）という三つのカテゴリーに大別されている．ここでの，聞こえに関するものとしては，主に聴覚障害を持った人を想定して分類されている．また，音声発話に関するものとしては，障害や個人性に関するものが想定されている．そのため，音環境バリアフリーという問題は，単純にこれらのカテゴリーに分別することは難しく，音環境バリアフリーの問題は音バリアフリーのカテゴリーにおいて横断的な問題だと考えられる．

例えば，駅構内などの残響環境において構内放送の明瞭性を確保するために，前処理した音声を呈示することで公共の場での音環境バリアフリーを実現する研究がある [14, 15]．この研究は，呈示前に音声を加工することから，音声発話に関するものに該当すると考えられる．一方，多くの場合，音環境バリアに対して収音系や情報加工といった雑音・残響除去や音源分離などの信号処理技術を用いて音環境バリアフリーの実現を目指している．このような音環境バリアフリーは，収音した音を加工するため，聞こえに関するものである．ここまでは，電気音響・音声における音環境バリアフリーについて述べたが，建築音響においても音環境バリアフリーの研究が様々取り組まれている．例えば，案内放送は，必要な人にとっては重要な情報だが，情報を必要としない人にとっては騒音でしかないため，案内放送を必要最小限にすることで音環境バリアフリーに貢献できる可能性も示されている [16]．

1.3.2　本研究で扱う音環境バリアフリーの範囲

本研究では，ユビキタス音声コミュニケーションの「どこでも」に着目していることから，音環境バリアは雑音残響として，雑音残響環境を音環境として扱う．扱う音環境バリアフリーは，人が発話した音声を人や機械が聴くことを前提に，発話者と聴取者の間の経路に存在する音環境バリアを取り除くことを行う．収音した雑音残響の影響を受けた観測信号に対して雑音・残響除去といった音環境バリアを相殺する信号処理によって音環境バリアフリーを実現する．

1.4　音声環境バリアフリーのための音声信号処理技術

1.4.1　音環境に関する情報の測定・推定

音環境バリアフリーのための音声信号処理技術には，雑音や残響といった音環境に関する情報が必要不可欠である．雑音除去や残響除去では，雑音や残響に関する尺度や物理指標をパラメータとして利用している．これらのパラメータは，元々音環境を客観的に評価することを目的に提案され，室の特性を測定して計算により求められる．最近は，これらのパラメータをブラインドで簡易的に推定することで，室の簡易評価に役立てようとする研究もある．そのため，雑音・残響に関する物理指標や評価尺度をパラメータとして精度よく推定することが求められている．本研究の音環境バリアフリーにおいても，音環境のバリアである雑音や残響の影響がわからなければバリアフリー実現が困難であるため，はじめに音環境に関するバラメータの尺度や指標とその測定・推定法について述べ，本研究を実現していく上での問題点について指摘する．

まず，雑音に関する尺度について述べる．雑音と目的音声の関係を表す尺度として最も広く知られているものに SNR がある．SNR は，信号と雑音のパワーの比を取るものであり，主に global SNR（gSNR）と local SNR がよく用いられている．前者は観測信号全体の SNR であり雑音の影響を示すのに一般的によく利用され，後者は観測信号を帯域分割した時のある帯域での SNR であり，帯域によってパワーが大きく異なる音声信号処理，特に雑音除去でよく用いられている．gSNR の算出式を次式に示す．

$$\mathrm{gSNR} = 10 \log_{10} \left(\frac{P_S}{P_N} \right), \tag{1.1}$$

ここで，P_S は音声のパワー，P_N は雑音のパワーであり，単位は dB である．gSNR は，雑音の影響を直感的に知るのに有効な尺度であり，観測信号全体の雑音レベルを示すのによく利用される．雑音が定常であれば，定常雑音のパワーを事前測定しておくことで，gSNR の理論値は求まる．しかし，通常，観測音声では目的音声に雑音が付加されているため，目的音声のみと雑音のみのパワーが未知である．そのため，観測音声のみから SNR を推定するには，統計的モデルなどを用いて雑音のパワーを推定し，観測音声のパワーから雑音のパワーを減算して目的音

声のパワーを求めてSNRを算出する．gSNRを推定する方法として，音声区間検出（VAD: Voice Activity Detection）を利用したgSNR推定法 [17] や瞬時振幅に基づくgSNR推定法 [18]，統計的分布と帯域分割処理によるgSNR推定法 [19], 変調スペクトルを用いたニューラルネットワークによるgSNR推定法 [20]，ガンマ分布に基づくgSNR推定法 [21]，計算論的聴覚情景分析（CASA: Computational Auditory Scene Analysis）に基づくgSNR推定法 [22] など様々な方法が提案されている．

local SNRの算出式を次式に示す．

$$\text{local SNR}_k = 10\log_{10}\left(\frac{P_{Sk}}{P_{Nk}}\right), \tag{1.2}$$

ここで，P_{Sk}はk番目の帯域での音声のパワー，P_{Nk}はk番目の帯域での雑音のパワーである．雑音除去などの音声信号処理においては，帯域分割処理がよく利用されており，ウィナーフィルタリング [23] やMMSE (Minimum Mean Square Error)-STSA (Short-Time Spectral Amplitude) [24] などの雑音除去などにおいては一種のlocal SNRがよく利用されている．短時間フレーム内のlocal SNRとして*a priori* SNRがある．この推定法として，DD (decision-directed) *a priori* SNR推定法 [24]，VADを用いた *a priori* SNR推定法 [25]，ニューラルネットワークを用いたデータ駆動型の *a priori* SNR推定法 [26]，多重線形回帰に基づく *a priori* SNR推定法 [27] など様々な推定法が提案されている．雑音除去や音声伝達指標（STI: Speech Transmission Index）[28, 29] 推定などの音声信号処理においては，雑音に関する尺度としてgSNRやlocal SNRを推定や予測して，パラメータとして利用している．本研究においても他の研究と同様に，雑音の影響を知る尺度としてSNRを用いる．SNRは前述の通り，目的音声に対してどの程度雑音の影響があるかを知るのに効果的な尺度であり，雑音の影響が完全に取り除かれた音環境では，SNR $= \infty$ dBになる．

次に，残響に関する尺度・物理指標について述べる．残響の特性として，多くの室のRIR特性は指数関数的に減衰することが知られている．残響に関する尺度の一つに，残響時間（T_{60}）がある．残響時間は，音源を停止してからパワーが60 dB減衰するまでに要する時間のことである．残響時間を含む残響に関する尺度は主に建築音響において利用される．残響時間の精密測定は，インパルス音を用いた測定である．インパルス音を用いた測定では，暗騒音に対して60 dB以上大き

い音を出す必要があり，容易な測定が出来ない．そこで，様々な残響時間推定法が提案されており，最尤法による残響時間推定法 [30] や混合ガウス分布（GMM: Gaussian Mixture Model）を利用した残響時間推定法 [31, 32]，スペクトル減衰分布を利用した残響時間推定法 [33]，ニューラルネットワークを利用した残響時間推定法 [34, 35, 36]，変調伝達関数（MTF: Modulation Transfer Function）に基づく残響時間推定法 [37, 38, 39] などがある．これらの残響時間推定法では，雑音の影響の無い残響環境において高い精度で残響時間を推定できるものの，残響時間が長くなるにつれてその推定誤差が大きくなる．他の残響に関する尺度として，直接音と残響音のパワー比をとる DRR（Direct to Reverberant Ratio）[40] がある．DRR は，残響の影響に関する尺度として広く利用されており，DRR 推定法には，空間的自己相関モデルを用いた DRR 推定法が提案されている [41]．近年，国際会議 WASPAA のワークショップとして ACE challenge が開催され，残響時間および DRR の推定アルゴリズムについて比較評価がなされた [42, 43]．この中で，雑音環境下における残響時間と DRR の推定性能は，SNR が高い環境では残響時間の推定誤差は小さいが，DRR の推定誤差は大きいという結果が示されている．SNR が低い環境下における残響時間や DRR の推定が難しいことも報告されている．残響時間や DRR 以外にも音の響きに関して，RIR の全体のパワーと 50 ms までのパワーから求める音の明瞭性に関する物理指標である D 値（Deutlichkeit）があり，次式で表現される．

$$D = \frac{\int_0^{50\mathrm{ms}} h^2(t)dt}{\int_0^{\infty} h^2(t)dt}. \tag{1.3}$$

ここで，$h^2(t)$ は RIR のパワーである．他には，室内音響の物理指標に，80 ms までのパワーと 80 ms 以降のパワーとの比から求める音の透明性に関する物理指標 C 値（C_{80}）があり，ホール設計などで利用されている [44]．建築音響分野において物理指標は，正確かつ適切に求めるために室内インパルス応答を測定して求められている．D 値や C 値といった物理指標も推定できれば便利であるが，これらの物理指標の推定法は報告されていない．これは，物理指標の導出過程を考えても，まず音声に関する評価に向いておらず応用分野が狭く，D 値や C 値の推定法の必要性が低いためと考えられる．残響時間や DRR は，STI 計算や残響除去などの音声信号処理においてパラメータとして利用されることから，残響時間や DRR

の推定は重要性が高い．本研究においても残響の影響を知るのに残響時間を利用する．

最後に，雑音と残響を同時に扱う物理指標について述べる．建築音響の現場において室の音声伝達性能を評価するのに用いられるSTIがある．STIは，MTFの概念 [45, 46, 47] そのものであり，雑音・残響によって音声がどの程度の影響を受けるかを示す．そのため，雑音と残響の両方が同時に存在する音場を効率よく評価できる．STIは，音声明瞭度と相関関係があり，聴き取りにくさとは相関が非常に高いことが報告されている [5, 10]．そのため，STIを知ることで明瞭度や聴き取りにくさがどの程度変化するのかを容易に知ることができる．また，STIの簡易版にRASTI（Rapid STI）があり，RASTIはSTIやSII (Speech Intelligibility Index) に比べて音環境の影響を予測するのに有効であることが報告されている [48]．しかし，通常のSTI計算では，音場のRIR測定を必要としており，音環境の事前測定が必要である．近年，RIRの事前測定を必要としない，観測音声からSTIを直接推定する試みがある [49]．STIは雑音と残響を同時に扱える室の音声伝送性能を示す物理指標であるため，音環境バリアフリーに有効である．音環境バリアフリーを実現するにあたり，STIもしくはSTIを求める際の伝達関数であるMTFを用いることが効果的であると考えられる．

実環境に最も近い雑音残響環境において雑音や残響の影響を知るのに，雑音と残響を同時に扱うことができる評価尺度や物理指標を用いることが望ましいと考える．従来のように雑音と残響を別々に扱うと推定精度の低下につながることが危惧されるが，STIやMTFなどのように雑音と残響を同時に扱うことができ，同時に推定が可能であれば，推定性能の向上が期待される．そのため，雑音と残響を同時に扱える物理指標であるSTIならびにMTFの概念は，本研究において有効なアプローチであると考えられる．

1.4.2 雑音・残響除去

ここでは，音環境バリアフリーとして，音環境そのものを取り除くことを目的とした雑音・残響除去法について述べる．雑音・残響除去法の利用目的は，音声回復やASRでの性能の向上である．

まずは，音環境バリアとして雑音のみを取り除く雑音除去法について述べる．雑音除去の代表的な手法として，Boll によって提案された SS (Spectral Subtraction) 法 [50] がある．この手法は，雑音の振幅スペクトルの推定平均を求めて，観測信号の振幅スペクトルから雑音の振幅スペクトルを減算処理することで雑音を除去する方法であり，改良型として，サブトラクション係数に対して前処理を行う SS 法 [51, 52] などが提案されている．最小二乗平均 (LMS: Least Mean Square) アルゴリズムを用いて適応フィルタを設計し，このフィルタを用いて雑音音声信号から雑音を除去する ANC (Adaptive Noise Cancelling）[53] が Samburによって提案されており，改良型の ANC 法 [54, 55] もある．Kalman filter を適用した雑音除去法 [56, 57, 58] や最大尤度を用いたフィルタ設計による雑音除去法 [59] などがフィルタ処理による雑音除去としてある．一方で，音源分離を用いたアプローチとして，独立成分分析 (ICA: Independent Component Analysis) に基づくブラインド音源分離 (BSS: Blind Source Separation) が提案されており，周波数領域での ICA を用いたブラインド雑音除去 [60] やマイクロフォンアレーによる SS 法を利用した ICA [61] が提案されている．Hermansky & Morgan は，音声の変調スペクトルの重要な周波数成分（約 1〜12 Hz）のみを通過させる IIR フィルタによる帯域制限フィルタ処理を行う RASTA (RelAtive SpecTrAl processing) 法 [62] を提案している．これらの手法は，雑音のみの音環境に対する手法であり，残響環境に即していない．

次に，音環境バリアとして残響のみを取り除く残響除去法について述べる．残響除去においては，RIR の逆フィルタ処理によって残響を相殺するという考え方に基づく手法が主流であり，様々なアプローチが提案されている．Neely & Allen によって提案された最小位相逆フィルタ法は，室内音場が最小位相特性を有している時に，RIR の逆フィルタ処理により残響を除去できる [63]．しかし，実際の室内音場は，非最小位相特性である場合がほとんどである．また，事前の RIR 測定を必要としており，未知の環境には適応できない．Miyoshi & Kaneda によって提案された MINT (Multiple-input/output inverse theorem) 法 [64] は，複数マイクロフォンを用いた逆フィルタ処理であり，音源から受音点までの RIR を事前に測定しておき，観測信号に RIR の逆フィルタを畳み込むというものである．RIR の事前測定を必要としない改良法として Semi-blind MINT 法 [65] や Semi-blind MINT 法で

取りきれなった残響成分を SS 法で取り除く手法 [66, 67] も提案され MOS (Mean Opinion Score) 値として音質の改善が確認されている．Wang & Itakura は，マルチマイクロフォンと MMSE を用いた帯域分割逆フィルタ処理法 [68] を提案し，各帯域での回復信号を合成することで音声回復を実現している．木下らは，後期残響の回復を目的として，マルチチャネルマルチステップ線形予測を利用した残響除去法 [69] を提案している．線形予測は，逆フィルタを推定するのに効果的な手段の一つである．この手法では，複数の入力からマルチステップ線形予測に予測係数を後部残響の推定に用い，スペクトル減算により残響除去を行う手法である．その結果，残響音声の音声認識率を飛躍的に向上させている．先に述べた ICA は残響環境にも適用することができ，残響に頑健な ICA としてマルチマイクロフォンを利用した ICA 法 [70] が提案されている．異なったアプローチとして，Nakatani et al. によって音声の調波構造に着目した HERB (Hermonic-based dEReverBeration) [12, 71] が提案されており，単一マイクロフォンでのブラインド残響除去を実現している．改良された HERB [72] も提案されている．RASTA も残響音声に対しても適用することができ，残響音声に対しても有効であることが示されている [73]. しかし，これらの多くの手法は，残響環境のみでしか音環境バリアフリーを実現できない手法であるとともに，RIR の事前測定が必要で，音環境が変わると性能が低下したり，ある程度のマイクロフォン間隔が必要なマルチマイクロフォンを要するといった問題がある．

最後に，音環境バリアとして雑音と残響の両者を取り除く雑音残響除去法について述べる．雑音残響除去法として，SS 法と Wiener filter による手法 [74] が提案されており，雑音残響環境において単一マイクロフォンで音声認識率を向上できることが示されている．先に述べたマルチチャネルマルチステップ線形予測を利用した残響除去法 [69] は，もともと雑音残響音声に対して有効な手法として提案されており，ASR による評価の結果，この手法が雑音残響音声に対しても有効であることが示されている [75]．また，RASTA は，雑音・残響それぞれの環境においては有効であるが，音声に着目した音声強調手法であるため雑音残響環境においての性能は大きく期待できない．他にも，位相に着目した Kalman Filter と CMN (Cepstrum Mean Normalization) の組み合わせによる雑音残響除去法を提案されており，音声を回復したり，音声認識率を向上させたりできることを示している

[76]．しかし，RASTA を除くこれらの手法のほとんどは，雑音と残響において異なった特徴を用いており，残響雑音と残響それぞれの問題に対して適応的な対処を単純に組み合わせた処理となっている．さらに残響については，後部残響を雑音とみなして低減するアプローチであり，残響の影響を除去するアプローチになっていない．

1.4.3 音声強調・音声回復

音環境バリアフリーとして，雑音や残響の影響により聴き取りにくくなった音声や認識しにくくなった音声を，人や機械が聴き取り易い音声に加工して呈示する音声強調や音声回復といった音声信号処理も重要なアプローチである．ここでは，その様々なアプローチについて解説し，どのような処理結果が望まれるかについて述べる．SS 法は音声強調としても利用することができ，時間変化に伴う雑音の変化に対応し，雑音スペクトルの推定精度を向上させるために，マイクロフォンアレーを用いた SS 法 [77, 78] が提案されており，MOS 値の改善が確認されている．SS 法の弱点として，雑音に関する推定誤差によりミュージカルノイズが生じるという点があるが，様々なアプローチでミュージカルノイズ発生の低減が行われており，非線形重み付け SS 法で取り残した雑音を経験的モード分解（EMD: Empirical Mode Decomposition）を利用して取り除く手法 [79] などが提案されている．変調スペクトル上での SS 法 [80] なども提案されており，客観評価尺度である PESQ (Perceptual Evaluation of Speech Quality) の改善が確認されている．ミュージカルノイズを発生させない雑音除去法に Ephraim & Malah によって提案された MMSE-STSA [24] があるが，音質が低下するという問題がある．そこで，改良型として音質が向上する重み付けを行う MMSE-STSA 法 [81] が加藤らによって提案されている．最適フィルタの設計による音声回復法が提案されており，Wiener Filter による音声強調法 [23, 82] があるが，繰り返し回数が多いとスペクトル歪が生じるという問題が残る．先に述べた音源分離によるアプローチも音声強調として利用することができる．音楽からの音声の抽出 [83] や，雑音環境下での雑音からの純音の抽出 [84]，雑音からの複合音の抽出 [85, 86] などがなされており，目的音声とそれ以外の音環境バリアを含む雑音を音源分離できれば，音声強調として

利用できると考えられるが，音環境バリアフリーとしての実現は遠い．

残響音声に対しては，RASTA は音声強調にも利用することができ，残響音声に対して RASTA 法を用いたところ明瞭度の改善が確認されている [14]．2014 年に Reverb Challenge が開催され，様々な音声強調法や残響除去法についての主観・客観評価及び ASR の評価が同一条件下において行われた [87, 88, 89]．雑音・残響除去法に関して主観・客観評価が行われており，その結果，主観評価では，単一マイクロフォンによる手法もマルチマイクロフォンによる手法も残響成分は低減でき，残響の影響が改善されているものの，音質が劣化する結果が示されている．また，単一マイクロフォンによる手法では，音質が著しく悪いという結果が示されている．このように，雑音残響音声に対する単一マイクロフォンによるブラインド雑音残響除去による音声強調・音声回復は非常に難しい問題であることがわかる．

1.4.4 音声認識，音声区間検出

ASR や音声区間検出（VAD）といった音声信号処理技術にも，音環境バリアフリーが必要不可欠である．そのため，これらの音声信号処理技術においてどのように音環境バリアフリーが組み込まれて，どの程度の性能を有しているのかについて述べる．

まず，ASR について述べる．ASR は，多言語翻訳や多機能操作において社会で広く利用されつつあり，実用的な音声信号処理の利用例として欠くことのできない重要な技術となっている．ASR においても，目的音声が雑音や残響の影響を受け，観測信号から ASR を行うため，実環境においては認識性能が低下する．ASR におけるバリアフリーのアプローチとして，前述の雑音・残響除去法を前処理として用いるアプローチ以外に，音響モデルに雑音・残響音声を含んで学習させるアプローチや雑音・残響の推定量を音響モデルに入力するアプローチなどがある．近年，同一の雑音条件の下で様々な雑音除去法の認識性能が比較されるイベントが行われた [90]．技術の発展と共にビッグデータを扱うことが可能となり，学習データにクリーン音声だけでなく，様々な種類の雑音音声を含むことで，認識率が大きく飛躍しており，SNR $= 0$ dB の条件下でも，90% 前後の認識率が得られる手法もある．雑音音声に対してだけでなく，残響音声を学習して ASR を行う取り組

みにおいても音声認識率が向上することが確認されている [91, 92]．しかしながら，先にも述べた Reverb Challenge における認識結果として，単一マイクロフォンの処理では，マルチマイクロフォンの処理に比べて認識率が非常に低いことが確認されている [87, 88, 89]．

次に VAD について述べる．VAD は，観測信号から音声信号の区間である音声区間を検出する要素技術であり，音声符号化や ASR の前処理で利用されている [93, 94]．本研究においても非常に重要な役割を担う要素技術でもあり，様々な音声信号処理に必要不可欠な技術である．VAD は音声の特徴が含まれる区間を音声区間，音声の特徴が含まれない区間を非音声区間として，何らかの基準に基づき判別を行う．観測信号は，雑音や残響の影響を受けるため，目的音声の原音声の音声区間を検出することは非常に困難な問題である．そこで，様々なアプローチによる頑健な VAD 法が提案されており，大別すると，音声/非音声検出のための特徴に基づく方法と音声/非音声検出のモデルや決定法を利用する方法に分けられる．特徴に基づく方法では，信号パワーに基づく VAD 法 [95] や信号の周期性と調波性に基づく VAD 法（G.729B）[96]，パワースペクトルの周期/非周期性に基づく VAD 法 [97]，長時間スペクトルを利用する VAD 法 [98], 長時間信号の変動性（LTSV: Long-term Signal Variability）を利用する VAD 法 [99]，スペクトル傾斜の時間変化を利用した VAD 法 [100]，帯域分割 SNR を利用する適応多重レート（AMR）オプション 2 で使われている VAD 法 [101]，時間-周波数変調に基づく VAD 法 [102]，変調スペクトルに基づく VAD 法 [103]，EMD と変調スペクトル分析（MSA）を利用した VAD 法 [104]，EMD と瞬時周波数を利用した VAD 法 [105]，Wavelet に基づく VAD 法 [106, 107, 108] が提案されている．モデルまたは決定法を利用する方法では，Otsu 法の二値化 [109] を利用することで信号パワーの閾値を柔軟に決定する VAD 法 [95] が提案されている．その他には，GMM に基づく VAD 法 [110, 111, 112, 113, 114, 115] や隠れマルコフモデル（HMM）に基づく VAD 法 [116, 117, 118, 119]，SVM (Support Vector Machine）に基づく VAD 法 [120, 121, 122]，遺伝的アルゴリズム（Genetic Algorithm: GA) に基づく VAD 法 [123]，ガンマ分布に基づく VAD 法 [124]，統計的 VAD 法 [125, 126, 127, 128, 129]，信号の高次統計量に基づく VAD 法 [130]，自己回帰モデルに基づく VAD 法 [131]，線形予測 (LP: Linear Prediction) と自己回帰条件付き不均一分散モデルによる VAD

法 [132] が提案されている．これらの VAD 法は，静音環境もしくは雑音環境ではよく機能するが残響環境や雑音残響環境においては，残響の影響により検出性能が著しく低下してしまう．この性能低下の主要因は，残響の重畳性の影響により，音声信号の終点が後退することで非音声区間を音声区間とする誤検出が増加するためである．この問題に対して，複数のマイクロフォンを利用した，HMM に基づく VAD 法 [133] や統計的信号処理による VAD 法 [134]，雑音除去を用いた GMM に基づく VAD 法 [135] が提案されている．また，単一マイクロフォンによる入力を観測信号とした残響に頑健な VAD 法として MTF に基づくパワーエンベロープ回復処理を用いた VAD 法 [136] がある．しかし，これらの雑音・残響に頑健な VAD は，雑音のみ，残響のみの音環境には対応できているが，雑音残響環境には対応できていない．雑音残響に頑健な VAD の構想はあるものの，観測信号の入力に単一マイクロフォンを想定した雑音残響に頑健な VAD 法は提案されていなかった．

1.5 問題意識と問題設定

これまでの音環境バリアフリーの多くは，雑音のみ，残響のみを音環境バリアとして捉えたものが主であり，音環境バリアフリーとして不十分であった．一方で，音環境を雑音残響環境として捉えた音環境バリアフリーでは，雑音と残響それぞれの音環境バリアに対して雑音除去と残響除去の組み合わせにより実現しているものが大半を占めた．そのため，回復した音声は音環境によって異なり，音環境のパラメータの推定誤差の影響により，雑音除去や残響除去で過少・過剰回復となる．その結果，ミュージカルノイズが発生したり音質が低下するといった影響で人が聴いた時には聴き取りにくく不快であったり，機械では認識性能が低下するという問題に陥る．これは，音環境と人の間に雑音残響除去を行う機械があるときに，音環境の変化に追随できないという，音環境と人の調和が取れていない処理であるために起こる問題である．仮に，音環境と人の間にある機械が，時々刻々と変化する音環境の情報を正確に推定して，音環境のバリアを相殺できれば，音環境と人の調和が取れ，このような問題は起こらない．現状の音環境バリアフリーのアプローチでは，推定された音環境の情報を雑音・残響除去でそのまま使っているだけで，人の音声聴取に有効な特徴を用いた処理や雑音環境と残響環境を

一つの音環境として捉えた処理，静音環境での特徴を規範としたような処理とはなっておらず，音環境と人の調和が取れた処理となっていない．そのため，本研究が目指す音環境バリアフリーは，過少・過剰回復が起こらずに，音環境が変化しても音環境の変化に追随して，聴き取りやすい・音声認識率を高く保てるような，機械が音環境と人の間に入った時に音環境と人の調和が取れる処理である．そのためには，雑音と残響をそれぞれの問題として雑音残響除去するようなこれまでのアプローチではなく，雑音と残響を同じ特徴で扱って最適に音環境バリアを相殺できるような規範に基づいた処理が必要であると考える．

現状でも音環境バリアフリーに対して取り組もうとしている問題は大きいため，本研究で扱う音環境について仮定や条件を設ける．優先すべきことは，音環境と人の調和が取れた処理を実現することである．音場としては雑音残響環境とするが，雑音には定常な雑音を仮定する．残響に関しても拡散音場を仮定する．さらに，音環境バリアフリーの利用対象者を本研究では機械のみとするが，人への呈示を将来的に実現できるようなアプローチを取る．本研究において機械に限定して取り組む理由は，機械と人の音声コミュニケーションの重要性が，近年急激に高まっていることが一つにある．訪日外国人の問題をはじめとした近年の社会問題において，多国籍言語に対応した ASR 技術を用いた自動翻訳の実用化は急務な問題として認知されており国家のプロジェクトとして実施されている [137]．また，空港や駅などでの対話型ロボットによる案内の試験運用が始まり，人と機械の音声コミュニケーションの重要性が昨今非常に高まっている．人と人のコミュニケーションにおいて，人は類推を用いて会話を行うため認識性能が高いが，現状の機械の認識では類推を用いた認識は行われておらず，人の方が認識性能が高いと考えられる．そのため，人と機械の音声伝達性能が向上すれば，必然的に人と人の音声伝達性能も向上するものと考える．扱う信号には，音声の知覚に重要であるとされており，線形で扱いやすい時間信号の包絡線 [138] のパワーを取ったパワーエンベロープを扱うこととする．時間包絡線は，前述の雑音と残響を同時に扱うことができる物理指標である STI においても利用されていることから，音環境バリアフリーにおいて効果的な信号であると考える．音声信号のパワーエンベロープは，帯域分割処理を行うことで ASR へも利用できることから，パワーエンベロープは音環境と人の調和が取れた処理による音環境バリアフリーで用いるのに非常

に有効であると考える．音源は 1 つとし，単一マイクロフォンによる入力を観測信号とする．マルチマイクロフォンではなく単一マイクロフォンとするのは，単一マイクロフォンを仮定した信号処理の場合，過去に収音された歴史的価値を持つモノーラル信号などに対しても後処理を施すことで音環境バリアフリーを実現することに役立つと考えるためである．実環境における雑音残響音声は非常に複雑であることから，本研究では雑音は加法性，残響は重畳性として，定常性を有した背景雑音 $n(t)$ と拡散音場を仮定した RIR $h(t)$ を用いた雑音残響信号 $y(t)$ とした．雑音残響信号 $y(t)$ は，次式で表現される．

$$\mathbf{y}(t) \quad = \quad \mathbf{x}(t) * \mathbf{h}(t) + \mathbf{n}(t) \tag{1.4}$$

ここで，$x(t)$ は原信号，“$*$” は畳み込み記号である．本研究では，このような雑音残響音声を観測信号として扱う．

1.6 本研究の目的

問題意識において述べた立場からもわかる通り，ユビキタス音声コミュニケーションの「どこでも」に着目し，音環境バリアフリーの実現を試みる．従って，本研究の目的は，音環境と人の調和が取れた処理による音声コミュニケーションの実現を目指し，パワーエンベロープ処理体系により音環境バリアフリーを実現することである．パワーエンベロープ処理体系は，MTF の概念を核として用いる．MTF は，雑音と残響を同時に扱うことができる STI を求める概念そのものであるため，これまでのように雑音と残響を分けて処理するのではなく，変調度に着目して雑音と残響を一元処理できる．パワーエンベロープ処理体系は，統合的音声信号処理により実現する．統合的音声信号処理は，変調度に着目して原信号のパワーエンベロープを規範とした処理として実現する．統合的音声信号処理は，音声情報である音声区間と非音声区間を検出する雑音残響に頑健な VAD，頑健な VAD を実現するために必要な雑音残響除去としてのパワーエンベロープ回復処理，雑音残響除去に必要な音環境の推定によって実現する．

1.7 本研究の構成

本論文は 5 章で構成される．本論文の構成を図 1.1 に示す.

第 1 章では，本論文が対象とする研究分野の背景を述べ，研究の範囲を示した上で問題点を指摘して目的を示す．第 1 章においてユビキタス音声コミュニケーション，音環境バリアフリー，音環境バリアフリーのためのパワーエンベロープ処理体系といった本研究の位置づけを図 1.2 に示すように述べる．

第 2 章では，本論文において本研究の核となる音環境バリアフリーを実現するための本研究のアプローチであるパワーエンベロープ処理体系を提案する．そして，音環境バリアフリーのためのパワーエンベロープ処理体系が音環境と人の調和が取れた処理となっているかについての理論を述べる．はじめに VAD での音環境バリアフリーを取り上げて，パワー閾値最適化により，頑健な VAD を実現できることを示す．そして，この処理体系の核である MTF について詳しく述べ，MTF の逆フィルタ処理による最適化を行うことにより，本研究の理論の妥当性と問題点に対する有効性を示す.

第 3 章では音環境バリアフリーのためのパワーエンベロープ処理体系を実現するための方策である，統合的音声信号処理を提案する．統合的音声信号処理の構成技術である，雑音残響に頑健な VAD，パワーエンベロープ回復処理（パワーエンベロープ減算と MTF の逆フィルタ処理），SNR 推定，残響時間推定がどのようなコンセプトで，どのように実現され，どのように音環境と人の調和が取れているかを示す．そして，それらの性能を評価して結果を示すことで，統合的音声信号処理の有効性を明らかにする.

第 4 章では第 3 章で示した統合的音声信号処理の応用として，統合的音声信号処理を ASR の前処理と STI 推定に組み込み，その性能がどの程度かを明らかにすることで，音環境バリアフリーとしての効果を明らかにする.

最後に第 5 章では結論および今後の展望について述べる.

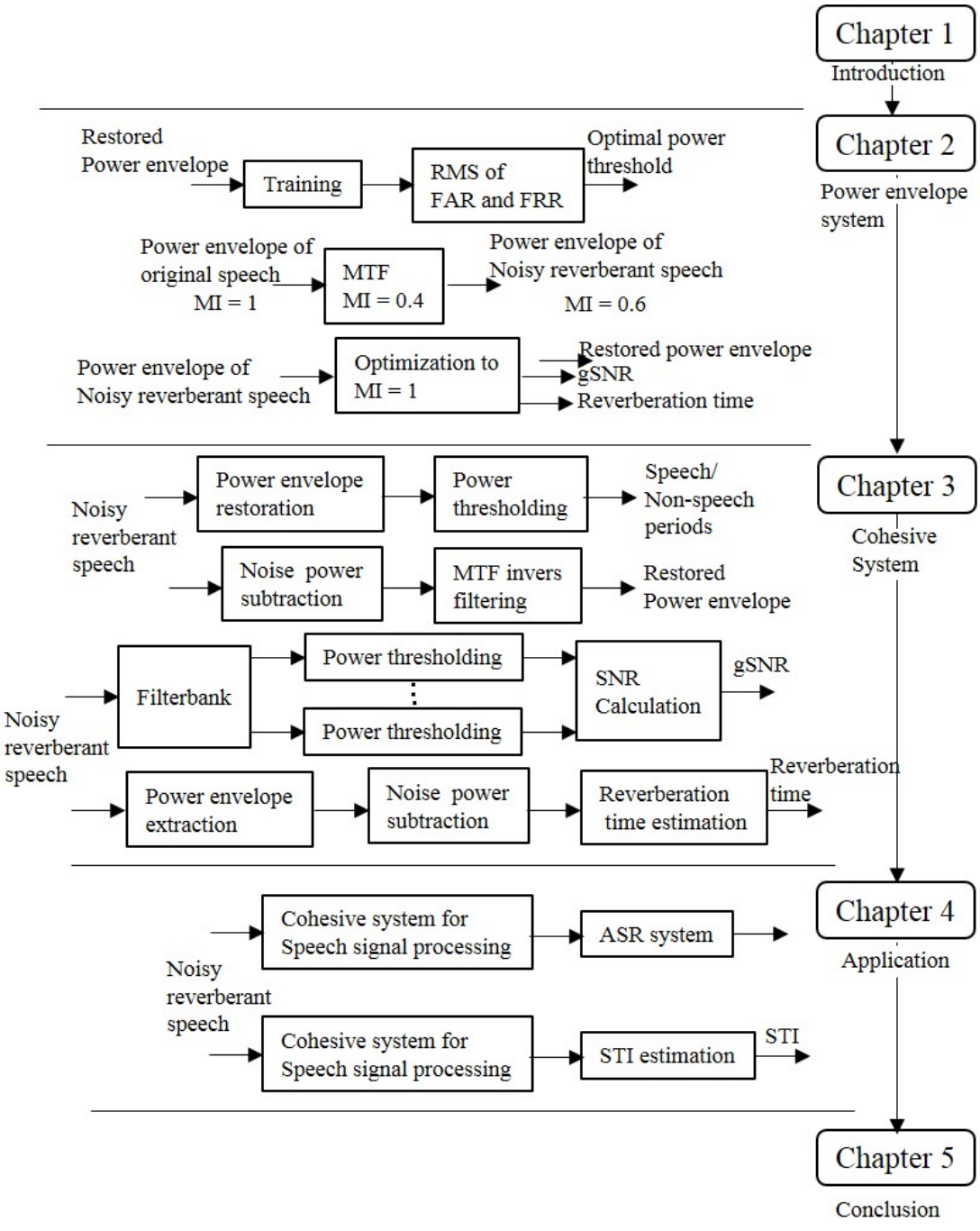

Chapter 1
Introduction
Chapter 2
Power envelope system
Restored Power envelope
Training
RMS of FAR and FRR
Optimal power threshold
Power envelope of original speech
MI = 1
MTF
MI = 0.4
Power envelope of Noisy reverberant speech
MI = 0.6
Power envelope of Noisy reverberant speech
Optimization to MI = 1
Restored power envelope
gSNR
Reverberation time
Chapter 3
Cohesive System
Noisy reverberant speech
Power envelope restoration
Power thresholding
Speech/ Non-speech periods
Noise power subtraction
MTF invers filtering
Restored Power envelope
Noisy reverberant speech
Filterbank
Power thresholding
Power thresholding
SNR Calculation
gSNR
Power envelope extraction
Noise power subtraction
Reverberation time estimation
Reverberation time
Chapter 4
Application
Noisy reverberant speech
Cohesive system for Speech signal processing
ASR system
Cohesive system for Speech signal processing
STI estimation
STI
Chapter 5
Conclusion

図 1.1: 論文の構成.

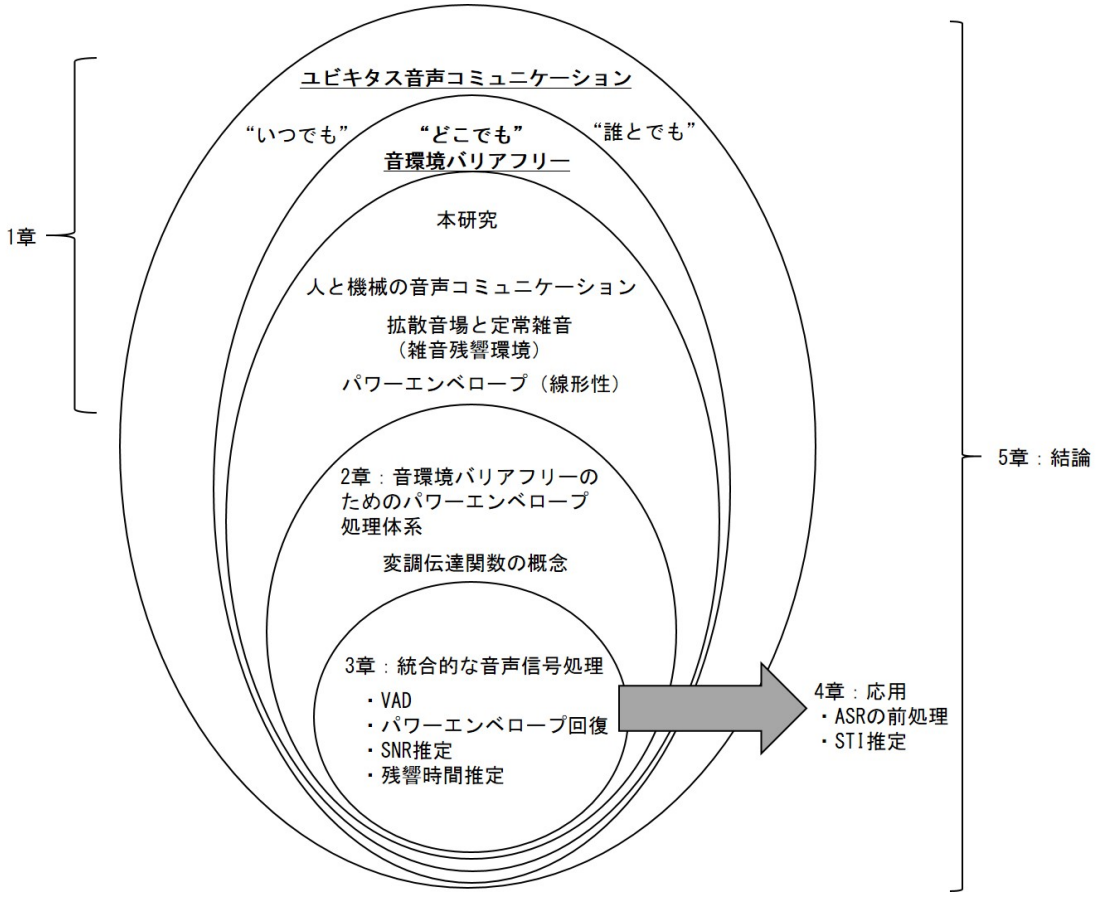


図 1.2: 本研究の位置づけ.

第 2章

変調伝達関数に基づいたパワーエンベロープ処理体系

本章では，変調伝達関数（MTF）に基づいたパワーエンベロープ処理体系について述べる．音環境バリアフリーにおいては，音環境バリアである雑音や残響に関する情報であるSNRや残響時間などを知ることが必要不可欠である．SNRは音声のパワーと雑音のパワーを用いて得られることから，音声区間と非音声区間の情報は非常に重要である．また，残響時間を例に取ると，残響時間は，音源が停止してから60 dBパワーが減衰するのに要する時間であることを考えれば，音声区間は非常に重要な情報である．そのために，音声区間と非音声区間を検出する音声区間検出（VAD）は，音環境バリアフリーにおいて非常に重要な音声信号処理技術と言える．しかし，VADへの入力信号は雑音や残響といった音環境バリアの影響を受けていることから，頑健なVADである必要があり，VADにも音環境バリアフリーが求められる．音環境バリアフリーにおいては，1.5節でも述べた通り，音環境と人の調和が取れた処理が必要とされる．そのためには，音声の明瞭度や聴き取りにくさとも関係があり，雑音と残響の両方の環境を同時に評価することができるSTIを参考にすることが有効であると考える．STIは，時間包絡線とMTFより求めることができる．そこで，時間包絡線のパワーを取ったパワーエンベロープを用いた処理体系による音環境バリアフリーを実現することを考えた．ここでは，VADにおける音環境バリアフリーの考え方について述べ，その後，この処理体系の重要な概念となるMTFについて解説する．そして，どのように音環

境と人の調和が取れるのかについて MTF の逆フィルタ処理を用いて述べる．

2.1 音声区間の検出と閾値最適化

VAD において音声区間と非音声区間を検出する際の最も重要な技術は，音声と非音声の判別処理（音声/非音声判別）である．音声/非音声判別には，スペクトルや時間信号のパワーを特徴として用いて判別を行う．音声/非音声判別に用いる判別閾値は，信号検出理論を用いて決定される．静音環境であれば判別閾値の決定が容易であるため音声/非音声判別も精度よくできるが，雑音・残響環境であればスペクトル上でも雑音・残響の影響で音声のスペクトルが埋もれる，もしくは歪むために音声や雑音・残響特有の特徴の抽出ができない．その結果，判別閾値が一意に定まらない判別閾値を使わざるを得ないために，音声/非音声の判別誤差が大きくなる．時間信号のパワーでも雑音のパワーに埋もれる，もしくは RIR が重畳される影響により音声区間の終点が遅れることで検出誤差が大きくなる．観測信号のパワーエンベロープに対して MTF に基づいた回復処理を行うことで，VAD に影響する音環境バリアを取り除き，最適化したパワー閾値を用いて音声/非音声判別を行う VAD を考えた．パワー閾値による VAD では，パワー閾値は音声/非音声判別において非常に重要であり，パワー閾値によって VAD の性能が大きく左右される．そこで，パワー閾値が一意に定まらない問題を解決すべく，パワー閾値の最適化というアプローチを考えた．

2.1.1 信号検出理論

信号検出理論は，画像処理を中心に発展を遂げ，音声信号処理でも利用されている．

信号検出理論においては，非信号区間を信号区間として誤判別した割合の誤受理率（FAR: False Acceptance Rate）と信号区間を非信号区間として誤判別した割合の誤棄却率（FRR: False Rejection Rate）を用いて評価を行う．このとき，信号検出のための判別閾値を変えた時の FAR と FRR を縦軸と横軸に描画した曲線である ROC (Receiver Operating Characteristic) 曲線を用いて，信号検出の性能

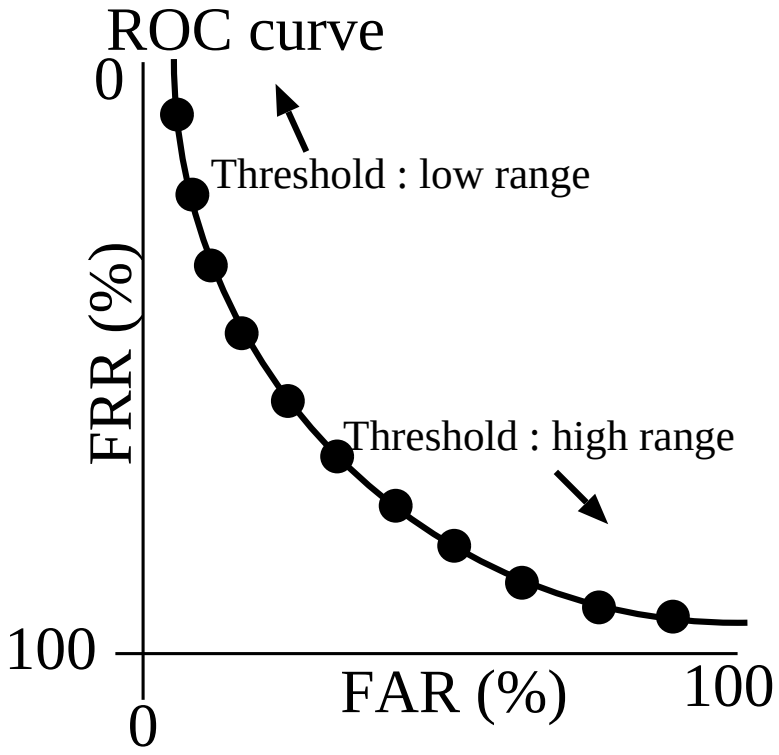


図 2.1: ROC 曲線のイメージ．

を明示することが一般的である．図 2.1 に ROC 曲線のイメージ図を示す．FAR と FRR はそれぞれ次式により求まる．

$$\mathrm{FAR} = \frac{N_{FA}}{N_{ns}} \times 100 \quad (\%) \tag{2.1}$$

$$\mathrm{FRR} = \frac{N_{FR}}{N_{s}} \times 100 \quad (\%) \tag{2.2}$$

ただし，N_s，N_{ns}，N_{FR}，N_{FA} は，それぞれ，音声サンプル点の総数，非音声サンプル点の総数，非音声と判断された音声サンプル点の数，音声と判断された非音声サンプル点の総数である．

ROC 曲線は，所望の VAD の性能を満たすための音声/非音声判別の閾値を決定するのに用いられる．VAD が最も利用される応用技術は，音声符号化と ASR であり，音声符号化と ASR においては，音声区間を 100 % 音声区間として判断する必要があるために，FRR は限りなく 0 % に近いところで閾値が設定される．しかしながら，本研究における VAD の優先的な利用目的は，SNR や残響時間といった音

環境バリアに関係する音環境の情報を知るのに利用する．そのため，従来の VAD のように FRR のみを重視して閾値を設定するのではなく，FAR と FRR の両方を重視して，少しでも全体としての検出誤差を軽減することが求められる．そこで次に述べるような，パワー閾値の最適化という方法によりパワー閾値を決定する．

2.1.2 パワー閾値の最適化

単純に雑音残響除去を行って信号検出理論で得た閾値を用いて音声/非音声判別を行っても，雑音残響除去による過少・過剰回復の影響により正確な音声区間を得ることは難しい．音声信号を変調で考えるとエンベロープとキャリアに分けられる．原信号のパワーエンベロープを入力信号，雑音残響信号のパワーエンベロープを出力信号とする．この時，原信号のパワーエンベロープが 100 % 変調であるとき，仮に雑音残響信号のパワーエンベロープは 50 % 変調とする．この時，雑音残響除去によってパワーエンベロープを回復しても，音環境の推定誤差などの影響によって，回復パワーエンベロープは 100 % 変調にはなっていない．そこで，本研究では，回復したパワーエンベロープに対して ROC 曲線を求め，正確に音声/非音声区間を判別するための最適な閾値を ROC 曲線から求める．回復したパワーエンベロープと最適なパワー閾値で音声/非音声判別を行うことは，少しでも原信号のパワーエンベロープが 100 % 変調に近い状態であると考えた．ここでの，パワー閾値決定の方法は，前述の通り，少しでも FAR と FRR の誤差を低減する方法を取る必要があるため，これより述べる最適化の方法によってパワー閾値を決定した．

図 2.2 にパワー閾値の最適化のイメージ図を示す．頑健な VAD は，回復パワーエンベロープに対するパワー閾値処理で構成した．ここで，パワー閾値は次の手順で最適化した．まず，回復パワーエンベロープに対してパワー閾値を変化させた時の VAD の性能を ROC 曲線に描画して求めた．次に，求めた ROC 曲線の FAR と FRR の二乗平均平方根（RMS）を次式のように取ることで音声区間と非音声区間の検出誤差を求めた．

$$\mathrm{RMS}=\sqrt{\frac{\mathrm{FAR}^2+\mathrm{FRR}^2}{2}}. \tag{2.3}$$

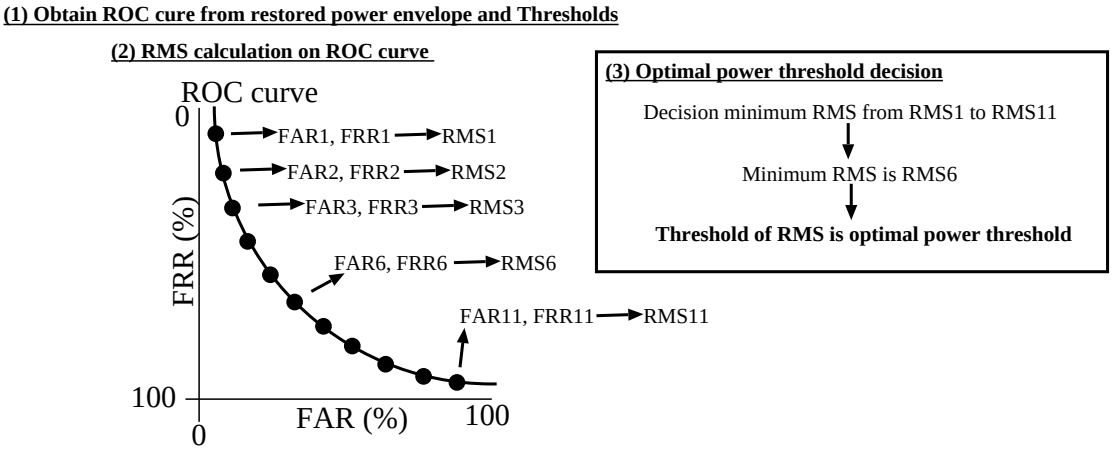


図 2.2: パワー閾値の最適化の流れ.

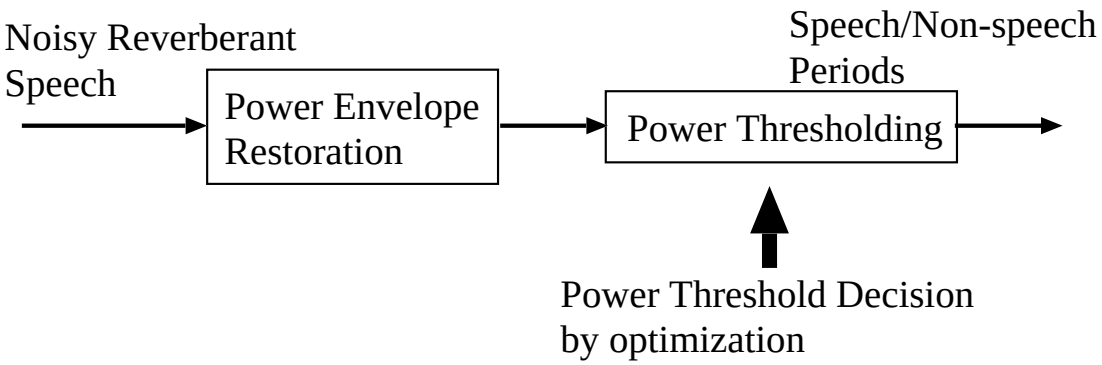


図 2.3: 頑健な音声区間検出の概要図.

そして，RMSが最小の時（誤差最小）のパワー閾値を，最適なパワー閾値として決定した．この時に，雑音や残響の条件を変えて最適化を行うことが，様々な音環境において精度よく音声区間と非音声区間を検出するために重要となる.

2.1.3 音声区間検出における頑健な処理

図 2.3 に頑健な VAD の概要図を示す．前述の通り，雑音残響に頑健な VAD は，回復パワーエンベロープと最適なパワー閾値により実現する．パワー閾値は 2.1.2 節の通りであるが，回復パワーエンベロープは，次節で述べる MTF の概念に基づき回復処理を行う．MTF の概念では，入力信号の変調度を 1（100 % 変調）と

して，雑音や残響の影響を受けることにより変調度が 1 未満になるという考えである．本研究におけるパワーエンベロープ処理体系においても限りなく変調度を 1 に近づけることを考えている．

変調度 1 というのは，原音声のパワーエンベロープと同等である完全に回復したパワーエンベロープのことを意味する．パワーエンベロープが変調度 1 の状態であれば，音声/非音声の判別も精度よくできるということになる．これまで述べてきた通り，本研究の雑音残響環境に頑健な VAD では，パワーエンベロープ回復と最適なパワー閾値により変調度を 1 に限りなく近づけるというアプローチを取っている．例として，原信号のパワーエンベロープの変調度は 1，雑音残響の MTF を 0.4 とすると雑音残響信号のパワーエンベロープの変調度は 0.6 となる．この時，パワーエンベロープ回復処理により変調度を 0.3 回復すると，回復したパワーエンベロープの変調度は 0.9 となる．このパワーエンベロープに対して原信号のパワーエンベロープに基づいて決定したパワー閾値で音声/非音声判別のパワー閾値処理を行っても正確な音声/非音声区間は検出できず検出誤差が生じることになる．そこで，回復パワーエンベロープと原信号のパワーエンベロープの差となる変調度 0.1 の誤差を，最適なパワー閾値を用いることによって音声/非音声判別の際に生じる誤差を埋めて，変調度 1 のパワーエンベロープに対するパワー閾値処理と同等の性能に近づける方法である．

頑健な VAD を含む，音環境バリアフリーのためのパワーエンベロープ処理体系においては，変調度を 1 に近づけるという概念が最も重要である．要は，変調度 1 というのは過少にも過剰にも回復しておらず，最適な状態（＝原信号のパワーエンベロープと同等）を意味しており，音環境と人の調和が取れた状態であるといえる．まずは，パワーエンベロープ処理体系において非常に重要な概念である MTF の概念を次節で詳しく解説する．

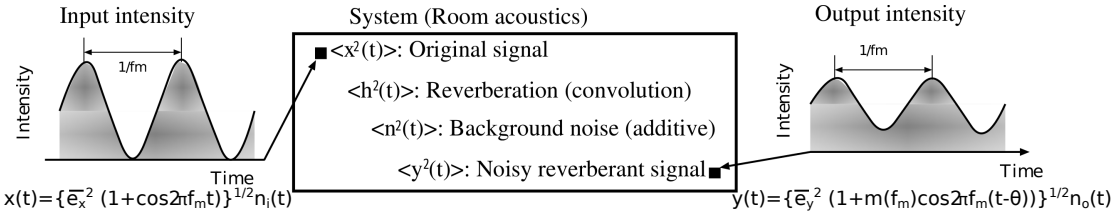


図 2.4: 雑音残響環境：入出力信号の信号強度とその伝達関数の関係.

2.2 変調伝達関数

2.2.1 変調伝達関数の概要

MTF の概念は，室内での音声伝送性能（音声の明瞭性）を予測し，客観的に評価するために，Houtgast & Steeneken によって音声明瞭度予測理論として確立された [45, 46, 47, 139]．彼らは，想定する室内音場を拡散音場と仮定し，変調度に関する音環境の問題を，残響による影響（重畳性）と背景雑音による影響（加法性）とした．彼らは，図 2.4 に示すように，室内において，話者の発話音声 $x(t)$ と聴取者の受聴音声 $y(t)$（入力信号 $x(t)$，出力信号 $y(t)$）の信号強度を，$\overline{I_i^2}$ と $\overline{I_o^2}$ と置き，次式で定義される信号のエンベロープの変化，つまり変調度の変化に着目した．

$$\text{Input:} \quad \overline{I_i^2}(1+\cos(2\pi f_m t)), \tag{2.4}$$

$$\text{Output:} \quad \overline{I_o^2}\{1+m(f_m)\cos(2\pi f_m(t-\theta))\}, \tag{2.5}$$

ただし，f_m は変調周波数，θ は位相，$m(f_m)$ は変調伝達関数である．ここで，入力信号をパワーエンベロープとみなすことで，入力信号は変調伝達関数 $m(f_m)=1$（100 % 振幅変調）の余弦波を有しているとする．出力パワーエンベロープでは，雑音や残響の影響を受けて変調伝達関数 $m(f_m)$ が 1 以下になり，$m(f_m)$ 倍に変化する．これが振幅変調で見た時の変調度に相当することから変調伝達関数（MTF）と呼ばれている．

MTF の概念を国際規格として確立させた客観評価尺度が STI である [28]．STI は，明瞭度と相関があることが Houtgast & Steeneken によって示されている [47]．一方で，戸井田らの報告によると STI と明瞭度の相関が高くないことも示されて

いる [140]．森本と佐藤の研究グループによって提案された聴き取りにくさという尺度に着目すると，STI と聴き取りにくさには高い相関があることが示されている [10]．また，Drullman et al. によると，エンベロープには聴こえに重要な情報を含んでいることが明らかにされている [138]．さらに，エンベロープと変調の関係に焦点を当てると，変調周波数の 2〜16 Hz が音声の明瞭性に重要であることが Arai et al. によって報告されている [141]．従って，STI と人の音声理解には関係があることから，STI を求める際に必要となる MTF もまた人の音声理解と関わりがあると言える．

次に，雑音・残響・雑音残響環境における MTF を示す．入力パワーエンベロープ $e_x^2(t)$ と出力パワーエンベロープ $e_y^2(t)$ は，それぞれ次式で表現される．

$$e_x^2(t) = \overline{e_x^2}(1+\cos(2\pi f t)), \tag{2.6}$$

$$e_y^2(t) = \overline{e_y^2}(1+m(f_m)\cos(2\pi f t)), \tag{2.7}$$

ここで，$\overline{e_x^2}$ と $\overline{e_y^2}$ は $e_x^2(t)$ と $e_y^2(t)$ の平均パワー，f_m は変調周波数，f は任意の変調周波数，$m(f_m)$ は変調伝達関数である．θ は MTF の定義に基づき便宜上 0 とした．入力信号，出力信号（雑音残響信号），統計的な RIR，雑音信号を，それぞれ，$\mathbf{x}(t)$，$\mathbf{y}(t)$，$\mathbf{h}(t)$，$\mathbf{n}(t)$ とする．統計的な RIR は Schroeder による指数減衰するモデル [142]，雑音には白色ガウス雑音のような定常な雑音を仮定している．これらは，次式のようにモデル化される．

$$\mathbf{x}(t) = e_x(t)\mathbf{c}_x(t), \tag{2.8}$$

$$\mathbf{h}(t) = e_h(t)\mathbf{c}_h(t) = a\exp(-6.9t/T_R)\mathbf{c}_h(t), \tag{2.9}$$

$$\mathbf{n}(t) = e_n(t)\mathbf{c}_n(t), \tag{2.10}$$

$$\mathbf{y}(t) = \mathbf{x}(t) * \mathbf{h}(t) + \mathbf{n}(t), \tag{2.11}$$

ここで，T_R は T_{60} で定義された残響時間，a は振幅項，“$*$” は畳み込み記号である．$e_x(t)$，$e_h(t)$，$e_n(t)$ は，それぞれ，$\mathbf{x}(t)$，$\mathbf{h}(t)$，$\mathbf{n}(t)$ のエンベロープである．$\mathbf{c}_x(t)$，$\mathbf{c}_h(t)$，$\mathbf{c}_n(t)$ は，$\mathbf{x}(t)$，$\mathbf{h}(t)$，$\mathbf{n}(t)$ のキャリアであり，それぞれ独立な白色ガウス雑音の特性を有するランダム変数である．このランダム変数は，次式の特性を有する．

$$\langle \mathbf{c}(t)\mathbf{c}(\tau)\rangle = \delta(t-\tau), \tag{2.12}$$

$\langle\cdot\rangle$ は集合平均の記号である [143]．また，δ はデルタ関数である．ここで, 集合平均の特性を利用することで，$\mathbf{x}(t)$ と $\mathbf{y}(t)$ の二乗集合平均は次式の通りである．

$$\langle \mathbf{x}^2(t) \rangle = e_x^2(t), \tag{2.13}$$

$$\langle \mathbf{y}^2(t) \rangle = e_y^2(t). \tag{2.14}$$

2.2.2 MTF と音環境

ここでは，MTF と音環境の関係性について述べる．MTF の特性が雑音・残響環境においてどのように表現でき，どのような特性を有しているのかについて述べる．

まず，雑音環境での MTF の特性について述べる．雑音には，式 (2.10) の定常加法性雑音 $\mathbf{n}(t)$ を考え，入力信号と出力信号の関係は，次式となる．

$$\mathbf{y}_N(t) = \mathbf{x}(t) + \mathbf{n}(t), \tag{2.15}$$

ここで，$\mathbf{x}(t)$ と $\mathbf{y}_N(t)$ は，原音声と観測信号（雑音音声）である．二乗集合平均を適用することで，次式が求まる．

$$\begin{aligned}
\langle \mathbf{y}_N^2(t) \rangle &= \langle \mathbf{x}^2(t) \rangle + \langle \mathbf{n}^2(t) \rangle, \\
e_{yN}^2(t) &= e_x^2(t) + e_n^2(t) \\
&= \overline{e_x^2}(1 + \cos(2\pi f_m t)) + \overline{e_n^2} \\
&= \left(\overline{e_x^2} + \overline{e_n^2}\right)\{1 + m_N(f_m)\cos(2\pi f_m t)\},
\end{aligned}$$

ただし，$\overline{e_n^2} = (1/T)\int_0^T e_n^2(t)dt$ である．T は，信号継続時間である．ここで，$e_n^2(t)$ は時間領域において変化が一定（定常）であると仮定することで，雑音環境における MTF は次式で表現できる．

$$m_N(f_m) = \frac{\overline{e_x^2}}{\overline{e_x^2} + \overline{e_n^2}} = \frac{1}{1 + 10^{-\frac{\mathrm{SNR}}{10}}}, \tag{2.16}$$

ここで, dB 表記の信号対雑音比（SNR）は，$\mathrm{SNR} = 10\log_{10}(\overline{e_x^2}/\overline{e_n^2})$ となる．雑音環境の MTF の特性を図 2.5 に示す．MTF の特性は，SNR の変数によって決まる．例えば，SNR = 10 dB の場合，全変調周波数帯域において変調度は 0.909 である，

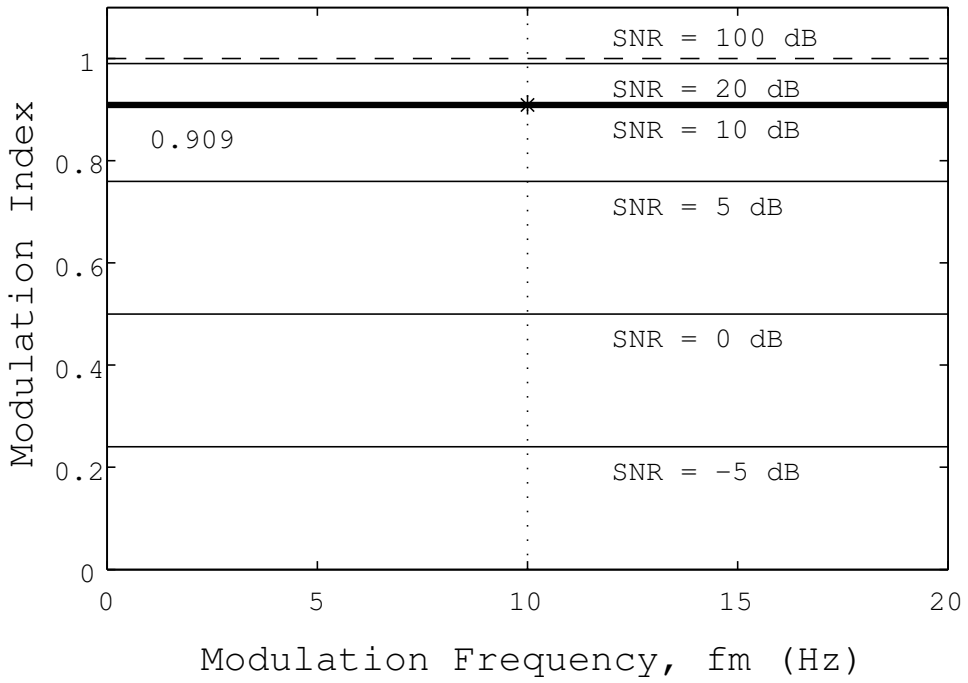


図 2.5: 雑音環境での MTF の特性：実線は SNR = 10 dB の MTF [144].

次に，残響環境における残響信号 $\mathbf{y}_R(t)$ は，原信号 $\mathbf{x}(t)$ と $\mathbf{h}(t)$ の畳み込みによって表現される．

$$\mathbf{y}_R(t) = \int_0^t \mathbf{h}(\tau)\mathbf{x}(t-\tau)d\tau. \tag{2.17}$$

この式は二乗集合平均により次式で表現される．

$$\begin{aligned} \langle \mathbf{y}_R^2(t) \rangle &= e_{yR}^2(t) = e_x^2(t) * e_h^2(t) \\ &= \frac{\overline{e_x^2}}{\alpha}\left(1 + \frac{\beta}{\alpha}\cos(2\pi f_m t)\right), \end{aligned} \tag{2.18}$$

ただし，$\alpha = \int_0^\infty \mathbf{h}^2(t)dt$ と $\beta = \int_0^\infty \mathbf{h}^2(t)\exp(-j\omega_m t)dt$ である．ここで，式 (2.7) と (2.18) から，$m(f_m) = \beta/\alpha$ であることがわかる．そして，複素表現の MTF は次式で表現される．

$$\mathbf{m}_R(\omega) = \frac{\int_0^\infty \mathbf{h}^2(t)\exp(-j\omega t)dt}{\int_0^\infty \mathbf{h}^2(t)dt}. \tag{2.19}$$

これは，$\mathbf{h}^2(t)$ の Fourier 変換である．式 (2.19) の $\mathbf{h}^2(t)$ を式 (2.9) で置き換えるこ

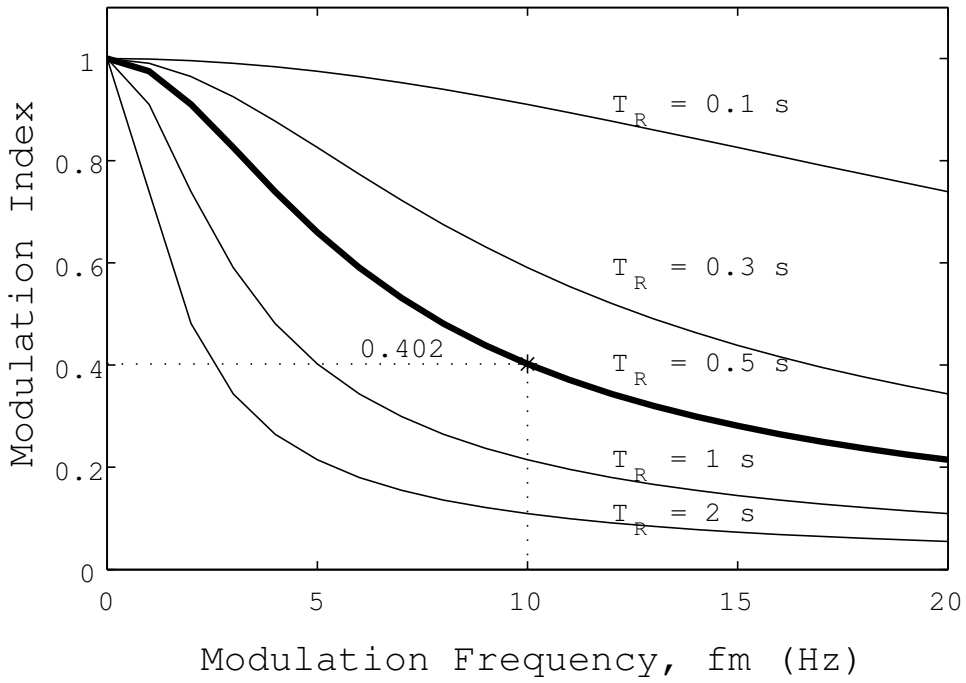


図 2.6: 残響環境での MTF の特性：実線の $T_R = 0.5$ s の MTF [144].

とにより，残響環境における MTF は次式として書き換えられる．

$$m_R(f_m) = |\mathbf{m}_R(\omega)| = \frac{1}{\sqrt{1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2}}. \tag{2.20}$$

残響環境における MTF の特性を図 2.6 に示す．残響環境における MTF の特性は，T_R によって決定される一種の低域通過特性を有している．図からもわかる通り，T_Rが長くなるにつれて，変調度が小さくなる．例えば，$T_R = 0.5$ s の時の変調周波数 $f_m = 10$ Hz における変調度は，0.402 であり，約 0.6 低下していることになる．

最後に雑音残響環境での MTF について考える．雑音残響環境における MTF の特性は，変調周波数領域での残響環境の MTF と雑音環境の MTF との掛け合わせで表現される．式 (2.11) を用いて，入力信号と出力信号の関係は次式で表現される．

$$\begin{aligned} \mathbf{y}(t) &= \int_0^t \mathbf{h}(\tau)\mathbf{x}(t-\tau)d\tau + \mathbf{n}(t), \\ e_y^2(t) &= e_x^2(t) * e_h^2(t) + e_n^2(t) \qquad (2.21) \\ &= \frac{\overline{e_x^2}}{\alpha}\left[1 + \frac{\beta}{\alpha}\cos(2\pi f_m t)\right] + e_n^2(t). \qquad (2.22) \end{aligned}$$

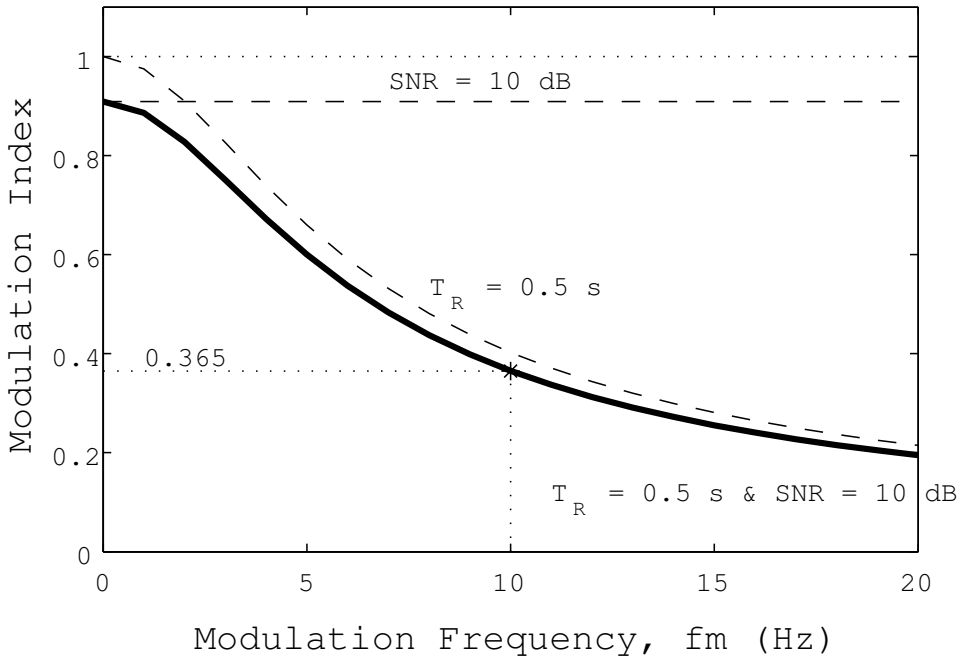


図 2.7: 雑音残響環境環境での MTF の特性：太線は T_R=0.5 s かつ SNR = 10 dB 条件下の MTF [144].

雑音残響環境の MTF は，式 (2.20) と 式 (2.16) から次式のように表現される．

$$\begin{aligned} m(f_m) &= m_R(f_m) \cdot m_N(f_m) \\ &= \frac{1}{\sqrt{1+\left(2\pi f_m \frac{T_R}{13.8}\right)^2}\left(1+10^{-\frac{\mathrm{SNR}}{10}}\right)}. \end{aligned} \tag{2.23}$$

ここで，雑音残響環境の MTF の特性を図 2.7 に示す．雑音残響環境の特性は，雑音環境と残響環境の両方の特性を有する．例えば，雑音残響環境の MTF の特性である図 2.7 は，図 2.6 と図 2.5 の掛け合わせであり，$T_R = 0.5$ s，SNR = 10 dB，$f_m = 10$ Hz での変調度は，0.365（$= 0.402 \times 0.909$）となる．

このように雑音には定常雑音を仮定し，残響環境には拡散音場を仮定しているという制約条件はあるものの，従来のアプローチでは雑音と残響を同時に扱うことが出来なかった雑音残響環境を，MTF により同時に扱うことができる．しかも，雑音残響環境を SNR と T_R のたった二つのパラメータで表現できることは，音環境バリアフリーを実現する上でも非常に有効と考える．雑音除去と残響除去をはじめから個別の問題として考える従来のアプローチによる雑音残響除去では，雑

音除去と残響除去（雑音に関するパラメータと残響に関するパラメータ）が連動していないために過少・過剰回復につながっていた．MTF を用いることにより雑音に関するパラメータである SNR と残響に関するパラメータである T_R の二つのパラメータを MTF 上で同時に扱うことができる．雑音残響環境において MTF を用いた二つのパラメータによるパワーエンベロープ処理体系を実現することが，音環境バリアフリーにおいて効果的であることを，次節で示す．

2.3 音環境バリアフリーと逆フィルタ処理

音声コミュニケーションにおいては，雑音や残響が音環境のバリアとなるため，音環境のバリアを測定・推定して，観測信号から音環境のバリアを相殺することの必要性は，1 章で述べた通りである．システムでの伝達系を考えると，入力が原信号，伝達関数が雑音残響（音環境のバリア）の時，出力は観測信号（雑音残響信号）であり，相殺するということは観測信号に対して伝達関数の逆関数を掛けることである．すなわち，観測信号に対する逆フィルタ処理により音環境バリアフリーが実現できることである．そこで，本節では，雑音残響環境において MTF に基づくことにより音環境バリアフリーのためのパワーエンベロープ処理体系が実現できることを，パワーエンベロープを Fourier 変換した変調スペクトル上において，逆フィルタを用いて最適解が求まることで示す．

2.3.1 音環境と変調スペクトル

ここでは，原信号のパワーエンベロープ $e_x^2(t)$，雑音残響信号のパワーエンベロープ $e_y^2(t)$，雑音信号のパワーエンベロープ $e_{yn}^2(t)$，残響信号のパワーエンベロープ $e_{yh}^2(t)$，雑音のパワーエンベロープ $e_n^2(t)$，RIR のパワーエンベロープ $e_h^2(t)$ を Fourier 変換して求めた変調スペクトルをそれぞれ，$E_x(f_m)$，$E_y(f_m)$，$E_{yn}(f_m)$，$E_{yh}(f_m)$，$E_n(f_m)$，$E_h(f_m)$，とする．また，原信号のパワーエンベロープ $e_x^2(t)$ と雑音残響信号のパワーエンベロープ $e_y^2(t)$ を式 (2.4) と式 (2.5) として話を進める．

まず，雑音信号の変調スペクトルについて示す．雑音として時間・周波数共に一定な雑音を仮定していることから，雑音のパワーエンベロープ $e_n^2(t) = \overline{e_n^2(t)}$ の変

調スペクトル $E_n(f_m)$ は，単位ステップ信号を Fourier 変換していることとなるため直流分の $E_n(0)$ にしかパワーを持っていないことになる．雑音信号の変調スペクトル $E_{yn}(f_m)$ は次式のような関係が成り立つ．

$$E_{yn}(f_m) = E_x(f_m) + E_n(f_m), \tag{2.24}$$

$$= E_x(f_m) \cdot E_N(f_m). \tag{2.25}$$

ここで，$E_N(f_m)$ は雑音に関する変調伝達関数であり，雑音の変調スペクトル $E_n(f_m)$ とは異なる．上記の関係から次式が成り立つ．

$$E_x(f_m) = E_{yn}(f_m) - E_n(f_m), \tag{2.26}$$

$$E_N(f_m) = \frac{E_{yn}(f_m)}{E_x(f_m)}. \tag{2.27}$$

この原信号と変調スペクトルの関係を解りやすく図2.8に示す．雑音の影響は $E_{yn}(0)$ のみに現れ，それ以外の変調周波数は原信号の変調スペクトル $E_x(f_m)$ の値を取ることとなる．雑音に関する伝達関数 $E_N(f_m)$ は次式のように求まる．

$$E_N(f_m) = G(0) \cdot m_N(f_m) \tag{2.28}$$

$$= \frac{E_{yn}(f_m)}{E_{yn}(f_m) - E_n(f_m)}. \tag{2.29}$$

ここで，$e_n^2(t)$ は直流分しかパワーを持たないことから，$G(0)$, $E_n(0)$, $m_N(0)$ となり次式のように求まる．

$$G(0) = \left(\frac{E_{yn}(f_m)}{E_{yn}(f_m) - E_n(0)} \right) \frac{1}{m_N(0)}, \tag{2.30}$$

$$= \left(\frac{E_x(f_m) + E_n(0)}{E_x(f_m)} \right) \frac{1}{m_N(0)}, \tag{2.31}$$

$$= \frac{1}{m_N^2(0)}. \tag{2.32}$$

雑音に関する伝達関数 $E_N(f_m)$ は，式 (2.28) と式 (2.40) より次式で求まる．

$$E_N(f_m) = G(0) \cdot m_N(0). \tag{2.33}$$

次に，残響信号の変調スペクトルについて示す．残響信号の変調スペクトル $e_{yh}(f_m)$ は次式のように表現できる．

$$E_{yh}(f_m) = E_x(f_m) \cdot E_h(f_m). \tag{2.34}$$

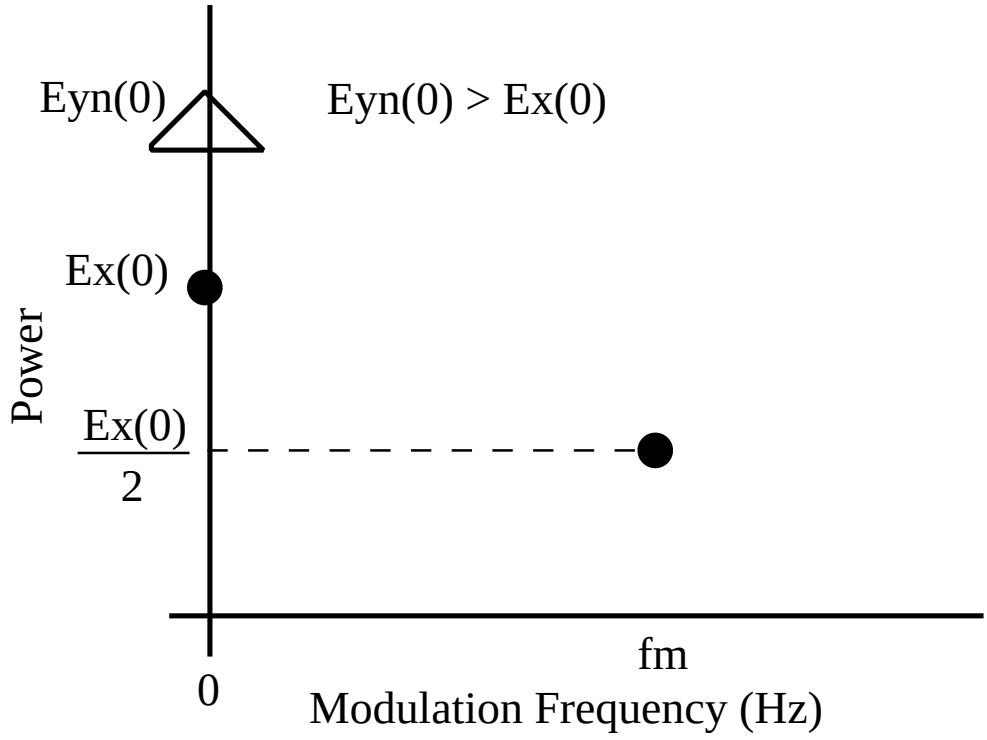


図 2.8: 原信号と雑音信号の変調スペクトルの関係.

ここで，$E_h(f_m)$ は RIR に関する伝達関数であり，次式で表現される．

$$E_h(f_m) \quad = \quad H \cdot m_R(f_m). \tag{2.35}$$

この時の H は次式となる．

$$H = a^2 = 1. \tag{2.36}$$

原信号と残響信号の変調スペクトルの関係を解りやすく図 2.9 に示す．変調スペクトルの直流分は等しく $E_{yh}(0) = E_x(0)$，残響の影響により $E_{yh}(0)$ 以外の変調周波数で残響時間に応じて変調度が変化することになる．

最後に雑音残響環境である．ここでは，MTF と同様に雑音に関する伝達関数と残響に関する伝達関数を掛け合わせた特性となり，図 2.10 に示す通りである．雑音の影響により直流分が増加し，残響の影響により $E_y(0)$ 以外の変調周波数で減衰する特性となる．

$$E_y(f_m) \quad = \quad E_x(f_m) \cdot E_h(f_m) \cdot E_N(f_m). \tag{2.37}$$

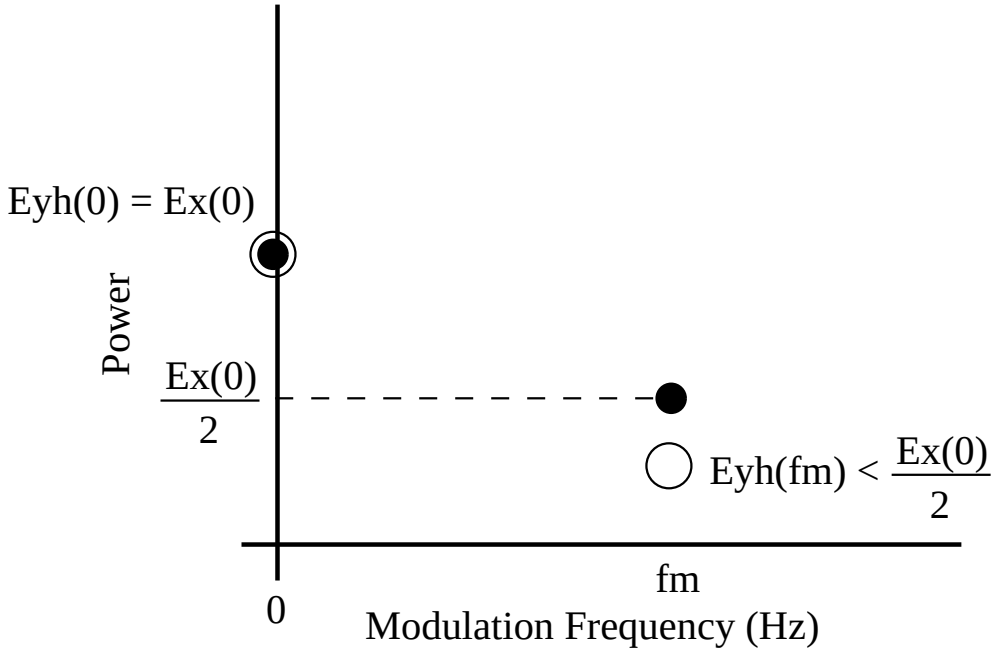


図 2.9: 原信号と残響信号の変調スペクトルの関係.

$E_h(f_m)$ と $E_N(f_m)$ は $m_R(f_m)$ と $m_N(f_m)$ を通して求められることから，T_R と SNR をパラメータとして求まることがわかる．このような変調スペクトル上での特徴を利用した規範を用いた逆フィルタ処理による最適化問題を次で述べる．

2.3.2 雑音残響環境での最適化問題

雑音残響信号のパワーエンベロープから原信号のパワーエンベロープと相関係数が 1 の回復パワーエンベロープを求めるということは，回復パワーエンベロープの変調度が 1（100 % 変調）の状態である．これは回復パワーエンベロープが最適にな状態（過少・過剰回復になっていない）である．この最適問題を解くことができれば，音環境バリアフリーのためのパワーエンベロープ処理体系の理論を確立することとなる．そこで，本節では，雑音残響環境において最適化問題を解くためのアプローチを述べる．

本研究の雑音残響音声に対する音声信号処理は，MTF の概念に基づき雑音残響音声のパワーエンベロープが 100 % 変調となるように SNR と残響時間，音声区間を最適化することで，音環境バリアフリーを実現できる．言うなれば，SNR と残

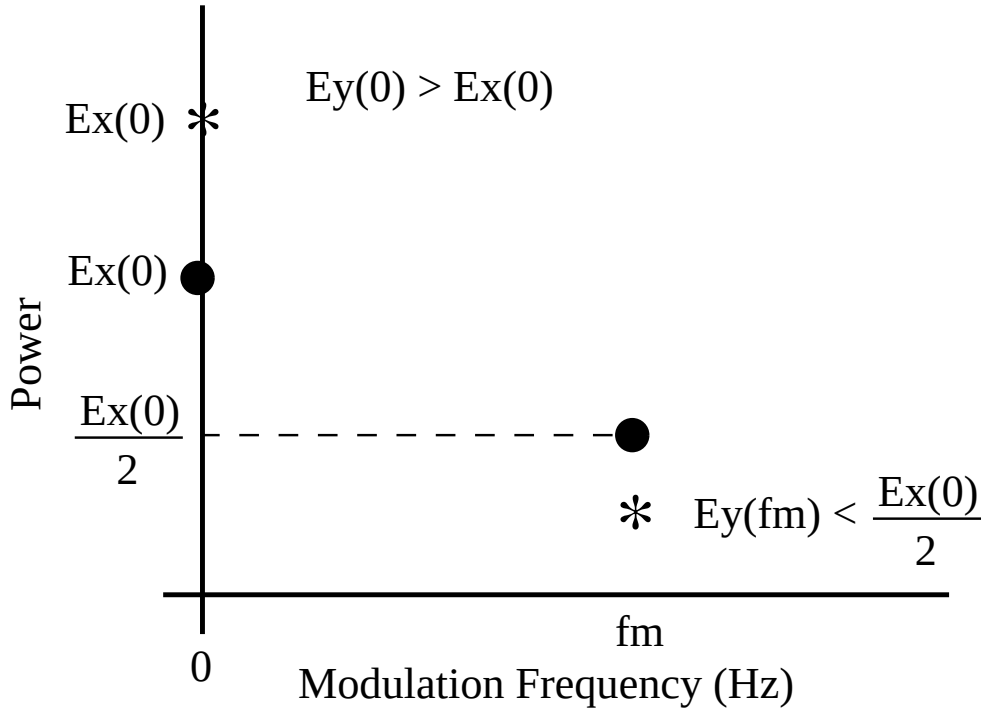


図 2.10: 原信号と雑音残響信号の変調スペクトルの関係.

響時間，回復パワーエンベロープ，音声区間を逆フィルタ処理によってまとめて得られるということを意味する．

ここでは，推定されたパラメータ（推定された SNR $\hat{SNR}$，推定された残響時間 $\hat{T}_R$，検出された音声区間 $\hat{S}_{VAD}$）を変数として MTF に基づく逆フィルタ処理により，回復パワーエンベロープ $\hat{e}_x^2(t)$ が求まる．

$$\hat{e}_x^2(t) = \mathrm{IMTF}\left[e_y^2(t; \hat{S}_{VAD}(t)), \hat{SNR}, \hat{T}_R\right]. \tag{2.38}$$

ここで，IMTF[·] は MTF に基づく逆フィルタ処理であり，$\hat{S}_{VAD}$ は音声区間の場合 1，非音声区間の場合 0 の値を持つ．

最適化についての詳細を述べる．回復パワーエンベロープ $\hat{e}_x^2(t)$，任意のパワーエンベロープ $\tilde{e}_x^2(t)$ を Fourier 変換した変調スペクトルを $\hat{E}_x$，$\tilde{E}_x(f_m)$ と表現する．MTF に基づく逆フィルタ処理は，次式のようになる．

$$\mathrm{IMTF}[\cdot] = \frac{E_y(f_m; \hat{S}_{VAD}(t))}{E_h(f_m; \hat{T}_R) E_N(f_m; \hat{SNR})}. \tag{2.39}$$

ここで，変調スペクトル上での逆フィルタ処理について述べる．雑音除去は，次

式のように雑音に関する伝達関数 $E_N(f_m)$ の逆フィルタで回復した変調スペクトルが得られる．

$$\hat{E}_x(f_m) = \frac{E_{yn}(f_m)}{E_N(f_m)}, \tag{2.40}$$

$$= E_{yn}(f_m)\left(\frac{E_{yn}(f_m) - E_n(f_m)}{E_{yn}(f_m)}\right). \tag{2.41}$$

このように，$E_N(f_m)$ が Wiener filter になっていることがわかる．そして次式の逆フィルタにより回復できることがわかる．

$$\hat{E}_x(f_m) = \frac{E_{yn}(f_m)}{G(0)\cdot m_N(0)} \tag{2.42}$$

次に，残響除去は次式のように RIR に関する伝達関数 $E_h(f_m)$ の逆フィルタ処理で求まる．

$$\hat{E}_x(f_m) = \frac{E_{yh}(f_m)}{E_h(f_m)}. \tag{2.43}$$

最後に雑音残響環境について考える．この時の雑音残響信号の変調スペクトル $E_y(f_m)$ と回復変調スペクトル $\hat{E}_x(f_m)$ は，次式で表現される．

$$E_y(f_m) = E_x(f_m)\cdot E_h(f_m)\cdot E_N(f_m), \tag{2.44}$$

$$\hat{E}_x(f_m) = \frac{E_y(f_m)}{E_h(f_m)E_N(f_m)} \tag{2.45}$$

このように，雑音・RIR に関する伝達関数の逆フィルタ処理により回復変調スペクトルが得られる．わかりやすくするために音声区間 S_{VAD} についての記述を省略したが，これらの処理は，S_{VAD} が 1 の音声区間について行われる必要がある．

IMTF による回復処理は，次式のような SNR, 残響時間, 音声区間をパラメータとして変調スペクトル上で 100 % 変調となるように最適化することを意味している．

$$\hat{E}_x\left(f_m; \hat{SNR}, \hat{T}_R, \hat{S}_{VAD}(t)\right) = \underset{\substack{SNR_{\min}\le SNR\le SNR_{\max}\\ 0\le T_R\le T_{R,\max}\\ 0\le S_{VAD}(t)\le 1}}{\arg\min} \left\{\epsilon_{MSR}\left(\tilde{E}_x\left(f_m; SNR, T_R, S_{VAD}(t)\right)\right)\right\}, \tag{2.46}$$

ここで，$\epsilon_{MSR}\left(\tilde{E}_x\left(f_m; SNR, T_R, S_{VAD}(t)\right)\right)$ は，SNR, T_R, $S_{VAD}(t)$ の時の任意の変調スペクトル $\tilde{E}_x(f_m)$ の変調周波数 f_m での誤差である．このように，誤差最小

にすることで SNR, RT, 音声区間と回復変調スペクトル（逆 Fourier 変換による回復パワーエンベロープ）をまとめて得ることができる．誤差は RMS (Root Mean Square) を用いて次式により求まる．

$$\begin{aligned}&\epsilon_{MSR}\left(\tilde{E}_x\left(f_m; SNR, T_R, S_{VAD}(t)\right)\right)\\ &= \sqrt{\left(\frac{\tilde{E}_x(0; SNR, T_R, S_{VAD}(t))}{2} - \tilde{E}_x(f_m; SNR, T_R, S_{VAD}(t))\right)^2}. \quad (2.47)\end{aligned}$$

原信号の変調スペクトルは，$E_x(0)/2 = E_x(f_m)$ という関係が成り立つ．一方で，残響信号の変調スペクトル $E_{yh}(f_m)$ では，$E_x(0) = E_{yh}(0)$ と $E_x(f_m) > E_{yh}(f_m)$ という関係，雑音の変調スペクトル $E_{yn}(f_m)$ では，$E_x(0) < E_{yn}(0)$ と $E_x(f_m) = E_{yn}(f_m)$ という関係が成り立つ．そこで，任意の $\tilde{E}_x(f_m)$ の T_R と SNR を変化させ，変調度 1 の $\tilde{E}_x(0)/2 = /\tilde{E}_x(f_m)$ が成り立つ時に誤差が 0 となるように設定した．式 (2.46) において各パラメータの値を変えることで $\tilde{E}_x(f_m; SNR.T_R, S_{VAD}(t))$ が求まる．

一方，パラメータは，VAD における誤差を最適化することで推定することができ，次式のように表現される．

$$\left\{\hat{SNR}, \hat{T}_R, \hat{S}_{VAD}(t)\right\} = \underset{\substack{SNR_{\min}\le SNR\le SNR_{\max}\\ 0\le T_R\le T_{R,\max}}}{\arg\min} \left\{\epsilon_{VAD}\left(SNR, T_R, S_{VAD}(t)\right)\right\}, \quad (2.48)$$

推定された VAD の誤差が 0 になるように，SNR と T_R を最適化して回復されたパワーエンベロープ $\tilde{e}_x^2(t)$ を得ることで，これらのパラメータをまとめて得られる．ここで，VAD の誤差 ϵ_{VAD} は次式のようになる．

$$\begin{aligned}&\epsilon_{VAD}\left(SNR, T_R, \tilde{S}_{VAD}(t)\right)\\ &= \sqrt{\left(\frac{FAR^2 + FRR^2}{2} - \tilde{E}_x(f_m; SNR, T_R, S_{VAD}(t))\right)^2}. \quad (2.49)\end{aligned}$$

表記の関係上，上式では省略して記述したが，FAR^2 は $FAR^2(\tilde{S}_{VAD}(t; B, \tilde{e}_x^2(t; SNR, T_R)))$，$FRR^2$ は $FRR^2(\tilde{S}_{VAD}(t; B, \tilde{e}_x^2(t; SNR, T_R)))$ である．B はパワー閾値である．誤

受理率 FAR と誤棄却率 FRR は次式のようになる．

$$
\begin{aligned}
&FAR(\tilde{S}_{VAD}(t;B,\tilde{e}_x^2(t;SNR,T_R))) \\
&= \frac{N_{FA}\left(\tilde{S}_{VAD}(t;B,\tilde{e}_x^2(t;SNR,T_R))\right)}{N_{ns}(S_{VAD}(t;e_x^2(t)))} \times 100, \qquad (2.50) \\
&FRR(\tilde{S}_{VAD}(t;B,\tilde{e}_x^2(t;SNR,T_R))) \\
&= \frac{N_{FR}\left(\tilde{S}_{VAD}(t;B,\tilde{e}_x^2(t;SNR,T_R))\right)}{N_s(S_{VAD}(t;e_x^2(t)))} \times 100. \qquad (2.51)
\end{aligned}
$$

N_{FA} は音声として判別された非音声信号の時間，N_{FR} は非音声として判別された音声信号の時間，N_{ns} は非音声としての信号時間，N_s は音声としての信号時間であり，次式で示される．

$$
\begin{aligned}
&N_{FA}\left(\tilde{S}_{VAD}(t;B,\tilde{e}_x^2(t;SNR,T_R))\right) \\
&= \int_0^T \left(\tilde{S}_{VAD}\left(t;B,\tilde{e}_x^2(t;SNR,T_R)\right) - S_{VAD}(t;e_x^2(t))\right) dt, \qquad (2.52) \\
&N_{FR}\left(\tilde{S}_{VAD}(t;B,\tilde{e}_x^2(t;SNR,T_R))\right) \\
&= \int_0^T \left(S_{VAD}(t;e_x^2(t)) - \tilde{S}_{VAD}\left(t;B,\tilde{e}_x^2(t;SNR,T_R)\right)\right) dt, \qquad (2.53) \\
&N_{ns}\left(S_{VAD}(t;e_x^2(t))\right) \\
&= \int_0^T S_{VAD}(t;e_x^2(t))dt, \qquad (2.54) \\
&N_s\left(S_{VAD}(t;e_x^2(t))\right) \\
&= \int_0^T \left(1 - S_{VAD}(t;e_x^2(t))\right) dt. \qquad (2.55)
\end{aligned}
$$

そして，$S_{VAD}(t;e_x^2(t))$ は正解音声区間，$\tilde{S}_{VAD}\left(t;B,\tilde{e}_x^2(t;SNR,T_R)\right)$ は次のようにパワー閾値 B により求めることができ，1 か 0 の値を取る．

$$
\tilde{S}_{VAD}\left(t;B,\tilde{e}_x^2(t;SNR,T_R)\right) = \begin{cases} 1 & (\tilde{e}_x^2(t;SNR,T_R) > 0) \\ 0 & (\tilde{e}_x^2(t;SNR,T_R) = 0) \end{cases} \qquad (2.56)
$$

このような音声信号処理では，変調度 1 に基づく逆フィルタ処理を用いて，SNR, T_R, S_{VAD} を最適化することで，回復パワーエンベロープと各パラメータをまとめて得られることがわかる．

2.3.3 MTF の逆フィルタ処理による最適化

ここでは，実際に前述の最適化問題について，大域的最適解に陥ることを示す．まず，原信号のパワーエンベロープと雑音のパワーエンベロープ，残響のパワーエンベロープを次式に示す通りとする．

$$e_x^2(t) = 2(1 + 1 \times \sin(2\pi \tilde{f_m} t - \frac{\pi}{2}), \tag{2.57}$$

$$e_n^2(t) = \overline{e_n^2(t)} = 3, \tag{2.58}$$

$$e_h^2(t) = a^2 \exp\left(-\frac{13.8t}{T_R}\right), \tag{2.59}$$

$$a^2 = \frac{1}{\int_0^T \exp\left(-\frac{13.8t}{T_R}\right) dt} \tag{2.60}$$

ここで，$\tilde{f_m}$ は任意の変調周波数で 5 Hz に設定し，残響時間 T_R は 0.4 s とした．また，サンプリング周波数 f_s は 40 Hz とした．

最適化を行うにあたり，残響時間の変化させる範囲は 0.00〜3.00 s として刻み幅は 0.05 s とした．SNR の範囲は −10〜20 dB として刻み幅は 0.5 dB とした．任意の変調周波数は，$E_y(f_m)$ から次式によりブラインドで決定した．

$$\tilde{f_m} = \underset{0<f_m \leq f_{m,max}}{\arg\max} \tag{2.61}$$

また，変調度 1 を規範とした誤差を求めるために，0 Hz における差分 ϵ_0 と 5 Hz における差分 $\epsilon_{\tilde{f_m}}$ は，次式とした．

$$\epsilon_0 = 10 \log_{10} \tilde{E}_x(0) - 10 \log_{10} E_x(0), \tag{2.62}$$

$$\epsilon_{\tilde{f_m}} = 10 \log_{10} E_x(\tilde{f_m}) - 10 \log_{10} \tilde{E}_x(\tilde{f_m}). \tag{2.63}$$

そして，ϵ_0 と 5 Hz における差分 $\epsilon_{\tilde{f_m}}$ から RMS を次式により計算する．

$$RMS = \sqrt{\frac{\epsilon_0^2 + \epsilon_{\tilde{f_m}}^2}{2}}. \tag{2.64}$$

SNR と残響時間 T_R を変化させたときの RMS の結果を図 2.11 に示す．大局的最適解が得られることがわかる．この時の残響時間は 0.40 s，SNR は −3.00 dB であり，正確な残響時間は 0.40 s，SNR は −3.22 であり，刻み幅を考慮すると正しい所で大局的最適解に陥っていることがわかる．このように，SNR と残響時間とい

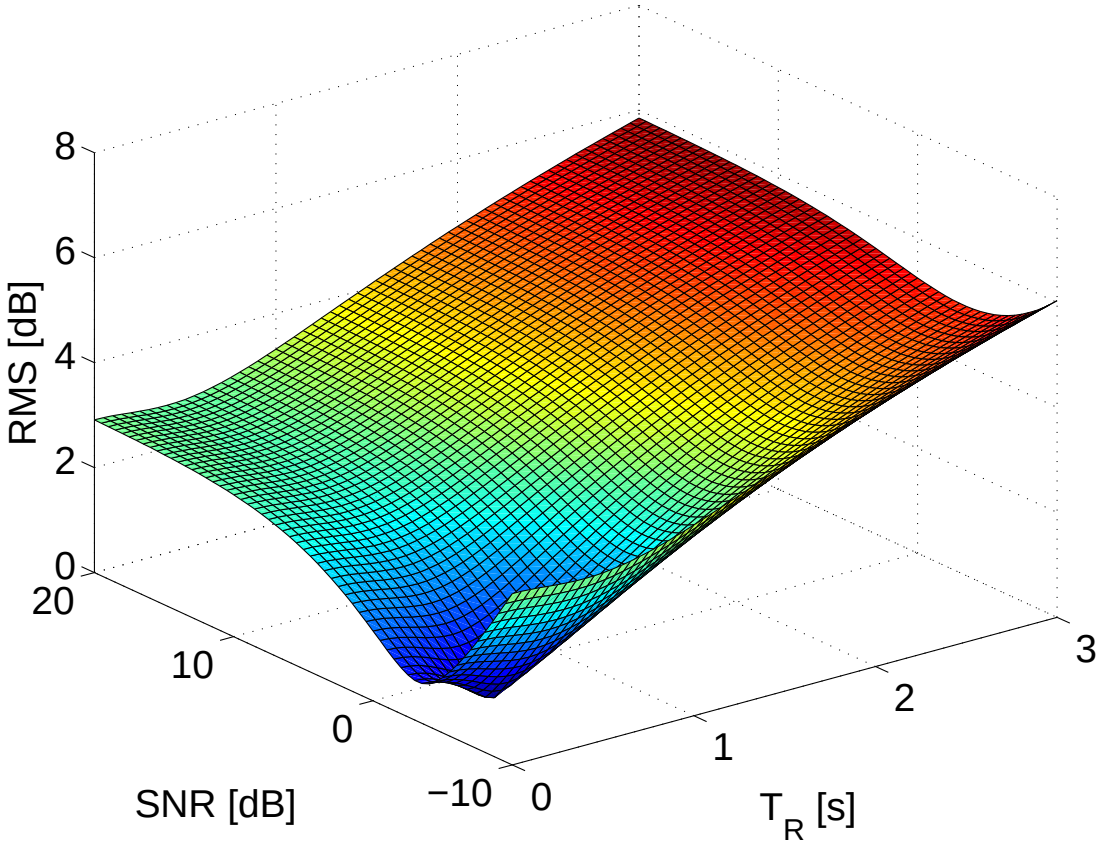


図 2.11: 最適化の結果.

う雑音と残響の二つのパラメータを用いた MTF に基づく逆フィルタ処理により，変調度 1 に近い所（100 % 回復であり，ほとんど過少・過剰回復になっていない）へ回復が可能であり，雑音と残響を同時に扱える処理であることから，このパワーエンベロープ処理体系が音環境バリアフリーの実現に有効なアプローチであることが示せた.

第 3章

パワーエンベロープ処理体系に基づく統合的音声信号処理

3.1 統合的音声信号処理の概要

前章では，音声バリアフリーのためのパワーエンベロープ処理体系についての理論を変調周波数上で述べたが，本章では，パワーエンベロープ処理体系を実現するための時間領域における統合的音声信号処理を提案する．統合的音声信号処理は，音声区間検出（VAD），パワーエンベロープ回復処理，パラメータ推定（SNR，残響時間）によって構成される．この概要図を図 3.1 に示す．パワーエンベロープ回復処理は，回復パワーエンベロープ単体だけでなく，音声区間検出にも利用される．そして，パラメータ推定は音環境の情報を推定するために必要であるため，パワーエンベロープ回復処理にも必要不可欠である．

統合的音声信号処理は，様々な音声信号処理技術によって構成されるが，全ての処理をパワーエンベロープ処理体系における変調度 1 を規範とした処理を踏襲することで実現している．VAD ならびに SNR 推定ではパワー閾値最適化を用い，残響時間推定では時間領域における変調度 1 を規範とした処理を行っている．これら，VAD と SNR 推定，残響時間推定という要素技術を利用して，変調度 1 へ回復するのがパワーエンベロープ回復処理である．

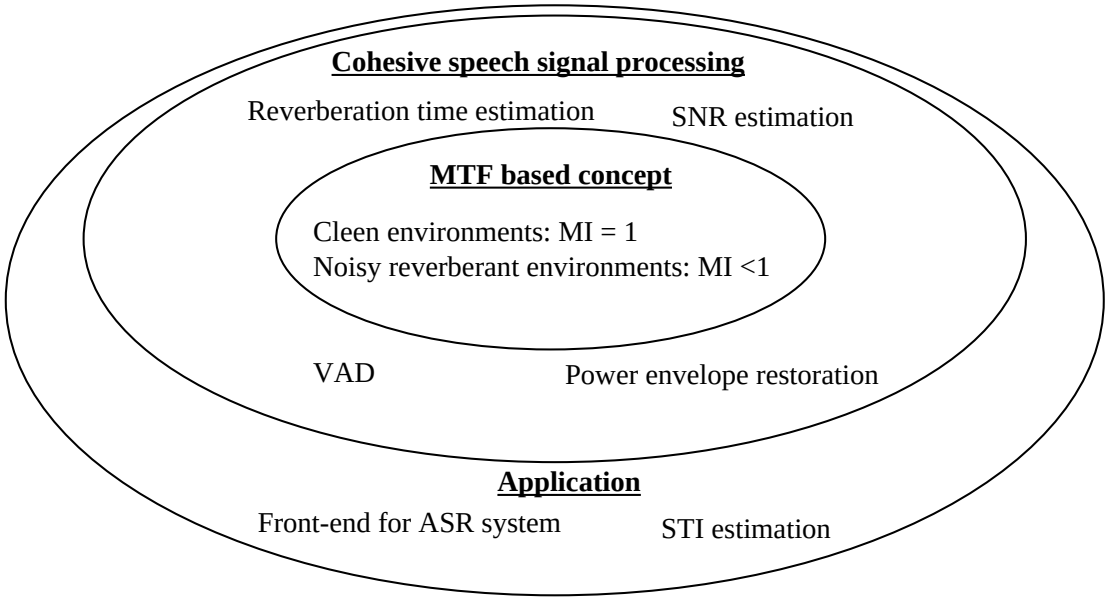


図 3.1: 統合的音声信号処理の概要図

3.2 音声区間検出

本節では，統合的音声信号処理において，パワーエンベロープ処理体系の全体を現わしている雑音残響に頑健な音声区間検出法 [145] について述べる．本 VAD 法では，パワーエンベロープ回復処理により回復されたパワーエンベロープに対してパワー閾値処理を行うことで雑音残響音声から精度よく音声区間を検出するというものであり，そのコンセプトは 2.1 節で述べた．

3.2.1 雑音残響に頑健な音声区間検出法の概要

これまでの音声区間検出（VAD）法は，1 章でも述べた通り音声符号化や ASR の前処理において利用されていた．このような目的では，音声区間を確実に検出する必要があり，FRR が 0 % となるように音声/非音声判別を設計する必要がある．しかし，2.1 節でも述べた通り，本研究では少しでも音声区間を正確に検出することを念頭に置いた雑音残響に頑健な VAD の実現を目指しており，音声/非音声判別において最適化したパワー閾値を用いた．雑音残響音声に頑健な VAD では，パワーエンベロープ回復処理（雑音残響除去）を行った回復パワーエンベロープに

対して最適なパワー閾値により音声/非音声判別を行うことで効果的に音声区間を検出することができる．これは，回復したパワーエンベロープと最適なパワー閾値により音声/非音声判別を行うことで，変調度 1（100 % 変調）のパワーエンベロープに対して音声/非音声判別を行った時と同等の結果を得るためである．

雑音残響に頑健な VAD のブロックダイアグラムは，2 章の図 2.3 に示した通りである．この手法は，(1) SNR 推定法（3.4.1 節で後述），(2) パワーエンベロープ回復処理（3.3 節で後述），(3) パワー閾値処理の三つの処理で構成される．最初の二つの処理は，観測信号のパワーエンベロープ（雑音残響信号のパワーエンベロープ）の加法性雑音と残響の影響を抑圧するのに利用される．雑音残響音声のパワーエンベロープ回復では，最初に SNR 推定法を利用して SNR を推定し，推定された SNR を利用して音声パワーエンベロープに含まれる加法性雑音成分を減算することで雑音除去を行う．そして，雑音の影響を取り除いたパワーエンベロープに対して，推定した残響時間を用いて MTF の逆フィルタ処理を行うことで残響除去を行う．最後に，回復パワーエンベロープを特徴として利用し，パワー閾値処理により音声/非音声判別を行うことで音声区間を検出する．提案する頑健な VAD 法では，SNR 推定法として後述する加法性雑音の帯域分割型 SNR 推定法（3.4.1 節）を利用する．

3.2.2 回復パワーエンベロープ

パワーエンベロープ回復処理は，MTF の概念に基づき，雑音除去として SNR をパラメータとした雑音パワーの減算処理と，残響除去として MTF の逆フィルタ処理によって構成される．逆問題を解くという観点から順序性が重要であり，雑音除去を行い，次に残響除去を行った．ここでも，パワーエンベロープ処理体系の概念である変調度 1 となることを念頭にした処理としている．残響時間は，推定した SNR から求めた雑音のパワーを，観測信号のパワーエンベロープから減算して，変調度 1 を規範に推定した．このような方法により，回復パワーエンベロープを得た．

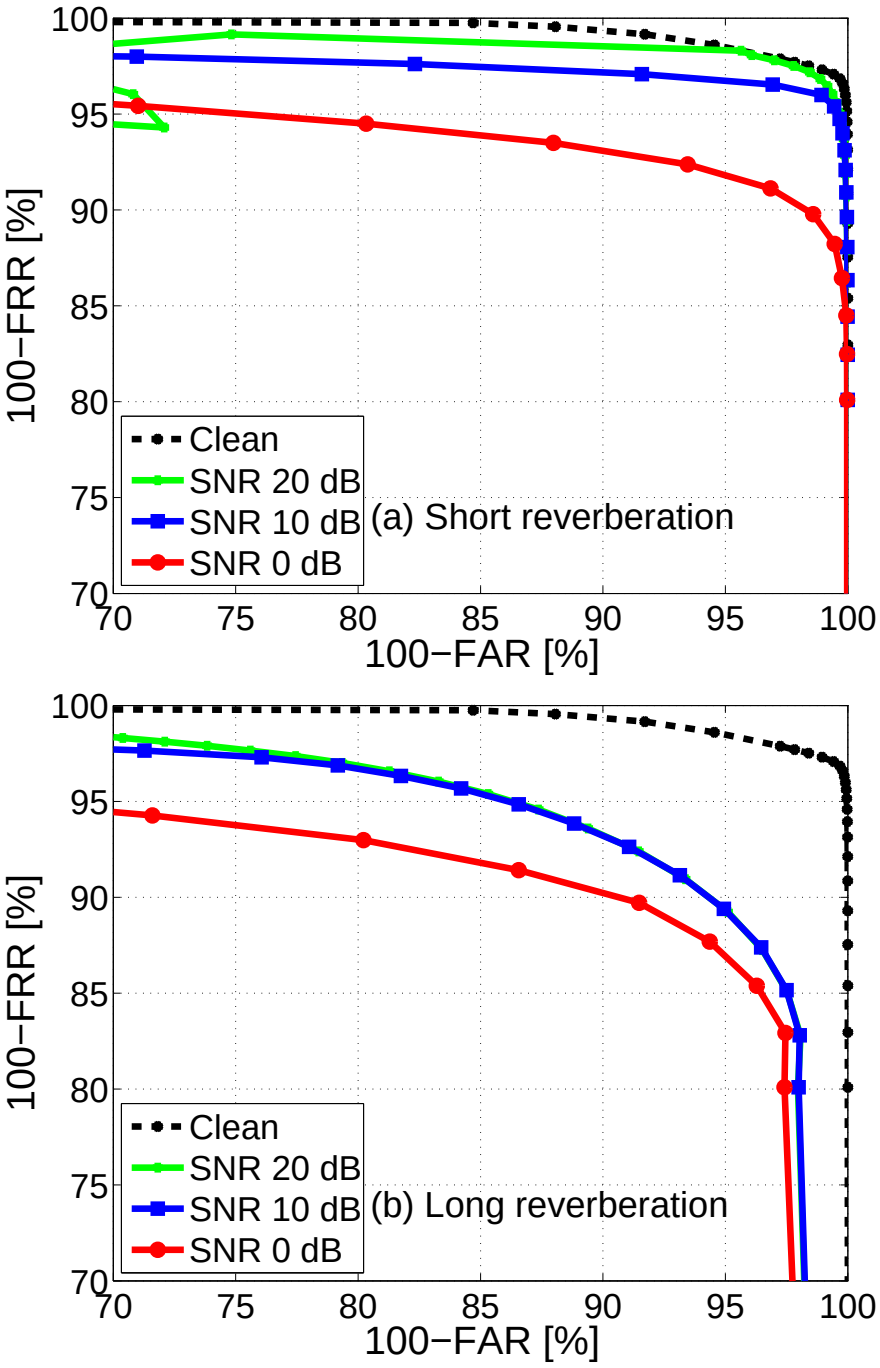


図 3.2: 雑音残響に頑健な VAD 法の ROC 曲線： (a) 残響時間が短い条件, and (b) 残響時間が長い条件.

3.2.3 閾値最適化

MTF に基づくパワーエンベロープ回復処理によって回復したパワーエンベロープを求め，音声・非音声判別は，回復パワーエンベロープに対するパワー閾値処

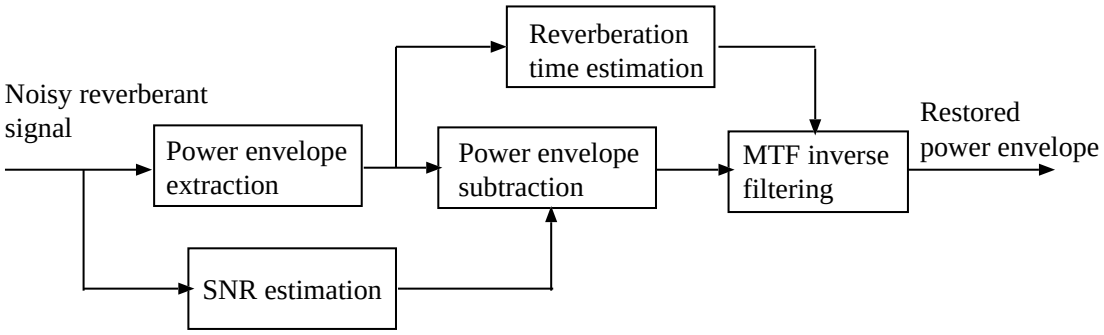


図 3.3: MTF に基づくパワーエンベロープ回復処理のブロックダイアグラム．

理によって行った．閾値の最適化は，2.1.2 節の手順によって求められ，パワー閾値は，ROC 曲線から決定した．

パワー閾値最適化に用いた雑音残響音声は，音声には AURORA-2J [146, 147] の学習データ 8440 音声，雑音には白色ガウス雑音を用いた．RIR $h(t)$ には，$T_R = 0.1, 0.3, 0.5, 1.0, 2.0, 3.0$ s の Schroeder の統計的な RIR [142] を利用した．雑音残響音声 $y(t)$ を式 (2.11) により求めた．残響環境の条件を，残響時間の短い環境（$T_R = 0.1, 0.3, 0.5$ s) と残響時間の長い環境（$T_R = 1.0, 2.0, 3.0$ s）に別けた．ROC 曲線を求めるにあたり，パワー閾値を -40〜0 dB で変化させた．

図 3.2 に雑音残響に頑健な VAD 法の ROC 曲線を示す．通常，縦軸・横軸には，FAR・FRR を示すが，ここでは，縦軸・横軸を 100−FRR・100−FAR とした．そのため，一般的な ROC 曲線と異なるが，曲線の角が右上に近いほど高性能であることを示す．残響が短い環境，長い環境の結果において，雑音の影響を SNR $= \infty, 20, 10, 0$ dB の 4 条件で示している．これらの結果より，雑音残響環境の全条件において，ROC 曲線の右凸の部分が 100−FRR・100−FAR ともに約 90 %以上であることから，非常によく機能していることがわかる．ROC 曲線の最適化から，音声/非音声判別のパワー閾値を -5 dB と決定した．

3.3 パワーエンベロープ回復処理

ここで，MTF に基づくパワーエンベロープ回復処理のブロックダイアグラムを図 3.3 に示す．パワーエンベロープ回復処理においても，変調度 1（100 % 変調）となるような雑音残響除去の処理が必要となる．2.3.2，2.3.3 節の理論通りの変調

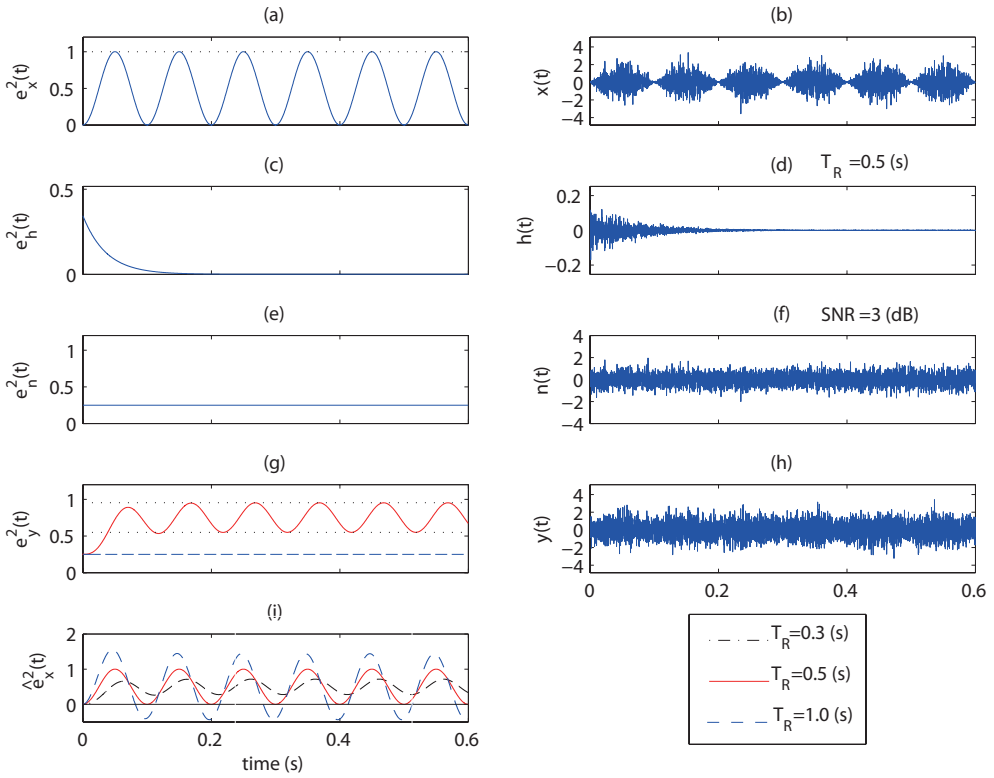


図 3.4: MTF に基づくパワーエンベロープ回復処理と各信号の関係例：(b) 原信号 $x(t)$ の (a) パワーエンベロープ $e_x^2(t)$，(d) インパルス応答 $h(t)$ ($T_R = 0.5$ s) の (c) パワーエンベロープ $e_h^2(t)$，(f) 雑音信号 $n(t)$ (SNR $= 3$ dB) の (e) $e_n^2(t)$，(g) $e_x^2(t) * e_h^2(t) + e_n^2(t)$ からの雑音残響信号のパワーエンベロープ $e_y^2(t)$，(h) $x(t) * h(t) + n(t)$ からの雑音残響信号 $y(t)$，(i) 回復パワーエンベロープ $\hat{e}_x^2(t)$

スペクトル上での最適化によるパワーエンベロープ回復処理は，音声信号に適用するにはまだまだ課題が多い．一方，時間領域でのパワーエンベロープ回復に着目すると，Unoki et al. によって大まかな方策が示されている [144]．この方法は，単純にパワーエンベロープ上で雑音残響除去を行うことを目的に提案されており，SNR 推定や音声区間の取り方について検討がなされていないだけでなく，初めから変調度 1 に着目した手法ではないものの，変調度 1 という考え方に合致する手法であることから，本研究では Unoki et al. によるパワーエンベロープ回復処理の方策を利用する．

本研究の雑音残響信号のパワーエンベロープは，原信号のパワーエンベロープに RIR パワーエンベロープが畳み込まれ，その残響信号のパワーエンベロープに

雑音のパワーエンベロープが加法されることで雑音残響信号のパワーエンベロープが求まるという式 2.21 の関係が成り立っている．パワーエンベロープ回復処理では，この逆問題を解いていくこととなり，雑音残響信号のパワーエンベロープから雑音のパワーを減算し，減算したパワーエンベロープに対して MTF を逆畳み込みすることにより，回復パワーエンベロープが求まることとなる．

MTF の概念に基づいて雑音残響信号からパワーエンベロープをどのように回復するのかを例として図 3.4 に示す．式 (2.8) に基づき，原信号のパワーエンベロープとして正弦波 （図 3.4(a)）$e_x^2(t)$ $(= 0.5(1 + \sin(2\pi f_m t)))$ とキャリアとして白色雑音 $c_x(t)$ を求め，原信号（図 3.4(b)）$x(t)$ は $e_x^2(t)$ と $c_x(t)$ の掛け算により求めた，ここで，変調周波数 f_m は 10 Hz，変調度 $m(f_m)$ は 1 とした．図 3.4(c) と (d) は，$T_R = 0.5$ s における $e_h^2(t)$ と式 (2.9) の RIR $h(t)$ である．図 3.4(e) と (f) は，式 (2.10) の SNR= 3 dB における $e_n^2(t)$ と $n(t)$ である．観測信号である雑音残響信号 $y(t)(= x(t) * h(t) * n(t))$ とそのパワーエンベロープ $e_y^2(t)(= e_x^2(t) * e_h^2(t) + e_n^2(t))$ を図 3.4(h) と 図 3.4(g) に示す．図の左側 ((a),(c),(e), (g)) はパワーエンベロープであり，図の右側 ((b), (d), (f), (h)) はそれぞれの信号である．図 3.4(i) の実線は，雑音残響信号のパワーエンベロープ $e_y^2(t)$(図 3.4(g)) から，$T_R = 0.5$ s と SNR $= 3$ dB として式 (3.7) と (3.2) により，最適に回復されたパワーエンベロープ $\hat{e}_x^2(t)$ である．点線は逆フィルタ処理の残響時間のパラメータを $T_R = 0.3$ s と過小推定した場合の回復パワーエンベロープであり，十分に回復されていないことがわかる．また，破線は残響時間のパラメータを $T_R = 1.0$ s と過大推定した場合の回復パワーエンベロープであり，過剰回復していることがわかる．したがって，パラメータを正確に推定することにより，最適なパワーエンベロープ回復を実現することができる．

パワーエンベロープ回復処理は，3.3.1 節で示すパワーエンベロープ抽出，3.3.2 節で示す雑音除去としてのパワーエンベロープ減算処理，3.3.3 節で示す残響除去としてのパワーエンベロープ逆フィルタ処理，3.4 節で示すパラメータ推定で構成されている．これから，それぞれの詳細を述べていく．

3.3.1 パワーエンベロープ抽出

音声信号のパワーエンベロープ $e_y^2(t)$ 抽出法は，ヒルベルト変換を用いて求める方法が Unoki et al. [37] により確立されており，次式で抽出できる．

$$e_y^2(t) = \mathbf{LPF}\left[|y(t) + j\mathbf{Hilbert}(y(t))|^2\right], \tag{3.1}$$

ここで，$\mathbf{Hilbert}(\cdot)$ は Hilbert 変換，$\mathbf{LPF}[\cdot]$ は低域通過フィルタである．この方法は，信号の瞬時振幅と低域通過フィルタを用いて瞬時振幅の高域成分を取り除くことで，パワーエンベロープを抽出する．低域通過フィルタの遮断周波数は，音声知覚や ASR で重要な変調周波数成分が約 20 Hz 以下であるという報告 [141, 148] に基づき 20 Hz としている．本研究でも，この方法を用いることとした．

3.3.2 パワーエンベロープ減算処理

パワーエンベロープ減算処理により，加法性雑音の雑音除去を行う．MTF に基づくパワーエンベロープ減算処理は，Yamasaki & Unoki により確立されている [149]．式 (2.16) において変調度と平均パワーは，雑音によって影響される．式 (2.22) の初項は，雑音残響信号 $e_y^2(t)$ から雑音の MTF $m_N(f_m)$ を用いて，次式で推定される．

$$\begin{aligned}\hat{e}_x^2(t) &= \overline{e_x^2}\left(1 + m_N(f_m)\cos(2\pi f_m t) \times \frac{1}{m_N(f_m)}\right) \\ &= e_y^2(t) - \overline{e_n^2}.\end{aligned} \tag{3.2}$$

雑音の平均パワー $\overline{e_n^2}$ は，観測信号のパワーエンベロープ $e_y^2(t)$ の非音声区間の平均値を用いるのが最も簡単な方法である．非音声区間を求めるにあたり，雑音残響に頑健な VAD 法が必要となり，非音声区間の正確な推定が雑音の平均パワーの推定値に直接影響し，パワーエンベロープ減算処理の性能に直結することがわかる．

頑健な VAD を提案するにあたり，頑健な VAD を利用することはできない．そこで，SNR と雑音の MTF $m_N(f_m)$ によるパワーエンベロープ減算処理する方法を提案した．

$$\overline{e_n^2(t)} = \overline{e_y^2(t)} \times (1 - m_N(SNR)), \tag{3.3}$$

$$\hat{e}_x^2(t) = e_y^2(t) - \overline{e_n^2(t)}. \tag{3.4}$$

ここで，$\overline{e_y^2(t)}$ は観測信号の平均パワーである．$m_N(SNR)$ は，変調周波数によらず，SNR から任意に求まる．この SNR によるパワーエンベロープ減算処理は，3.2 節の雑音残響に頑健な VAD において用いられている．SNR の推定法については，3.4.1 にて後述する．

3.3.3 パワーエンベロープ逆フィルタ処理

残響除去法としてのパワーエンベロープ逆フィルタ処理は，Unoki et al. [150] によって確立されている．残響の影響である残響の MTF を逆フィルタ処理することで，残響の影響を相殺するというもので，式 (2.17) から次式で表される．

$$\begin{aligned}\langle \mathbf{y}^2(t) \rangle &= \left\langle \left\{ \int_{-\infty}^{\infty} \mathbf{x}(\tau)\mathbf{h}(t-\tau)d\tau \right\}^2 \right\rangle \\ &= \int_0^t e_x^2(\tau)e_h^2(t-\tau)d\tau = e_y^2(t). \end{aligned} \tag{3.5}$$

式 (2.9) の $e_h^2(t)$ のパワーエンベロープは，z 変換により指数減衰の関数として表すことができ，$t < 0$ では $e_h^2(t) = 0$ となることから，最小位相特性を有している．そして，$e_h^2(n)$ の z 変換 $E_h(z)$ は，次式で表される．

$$E_h(z) = \frac{a^2}{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}}. \tag{3.6}$$

式 (3.5) の畳み込みの関係から，$e_x^2(t)$ の変調スペクトル $E_x(z)$ は，$e_y(t)$ の変調スペクトル $E_y(t)$ に対して $E_h(t)$ の逆特性を掛けることで求められる．したがって，$E_x(z)$ は次式により決定される．

$$E_x(z) = \frac{\hat{E}_y(z)}{a^2} \left\{ 1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1} \right\}, \tag{3.7}$$

ここで，f_s はサンプリング周波数である．回復パワーエンベロープ $\hat{e}_x^2(n)$ は，$\hat{E}_x(z)$ の逆 z 変換から求めることができる．式 (3.7) の 2 つのパラメータは，RIR の測定を必要とせずに，MTF の概念を用いて推定が可能であり，3.4.2 節で述べる．

3.3.4 帯域分割型パワーエンベロープ回復処理

音声信号は帯域分割して処理した方が効果的に回復処理を行えることから，定帯域幅フィルタバンクを用いて，各帯域でパワーエンベロープ回復を行う [144]．定

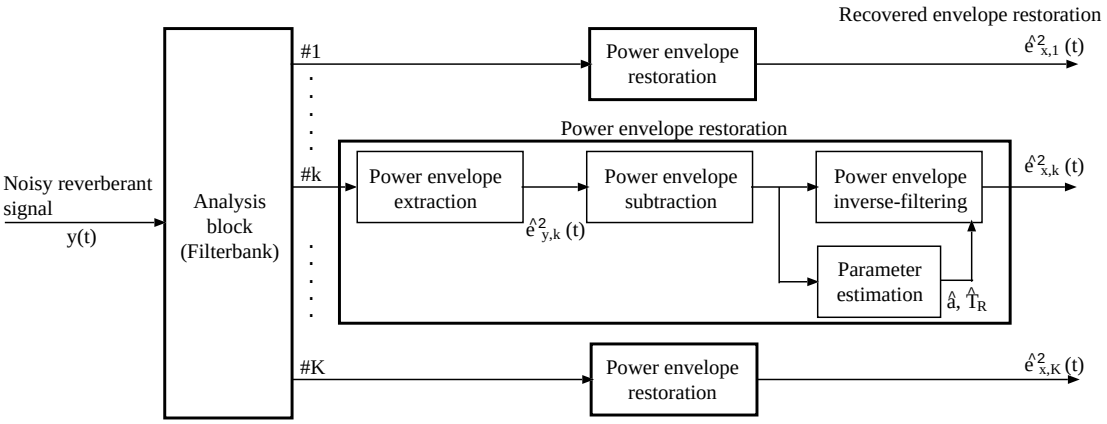


図 3.5: MTF に基づく帯域分割型パワーエンベロープ回復処理のブロックダイアグラム．

帯域幅フィルタバンクを用いて帯域分割を行うのは，帯域分割したパワーエンベロープ回復処理が ASR の前処理として利用できるためである [151]．

3.4 パラメータ推定

パラメータ推定は，雑音・残響の影響を知るのに雑音残響除去においては重要な技術である．本節では，MTF に基づくパワーエンベロープ回復処理のブラインド処理に必要不可欠な SNR 推定と残響時間推定について述べる．

3.4.1 SNR 推定

本節で扱う SNR は，MTF で利用する global SNR のことを指す．SNR 推定は，音声区間を用いて行うのが最もシンプルな方法であるため，頑健な VAD を必要とする．頑健な VAD のために SNR 推定が必要で，SNR 推定のために頑健な VAD が必要になる．そこで，提案した SNR 推定法では，多少の誤差は仕方がないものとして，帯域分割処理と最適なパワー閾値処理を組み合わせた上で，繰り返し処理を行うことで，SNR 推定を実現した．このアプローチは，帯域分割した処理の中で原信号（100 %）を仮定して，最適化したパワー閾値を用いることで誤差を低減するという考えである．帯域分割型パワーエンベロープ回復処理では，local SNR

を推定する方法が必要であると考えるが，現在はその提案にまでは至っておらず，ここでは，VAD を前提とした global SNR の推定法を紹介する．ここで紹介する global SNR の推定法は，帯域分割型 VAD 法の最適設計に基づく SNR 推定法 [152] を改良したものである．

まず，雑音音声の SNR 推定について考える．SNR は推定された音声区間 $\hat{S}_{VAD}$ かつ雑音が定常であれば，次式のように推定できる．

$$\hat{SNR} = 10\log 10\left(\frac{P_S(y(t),\hat{S}_{VAD}(t;e_y^2(t),\hat{B}))}{P_N(y(t),\hat{S}_{VAD}(t;e_y^2(t),\hat{B}))}\right) \quad [\mathrm{dB}] \tag{3.8}$$

$$\begin{aligned} & P_S(y(t),\hat{S}_{VAD}(t;e_y^2(t),\hat{B})) \\ = & \int_0^T \left(y^2(t)\hat{S}_{VAD}(t;e_y^2(t),\hat{B})\right)dt - P_N(\hat{S}_{VAD}(t;e_y^2(t),\hat{B})) \end{aligned} \tag{3.9}$$

$$P_N(\hat{S}_{VAD}(t;e_y^2(t),\hat{B})) = \int_0^T \left(\overline{P_N(\hat{S}_{VAD}(t;e_y^2(t),\hat{B}))}\hat{S}_{VAD}(t;e_y^2(t),B)\right)dt \tag{3.10}$$

$$\overline{P_N(\hat{S}_{VAD}(t;e_y^2(t),\hat{B}))} = \frac{\int_0^T \left(y^2(t)\left(1-\hat{S}_{VAD}(t;e_y^2(t),\hat{B})\right)\right)dt}{\int_0^T \left(1-\hat{S}_{VAD}(t;e_y^2(t),\hat{B})\right)dt} \tag{3.11}$$

ここで，$\overline{P_N}$ は雑音の平均パワー，P_N と P_S は雑音と音声のパワーである．実際には，音声区間は未知であり，音声区間を検出する必要がある．そこで，VAD の時と同様にパワー閾値を最適化することでこの問題を解決できると考え，SNR 推定においては多くの雑音残響信号のパワーエンベロープを用いてパワー閾値 $\hat{B}$ を最適化することを考えた．最適化の手順は VAD の時と同様であるが，回復パワーエンベロープ $e_y^2(t)$ なのか雑音残響信号のパワーエンベロープ $\hat{e}_x^2(t)$ なのかが異なる．本研究では，雑音信号のみでの最適化かつ帯域分割処理を用いて実現している．

global SNR を推定するにあたり音声成分と雑音成分を判別する必要があり，この判別に VAD がよく用いられているが，雑音残響の影響により VAD の性能が低下し，SNR の推定性能も低下する．音声に対する雑音の影響は，周波数帯域ごとに異なっていることから多くの音声信号処理では，帯域分割処理が用いられている．提案する SNR 推定法では，帯域分割処理を用いて帯域信号を求める．各帯域に異なった閾値を設定してパワー閾値による VAD を行うことで音声/非音声区間が得られ，帯域ごとに音声パワーと雑音パワーを求め，最終的に全帯域の音声パワーと雑音パワーを合算することで global SNR を推定できる．

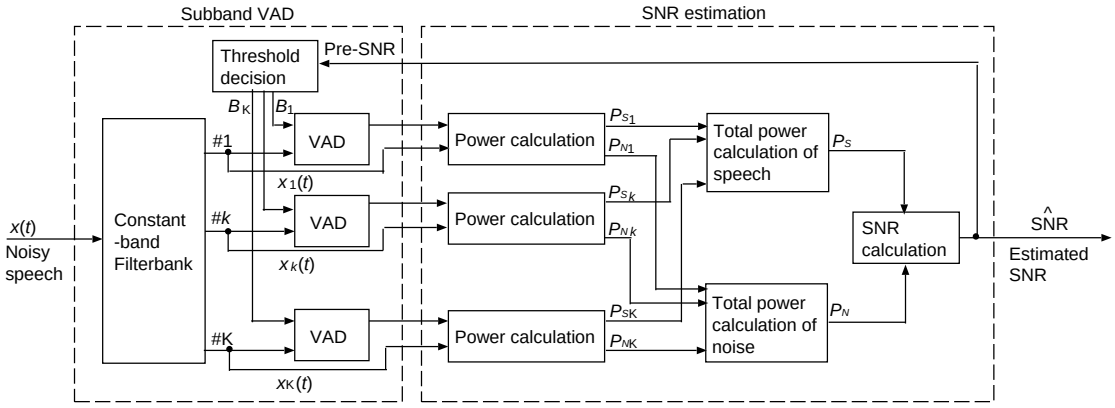


図 3.6: grobal SNR 推定法のブロックダイアグラム．

本 SNR 推定法のブロックダイアグラムを図 3.6 に示す．図中の k は帯域番号，K は帯域分割数を意味する．図 3.6 に示す通り，本 SNR 推定法は大きく分けて二つのブロック（二つの破線のブロック）で構成されている．図中の前半は帯域分割信号に対する VAD であり，図中の後半は各帯域における雑音パワーと音声パワーの推定と最終的な SNR の計算である．フィードバック処理では，推定された各帯域の local SNR を返し，各帯域のパワー閾値を再設定し，繰り返し処理を行う．そして，フィードバック処理を繰り返した後，最終的な global SNR が求まる．各帯域の閾値は，雑音音声を帯域分割し，ROC 曲線上での誤受理率と誤棄却率のトレードオフの関係を最適化し，最適な閾値と local SNR の関係をまとめることで求めた．

帯域分割処理には，定 Q フィルタバンク（CQFB: constant-Q filterbank）や定帯域幅フィルタバンク（CBFB: constant-bandwidth filterbank）がよく利用されている．本研究では，定帯域分割処理を利用して ASR を行っていることから，帯域分割処理に帯域幅 100 Hz 固定の CBFB を利用した．サンプリング周波数が 20 kHz の場合，帯域分割数 K は 100 帯域となる．

帯域分割信号は，帯域ごとに SNR が異なるため，帯域ごとに異なった閾値を設定して音声/非音声を判別する必要がある．2.1 節で述べた FAR と FRR を各帯域分割信号で置き換えると，FAR(B_k) と FRR(B_k) と表現され，B_k は k 番目の帯域の閾値を意味する．多くの VAD では，ROC 曲線上のある一点を決めることで，VAD の目的に合わせて性能を調整している．様々な条件の雑音音声を用いて学習することで求めた ROC 曲線上の最適な閾値を決定する．

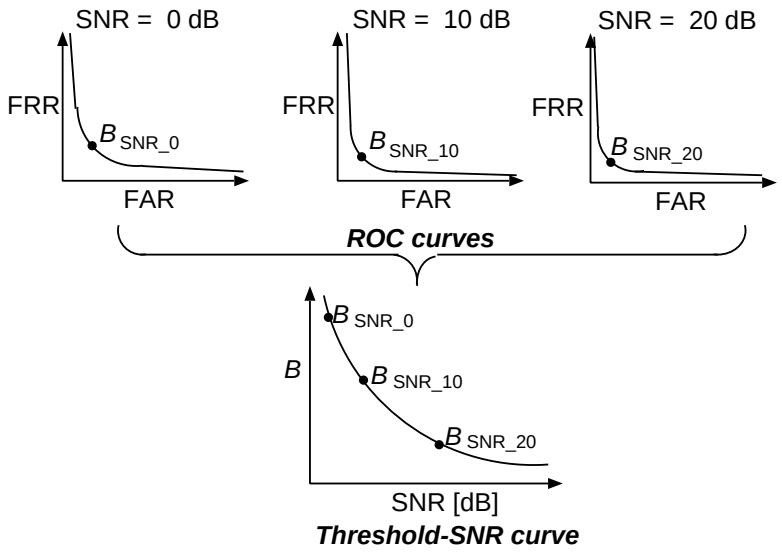


図 3.7: 各 SNR 条件下での VAD 閾値決定法.

k 番目の帯域及び SNR 条件ごとに，FAR(B_k) と FRR(B_k) から ROC 曲線を求めて，FAR と FRR の二乗平均平方根（RMS）を次式のように求めた.

$$\mathrm{RMS}(B_k)=\sqrt{\frac{\mathrm{FAR}^2(B_k)+\mathrm{FRR}^2(B_k)}{2}}, \tag{3.12}$$

また，VAD の性能を最大限発揮できる最適な閾値を次式を用いて求めた.

$$B_k^*=\arg\min_{B_k} \mathrm{RMS}(B_k). \tag{3.13}$$

式 (3.12) のように FAR と FRR の RMS を求める方法は，音声/非音声の検出性能を容易に予測できることから，VAD の性能を評価して最適な閾値を決定するのに有効な方法である.

様々な SNR 条件における雑音音声のデータベースを用いて学習を行うことで，SNR 条件ごと，帯域ごとの最適な閾値を求めた．すべての SNR 条件において学習で閾値を最適化することは困難である．そのため，図 3.7 に示すように，実際にはいくつかの SNR 条件において学習を行い，SNR 条件ごとに ROC 曲線を求めて最適な閾値を決定し，各 SNR 条件の最適な閾値を布置してシグモイド関数で近似することで Threshold-local SNR 曲線を描画した．そして，フィードバック処理で pre-local SNR を返し Threshold-local SNR 曲線を用いて各帯域の閾値 B_k を再設定することで，推定誤差の低減を図った.

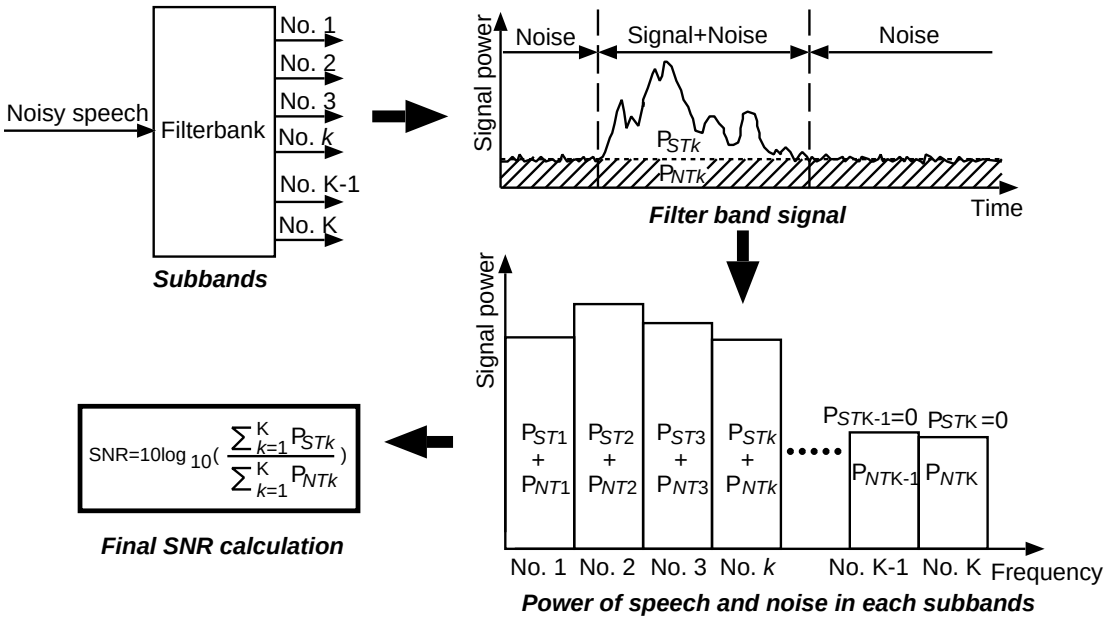


図 3.8: グローバル SNR 推定のパワー計算の流れ．

パーセバルの定理に基づき時間信号のパワーは，各帯域のパワーの合算として求めることができる．SNR 推定の流れを図 3.8 に示す．帯域ごとに音声に対する雑音の影響が異なっていることを考慮して，VAD による音声/非音声の判別とパワー計算は帯域ごとに行われた．そして，最終的な SNR はパーセバルの定理に基づいて全帯域の音声パワーと雑音のパワーを合算し比を取ることで，次式のように求めることができる．

$$\hat{\mathrm{SNR}} = 10\log_{10}\left(\frac{\sum_{k=1}^{\mathrm{K}} P_{ST_k}}{\sum_{k=1}^{\mathrm{K}} P_{NT_k}}\right), \tag{3.14}$$

$$P_{NT_k} = \int_0^T P_{N_k}(t)\, H_{N_k}(t)\, dt, \tag{3.15}$$

$$P_{ST_k} = \int_0^T P_{S_k}(t)\, H_{S_k}(t)\, dt - \int_0^T \overline{P_{NT_k}} H_{S_k}(t)\, dt, \tag{3.16}$$

ここで，P_{ST_k} は k 番目の帯域の音声パワーの合計，P_{NT_k} は k 番目の帯域の雑音パワーの合計である．$\overline{P_{NT_k}}$ は k 番目の帯域の非音声区間の平均雑音パワーである．

このコンセプトは，図 3.6 の二つ目のブロックを示している．k 番目の帯域の帯域分割雑音音声信号に対して，ある区間が非音声として判別された場合，雑音パ

ワーはこの区間の帯域分割信号の合計として計算される．そして，残りの区間が音声として分類された場合，その区間の帯域分割信号には，音声と雑音が存在している．本研究では，この音声と雑音が存在する区間の平均雑音パワーが，非音声区間の平均雑音パワーと等しいと仮定して，音声と雑音が存在する区間の帯域分割信号の合計パワーから，この区間の長さの平均雑音パワーの合計を減算することで，音声と雑音が存在する区間の音声のパワーの合計を求めることができる．各帯域でこの計算を行い，全帯域の音声パワーと雑音パワーの比を取ることで，最終的にSNRが求まる．

3.4.2 残響時間推定

残響信号の残響時間推定については，本研究で考える変調度1を規範とした処理という概念を用いている，Unoki et al.が提案したMTFの概念に基づく残響時間推定法[37]を利用する．残響の影響を受ける前の信号のパワーエンベロープは，少なくとも一つのゼロになる点（谷）を無音区間に有していることから，変調度は1であると言える．しかし，残響の影響を受けた信号のパワーエンベロープは，変調度が1以下であるため，谷がゼロではなく正の値となる．そこで，この谷が逆フィルタ処理によって負もしくはゼロになる時の変調度を求めることで，パワーエンベロープ回復に最適な残響時間T_RをMTFから推定できるという方法である．ただし，Unoki et al.の手法では雑音の影響は考慮されておらず，本研究の中で雑音残響信号に対応する残響時間推定法へと拡張している．

残響環境におけるT_Rのパラメータ推定について考える．残響音声のパワーエンベロープは次式で回復できる．

$$\hat{e}_x^2(t) \quad = \quad e_y^2(t) \otimes e_h^2(t), \\ . \tag{3.17}$$

ここで，$\otimes$は，逆畳み込みの演算子である．変調度1の原信号のパワーエンベロープ$e_x^2(t)$の音声区間での最小値を0とした時，残響信号のパワーエンベロープ$e_{yh}^2(t)$の音声区間での最小値が0以上になっていることを利用し，T_Rで最適化したときに任意のパワーエンベロープの負の面積が最小になるようにT_Rを推定した[37]．残響環境での推定式は次式の通りである．

$$\hat{T}_R = \underset{0\leq T_R\leq T_{R,\max}}{\arg\min} \left\{ \frac{dT_P(T_R)}{dT_R} \right\}, \tag{3.18}$$

$$T_P(T_R) = \min \left(\underset{t_{\min}\leq t\leq t_{\max}}{\arg\min} \left| \hat{e}_{x,n,T_R}(t)^2 - \theta \right| \right), \tag{3.19}$$

ここで，$\hat{e}_{x,n,T_R}$ は帯域 n と残響時間 T_R を変数として求めたパワーエンベロープ，θ は $e_y^2(t)$ の最大値から求めた閾値である．ここで，θ は，$t_{\min}$ 〜 $t_{\max}$ の時間範囲内における $e_y^2(t)$ の最大値に 0.01 をかけた値（-20 dB 低下した値）とした．$t_{\min}$ と $t_{\max}$ は，閾値決定のための下限と上限時間である．

パワーエンベロープ逆フィルタ処理で必要となる振幅項 a の推定法 [37] について述べる．室内における残響の影響は，ゲインの増幅よりも反射による遅延の重ね合わせ効果の方が大きい．インパルス応答のゲインは，インパルス応答のパワーの合算として近似される．したがって，a の値は $e_h^2(t)$ の総和から，推定した残響時間を用いた次式で決定される．

$$\hat{a} = \sqrt{1/\int_0^T \exp(-13.8t/\hat{T}_R)dt}. \tag{3.20}$$

また，式 (3.20) は残響のみの影響を考慮したものであり，雑音残響環境においては次式のように，雑音の影響を除去する必要がある．

$$\hat{T}_R = \underset{0\leq T_R\leq T_{R,\max}}{\arg\min} \left\{ \frac{dT_P(T_R)}{dT_R} \right\}, \tag{3.21}$$

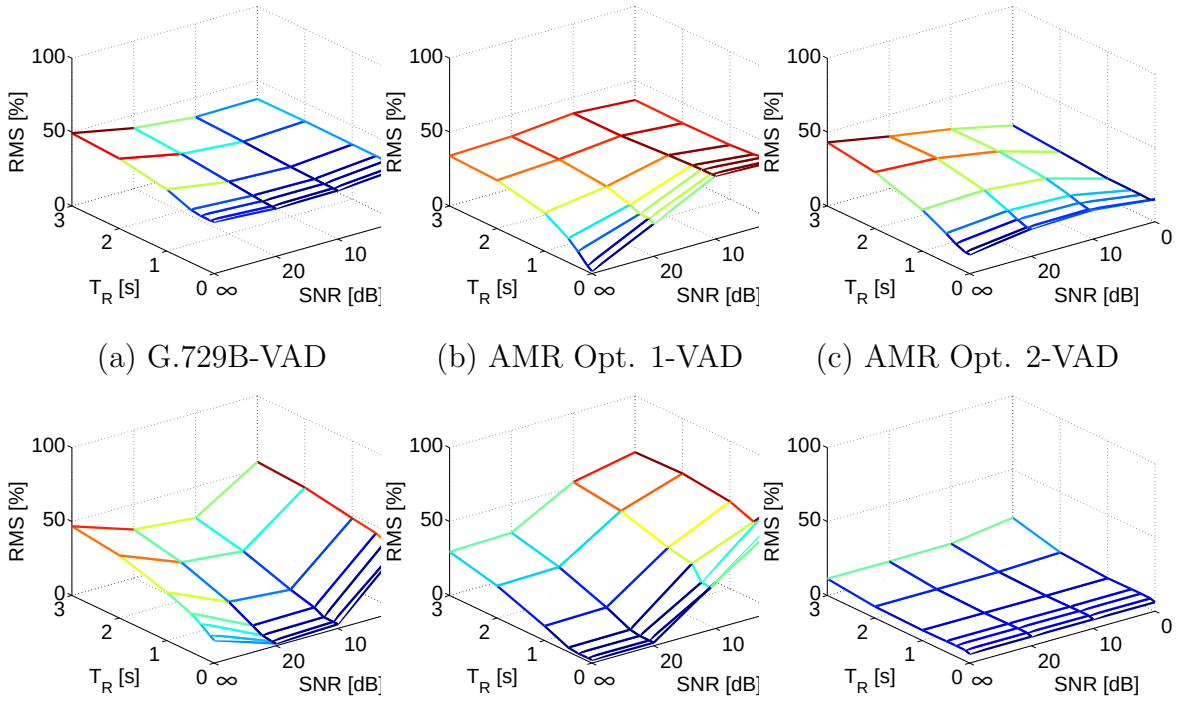
$$T_P(T_R) = \min \left(\underset{t_{\min}\leq t\leq t_{\max}}{\arg\min} \left| \left(\hat{e}_{x,T_R}(t)^2 - \overline{e_n^2} \right) \theta \right| \right). \tag{3.22}$$

ここで，θ は $\hat{e}_{x,T_R}(t)^2 - \overline{e_n^2}$ から求めた閾値である．実際には，本論文では雑音パワー減算処理の後に T_R を推定しており，実質的に上記と同様の処理を行っている．

3.5 性能評価

3.5.1 音声区間検出

ここでは，音声区間検出性能の比較評価を行う．まず，人工的な雑音残響環境において，雑音残響に頑健な VAD 法を含む 4 つの VAD 法の音声・非音声検出



(a) G.729B-VAD (b) AMR Opt. 1-VAD (c) AMR Opt. 2-VAD

(d) Otsu-VAD (e) IMTFRvb-VAD (f) Proposed method

図 3.9: 人工的な雑音残響環境における VAD の検出結果.

の性能評価を行った．比較手法として提案法の他に G.729B-VAD 法 [96]，AMR Opt. 1-VAD 法 [101]，AMR Opt. 2-VAD 法 [101]，音声/非音声判別に Otsu の二値化を用いた VAD 法（Otsu-VAD），従来法である残響除去のみを前処理とした IMTFRvb-VAD 法 [136] を用いた．評価条件は，AURORA-2J のテストデータ 1001 音声（学習データとは異なるオープンデータ）を利用し，雑音には白色ガウス雑音，RIR には Schroeder の RIR [142] を用いた．評価基準には，次式の FRR (%) と FAR (%) の二乗平均平方根（RMS）を用いた．RMS の値が低いときには検出性能が高いことを，RMS の値が低いときには検出性能が低いことを示す．

$$\mathrm{RMS} = \sqrt{\frac{\mathrm{FRR}^2 + \mathrm{FAR}^2}{2}} \tag{3.23}$$

図 3.9 に評価の結果を示す．G.729B-VAD 法と AMR Opt. 2-VAD 法の検出性能は，同様の傾向を示しており，背景雑音と残響の影響により低下していることがわかる．また，残響時間が長くなるにつれ検出性能が低下することも確認された．

他には，SNR = 0 dB かつ $T_R = 3$ s の条件では，SNR = 20 dB かつ $T_R = 3$ s の条件より検出性能が向上しており，残響の影響より雑音の影響が大きくなった場合に検出性能が向上していることがわかる．一方，IMTFRvb-VAD 法は，残響のみの条件（SNR = ∞）では検出性能が高い．しかし，SNR が低くなるにつれて検出性能が著しく低下しており，SNR = 0 dB では検出が困難であることがわかる．提案法は，全ての条件下において他の手法と比較して最も性能が優れており，T_R = 2 s までの RMS が非常に小さいことがわかる．これらの結果より，人工的な雑音残響に対して提案法が最も頑健であることがわかる．しかし，$T_R = 3$ s の雑音残響条件においては，提案法を含めたすべての手法で検出性能がほぼ同じであり，あまり検出性能が高くないことがわかった．残響時間 $T_R = 3$ s 以上の雑音残響環境に対する頑健性は，今後の課題である．

実環境を想定した雑音残響環境での評価として，SMILE2004 [153, 154] に収録されている実環境の室内で収録された 43 個の RIR $h(t)$ と NOISEX-92 [155] に収録されている白色雑音，ピンク雑音，バブル雑音，工場雑音を SNR = 20, 10, 0 dB で利用した．音声信号は，人工的な雑音残響での評価条件と同じであり，雑音残響音声は同様に式 (2.11) によって求めた．評価結果を図 3.12, 図 3.11, 図 3.12, 図 3.13 に示す．

実環境を想定した雑音残響環境での評価結果より，提案法は他の手法に比べて RMS が非常に小さく，優れた性能である．残響時間が長くなることによって，検出性能が若干低下するものの，他の手法に比べて検出性能が非常に高い．バブル雑音と工場雑音の SNR = 0 dB の条件下においては，音声区間の検出性能が著しく低下するものの，他の手法に比べると若干検出性能が高い条件もある．これらの雑音は白色雑音やピンク雑音に比べて非定常性が高いため，非音声区間の雑音のパワーがパワー閾値を上回るために生じる問題である．

SNR や残響時間の影響による検出性能の低下は，提案法が音声信号のパワーエンベロープのみで音声/非音声判別を行っていることにより，子音や無声音を音声区間として検出することが困難であるためと考えられる．雑音・残響環境においてパワーの小さい子音や無声音を検出することは，非常に難しい問題である．雑音残響環境においてこの問題を解決するためには，頑健な有声/無声判別やイベント検出など，他の音響特徴を頑健に推定して複合的に利用する必要があると考える．

定常的な雑音と音声による雑音音声のみを仮定している本 VAD 法は，パワー閾値に頼った手法であることから，突発性雑音などには対応できていない．本 VAD 法の発展を考えると，パワー閾値によって検出された音声/非音声区間が，音声の特徴を含んでいるのかどうかを変調スペクトルを用いて判別 [104] するアプローチが有効であると考える．

3.5.2 パワーエンベロープの回復

ここでは，帯域分割型パワーエンベロープ回復処理の性能評価を行う．まず，人工的な雑音残響環境下において，提案法のパワーエンベロープ回復性能の評価を行った．音声信号 $x(t)$ として，AURORA-2J 音声データベース [146] のテスト用の 1001 個のクリーン音声を利用した．音残響音声は，式 (2.11) を利用して求めた．RIR $h(t)$ には，残響時間 0.3, 0.5, 1.0, 1.5, 2.0 s の合計 5 種類の Schroeder の RIR を利用した．背景雑音 $n(t)$ には，SNR 20, 10, 0 dB の 3 種類の白色ガウス雑音を利用した．そのため，評価に用いた残響信号の刺激数は 5,005（$= 1001 \times 5$），雑音信号の刺激数は 3,003（$= 1001 \times 3$），雑音残響信号の刺激数は 15,015（$= 1001 \times 5 \times 3$）であった．サンプリング周波数 f_s は 8 kHz，40 チャンネル定帯域フィルタバンク（100 Hz 帯域幅）を利用した．

パワーエンベロープの誤差と類似性（振幅と形状）を評価するために，評価尺度は，パワーエンベロープに対する信号対復元誤差比（SER：S をオリジナル，E をオリジナルと雑音残響または回復したものとの差）と相関値（Corr）を利用した．

$$\mathrm{SER}(e_x^2, \hat{e}_x^2) = 10 \log_{10} \frac{\int_0^T (e_x^2(t))^2 dt}{\int_0^T (e_x^2(t) - \hat{e}_x^2(t))^2 dt} \tag{3.24}$$

$$\begin{aligned} &\mathrm{Corr}(e_x^2, \hat{e}_x^2) \\ &= \frac{\int_o^T \left(e_x^2(t) - \overline{e_x^2(t)}\right)\left(\hat{e}_x^2(t) - \overline{\hat{e}_x^2(t)}\right) dt}{\sqrt{\left\{\int_0^T \left(e_x^2(t) - \overline{e_x^2(t)}\right)^2 dt\right\}\left\{\int_0^T \left(\hat{e}_x^2(t) - \overline{\hat{e}_x^2(t)}\right)^2 dt\right\}}} \end{aligned} \tag{3.25}$$

ただし，$\overline{e_x^2(t)}$ は，$e_x^2(t)$ の平均値を示す．そして，改善量を解り易くするために，$\mathrm{SER}(e_x^2, \hat{e}_x^2)$ と $\mathrm{Corr}(e_x^2, \hat{e}_x^2)$ の改善量を，それぞれ $\mathrm{SER}(e_x^2, \hat{e}_x^2) - \mathrm{SER}(e_x^2, e_y^2)$ と $\mathrm{Corr}(e_x^2, \hat{e}_x^2) - \mathrm{Corr}(e_x^2, e_y^2)$ と定義して，評価に用いた．そのため，これらの評価尺度の改善量（Imp SER と Imp Corr）は，改善すると正の値になり，改悪すると負の値になる．

各条件下での SER と Corr の改善量の平均値を，図 3.14 と図 3.15 に示す．結果より，残響環境（SNR= ∞ dB）では，残響時間が長くなるにつれて ICorr が大きくなることがわかる．残響の影響は残響時間が長くなるにつれて大きくなるため，改善量も適切に逆フィルタ処理が機能することで残響時間が長くなるにつれて大きくなる．得られた結果より，適切に逆フィルタ処理が機能していることがわかる．雑音環境（$T_R = 0$ s）では，SNR が低くなるにつれて ISER が大きくなることがわかる．これは，SNR が低くなるにつれて雑音のパワーが増加するため，適切に雑音除去が行われると改善量が大きくなることで，ISER も大きくなる．したがって，雑音・残響環境において，うまく機能していることがわかる．雑音残響環境では，雑音除去と残響除去がうまく機能することで，雑音と残響の影響が大きくなることで ISER と ICorr が大きくなることがわかる．これらの結果より提案法は，雑音除去と残響除去をうまく統合できており，雑音と残響の影響を同時に除去できることを示した．

実環境を想定した雑音残響環境下で提案法のパワーエンベロープ回復の評価シミュレーションを行った．音声信号 $x(t)$ には，AURORA-2J 音声データベース [146] のテスト用の 1001 個のクリーン音声を利用した．室内インパルス応答 $h(t)$ には SMILE2004 [153] の実環境で集音された 8 個の RIR を利用し，背景雑音 $n(t)$ には NOISEX-92 [155] の白色雑音，ピンク雑音，バブル雑音，工場雑音 2 種類を利用した．ここでは，雑音残響に頑健な VAD は用いず，音声区間を既知とした時の結果を示している．評価尺度には，人工的な雑音残響環境における評価と同様に，SER と Corr の改善量である ISER と ICorr の平均値を利用した．

表 3.1 と表 3.2 に，各条件での評価結果を示す．結果より，ISER はほとんどの条件において正の値を取り，提案法による改善が確認された．また，ISER は雑音の影響が大きくなるにつれて大きくなっており，雑音除去がうまく機能していることがわかる．一方，ICorr は，雑音の影響が大きい環境では改善が見られない場合や改悪になる場合もあったが，ほとんどの条件で提案法による改善が確認された．雑音の影響が小さい場合には，ICorr が大きいことから，SNR が低いときにパワーエンベロープ逆フィルタ処理の効果が大きいことがわかった．

表 3.1: 実環境を想定した雑音残響環境での ISER，IRdata No. は SMILE2004 [153] のファイル番号である．

Room condition (Impulse response)	IRdata	T_R	ISER (dB)														
	No.	(s)	white			pink			babble			factory1			factory2		
SNR (dB)			20	10	0	20	10	0	20	10	0	20	10	0	20	10	0
Noisy environments		0.00	3.0	7.6	10.9	1.2	6.0	10.3	-7.2	-1.3	4.7	-0.1	4.2	7.6	-8.5	-1.9	-1.9
Living room (wooden)(capacity: 110 m^3)	411	0.36	0.9	3.7	9.1	0.4	2.5	7.8	-0.1	0.9	4.4	0.2	1.9	6.0	0.2	0.8	3.1
Church1 (capacity: 1,200 m^3)	405	0.71	1.1	3.1	7.8	0.6	1.8	6.9	0.3	1.0	4.2	0.4	1.3	5.6	0.5	0.9	2.8
MPH1 (with RB)(capacity: 2,000 m^3)	301	1.09	1.5	3.2	6.8	1.1	2.0	6.0	0.8	1.4	4.1	0.9	1.4	5.0	1.1	1.3	2.6
GSH (capacity: 11,000 m^3)	404	1.54	0.8	2.9	6.4	0.3	1.5	5.6	-0.1	0.8	3.9	0.1	0.9	4.8	0.2	0.5	2.2
MPH3 (with RB)(capacity: 7,200 m^3)	305	1.93	1.0	2.8	6.2	0.6	1.5	4.9	0.2	0.9	3.7	0.4	1.0	3.8	0.5	0.8	1.9
CCH1 (capacity: 5,600 m^3)	309	2.35	1.2	2.9	6.0	0.7	1.7	4.8	0.4	1.1	3.9	0.5	1.1	4.3	0.6	0.9	2.1
Event hall1 (capacity: 28,000 m^3)	407	3.03	0.7	2.5	5.8	0.2	1.2	4.2	-0.1	0.7	3.5	0.0	0.7	3.0	0.2	0.5	1.4
Event hall2 (capacity: 41,000 m^3)	408	3.62	0.7	2.5	5.7	0.2	1.2	4.2	-0.1	0.7	3.5	0.0	0.7	3.0	0.1	0.4	1.5

3.5.3 パラメータ推定

ここでは，SNR 推定と残響時間推定の性能評価を行う．

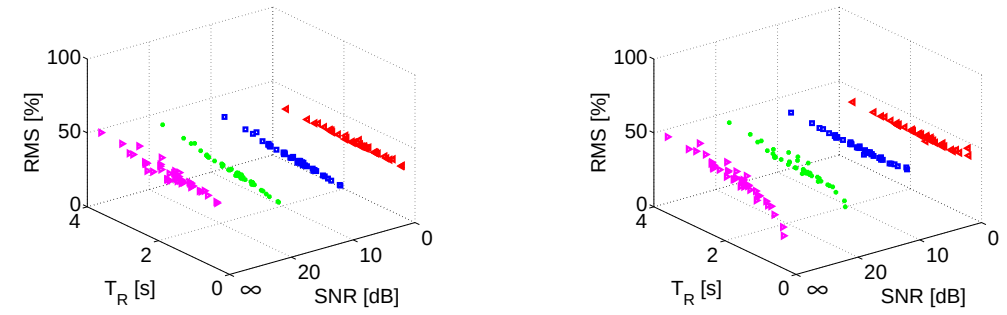
はじめに，雑音環境における本 SNR 推定法の評価を行う．音声は AURORA-2J [146] からオープンデータとして学習データとは異なるテストデータ 1001 発話，雑音には SNR は 20, 10, 0 dB の白色ガウス雑音を利用した．本 SNR 推定法の繰り返し処理の回数は，1 回に設定した．図 3.16 に本 NSR 推定法の評価結果を示す．

本 SNR 推定法は，SNR $= 0$ dB 以上の条件において，誤差約 1 dB 以下で高精度に SNR を推定できることがわかる．SNR $= 0$ dB においても推定精度が高いのは，音声パワーの大きい帯域において local SNR が高くなることで音声区間検出性能がよく，音声/非音声判別の精度が向上しているためと考えられる．このように，帯域分割処理と VAD，雑音レベルに合わせた閾値設定を行うことにより，SNR を精度よく推定できることを示した．

次に雑音残響環境における本 SNR 推定法の評価を行う．音声と雑音は，雑音環境と同じデータを用いた．RIR は，残響時間 $T_R = 0.1, 0.3, 0.5, 1.0, 1.5, 2.0$ s の Schroeder の RIR [142] を利用した．この時の SNR の本 SNR 推定法の評価結果を図 3.17-図 3.19 に示す．

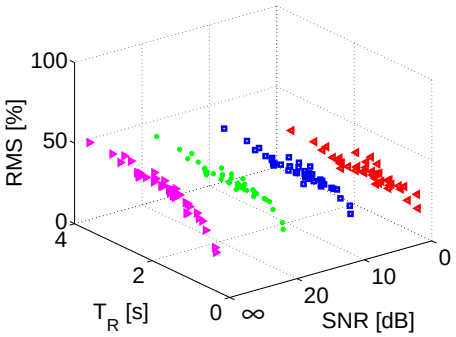
表 3.2: 実環境を想定した雑音残響環境での ICorr，IRdata No. は SMILE2004 [153] のファイル番号である．

Room condition (Impulse response)	IRdata	T_R	ICorr														
	No.	(s)	white			pink			babble			factory1			ffactory2		
SNR (dB)			20	10	0	20	10	0	20	10	0	20	10	0	20	10	0
Noisy environments		0.00	0.00	0.01	0.01	0.00	0.01	0.02	-0.04	-0.05	-0.05	0.00	0.00	0.00	-0.04	-0.04	-0.04
Living room (wooden)(capacity: 110 m^3)	411	0.36	0.00	-0.01	0.00	0.01	0.00	0.01	0.01	0.00	-0.03	0.00	0.00	0.00	0.01	0.00	0.00
Church1 (capacity: 1,200 m^3)	405	0.71	0.07	0.03	0.00	0.09	0.05	0.01	0.09	0.05	-0.02	0.08	0.03	0.00	0.10	0.08	0.02
MPH1 (with RB)(capacity: 2,000 m^3)	301	1.09	0.14	0.08	0.02	0.16	0.12	0.03	0.16	0.12	0.01	0.16	0.10	0.00	0.17	0.15	0.08
GSH (capacity: 11,000 m^3)	404	1.54	0.01	-0.01	-0.02	0.02	0.00	-0.01	0.02	0.61	-0.03	0.02	0.00	-0.01	0.03	0.02	0.00
MPH3 (with RB)(capacity: 7,200 m^3)	305	1.93	0.08	0.04	0.01	0.09	0.06	0.02	0.10	0.06	-0.01	0.09	0.05	0.00	0.11	0.09	0.05
CCH1 (capacity: 5,600 m^3)	309	2.35	0.12	0.08	0.03	0.14	0.11	0.05	0.14	0.10	0.02	0.14	0.10	0.02	0.15	0.14	0.09
Event hall1 (capacity: 28,000 m^3)	407	3.03	0.03	0.00	0.01	0.04	0.03	0.02	0.03	0.03	0.01	0.04	0.03	0.02	0.04	0.04	0.04
Event hall2 (capacity: 41,000 m^3)	408	3.62	0.03	0.02	0.01	0.03	0.03	0.02	0.03	0.02	0.01	0.04	0.03	0.02	0.04	0.04	0.04

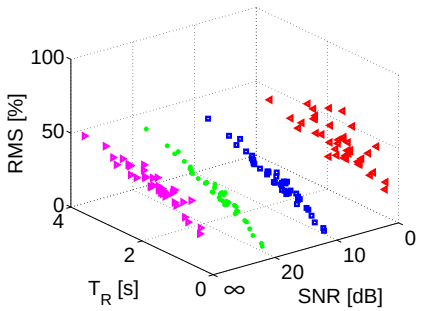


(a) G.729B-VAD

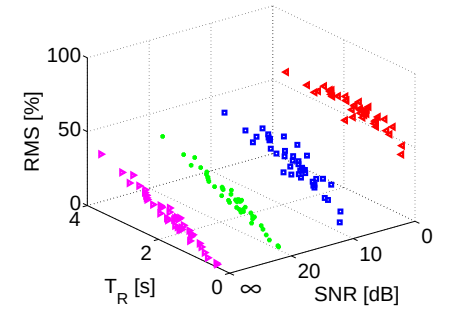
(b) AMR Opt. 1-VAD



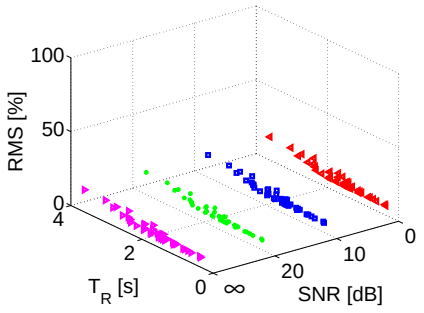
(c) AMR Opt. 2-VAD



(d) Otsu-VAD

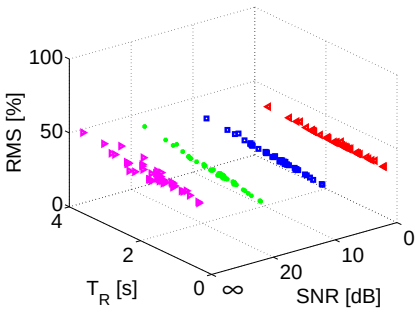


(e) IMTFRvb-VAD

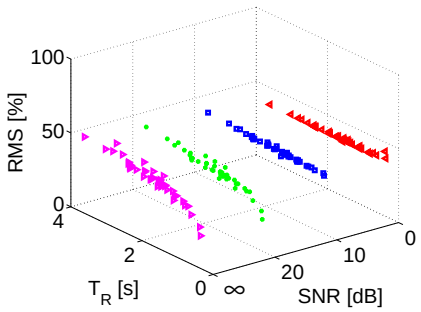


(f) Proposed method

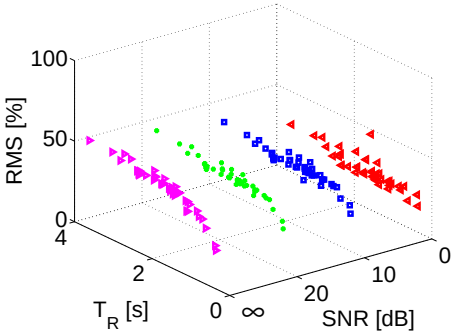
図 3.10: 実環境を想定した雑音残響環境における VAD の検出結果（白色雑音）: マジェンタ ▷ : SNR $= \infty$ dB (雑音なし), 緑 ● : SNR $= 20$ dB, 青の白抜きの三角 : SNR $= 10$ dB, 赤 ◁ : SNR $= 0$ dB.



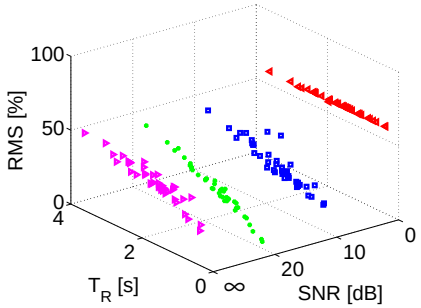
(a) G.729B-VAD



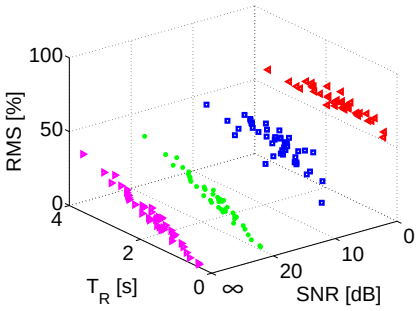
(b) AMR Opt. 1-VAD



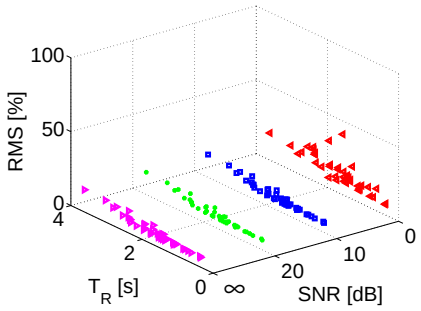
(c) AMR Opt. 2-VAD



(d) Otsu-VAD

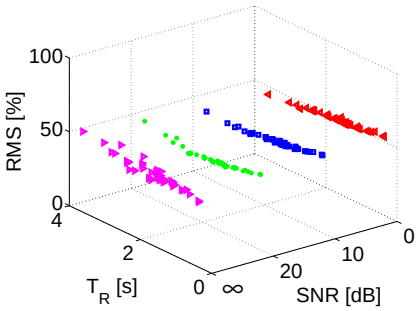


(e) IMTFRvb-VAD

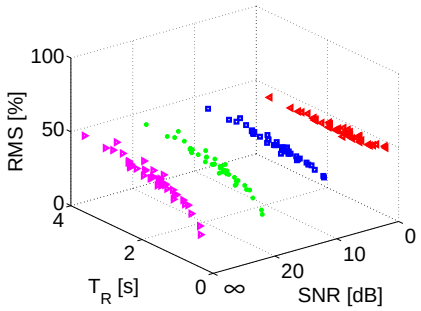


(f) Proposed method

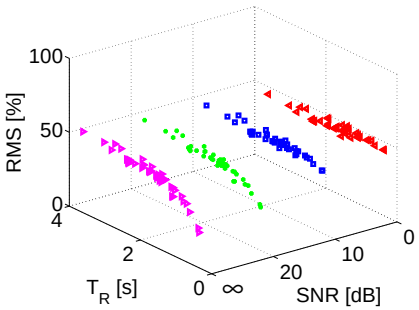
図 3.11: 実環境を想定した雑音残響環境における VAD の検出結果（ピンク雑音）: マジェンタ ▷ : SNR = ∞ dB (雑音なし), 緑 ● : SNR = 20 dB, 青の白抜きの三角 : SNR = 10 dB, 赤 ◁ : SNR = 0 dB.



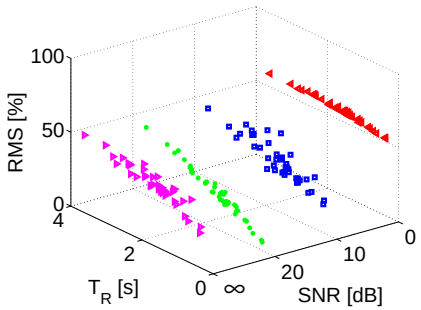
(a) G.729B-VAD



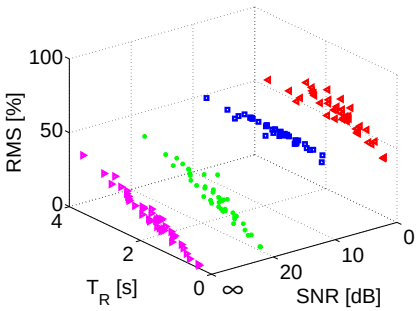
(b) AMR Opt. 1-VAD



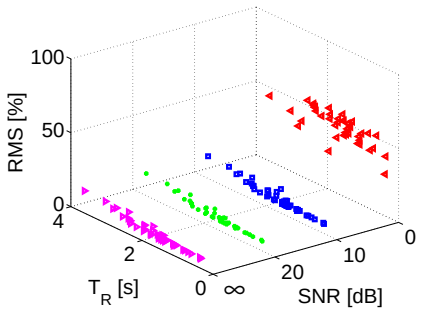
(c) AMR Opt. 2-VAD



(d) Otsu-VAD

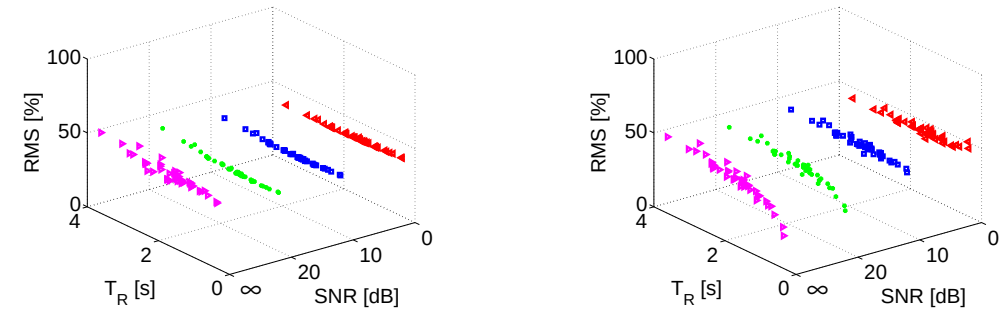


(e) IMTFRvb-VAD



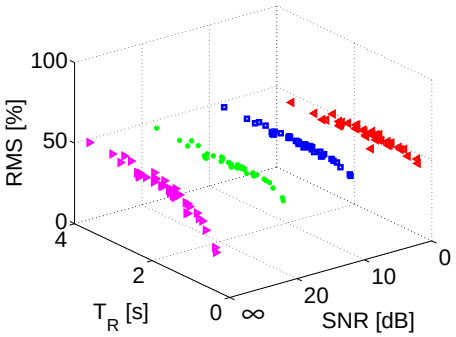
(f) Proposed method

図 3.12: 実環境を想定した雑音残響環境における VAD の検出結果（バブル雑音）: マジェンタ ▷ : SNR = ∞ dB (雑音なし), 緑 ● : SNR = 20 dB, 青の白抜きの三角 : SNR = 10 dB, 赤 ◁ : SNR = 0 dB.

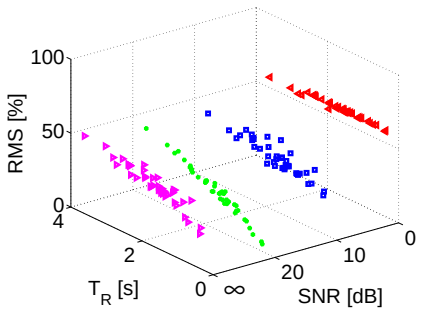


(a) G.729B-VAD

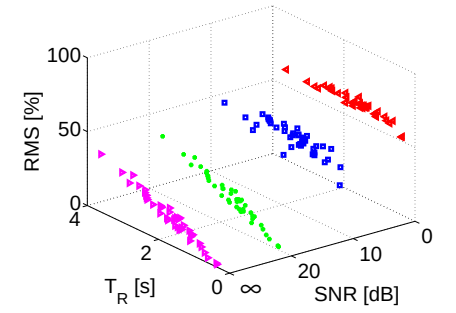
(b) AMR Opt. 1-VAD



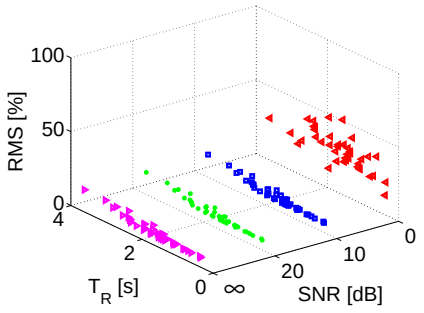
(c) AMR Opt. 2-VAD



(d) Otsu-VAD



(e) IMTFRvb-VAD



(f) Proposed method

図 3.13: 実環境を想定した雑音残響環境における VAD の検出結果（工場雑音）: マジェンタ ▷ : SNR = ∞ dB (雑音なし), 緑 ● : SNR = 20 dB, 青の白抜きの三角 : SNR = 10 dB, 赤 ◁ : SNR = 0 dB.

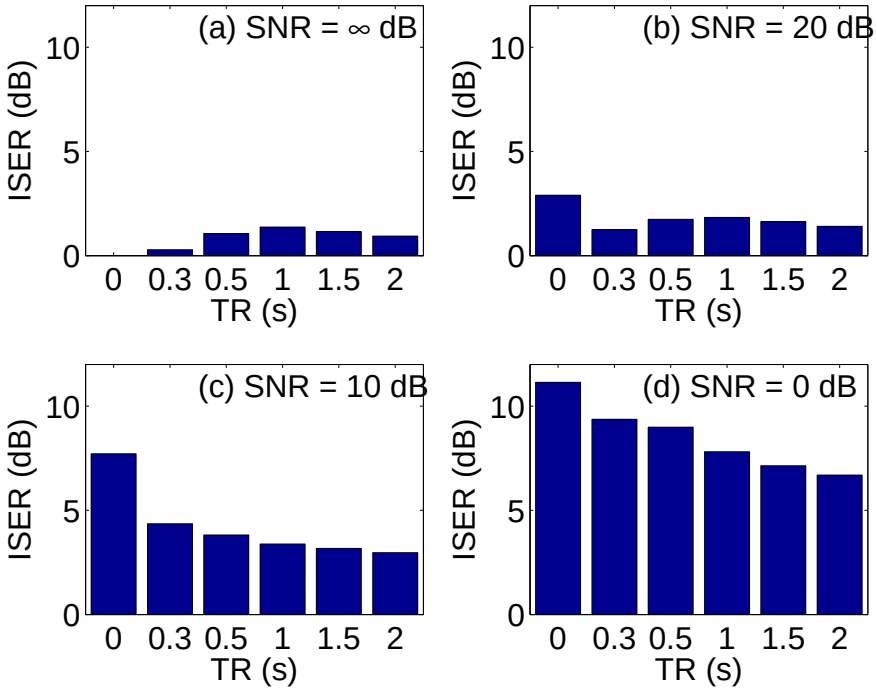


図 3.14: 人工的な雑音・残響環境での ISER：(a) SNR $= \infty$ dB, (b) SNR $= 20$ dB, (c) SNR $= 10$ dB, (d) SNR $= 0$ dB.

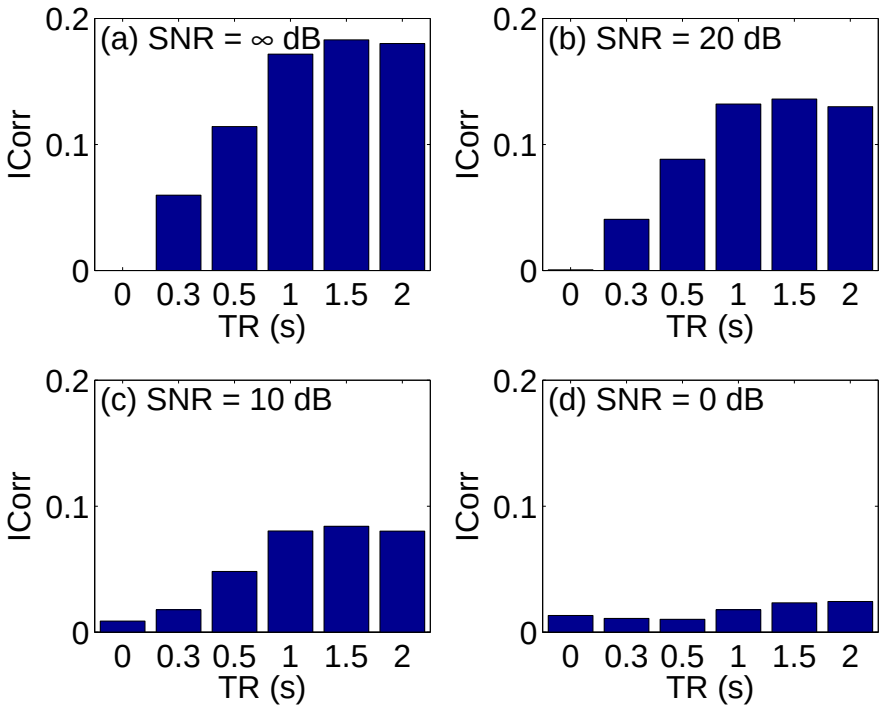


図 3.15: 人工的な雑音・残響環境での ICOR：(a) SNR = ∞ dB, (b) SNR = 20 dB, (c) SNR = 10 dB, (d) SNR = 0 dB.

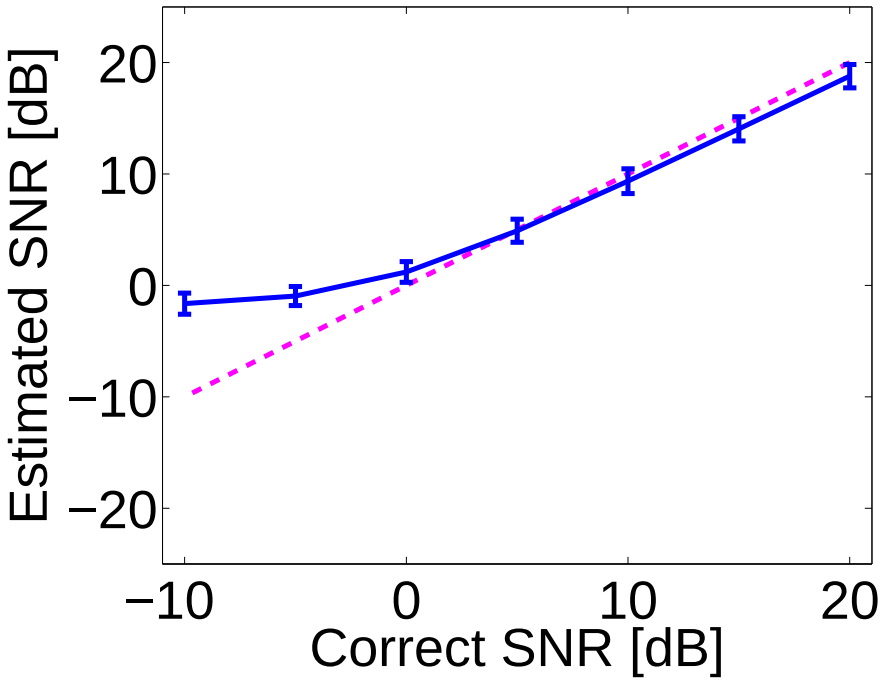


図 3.16: SNR の推定結果.

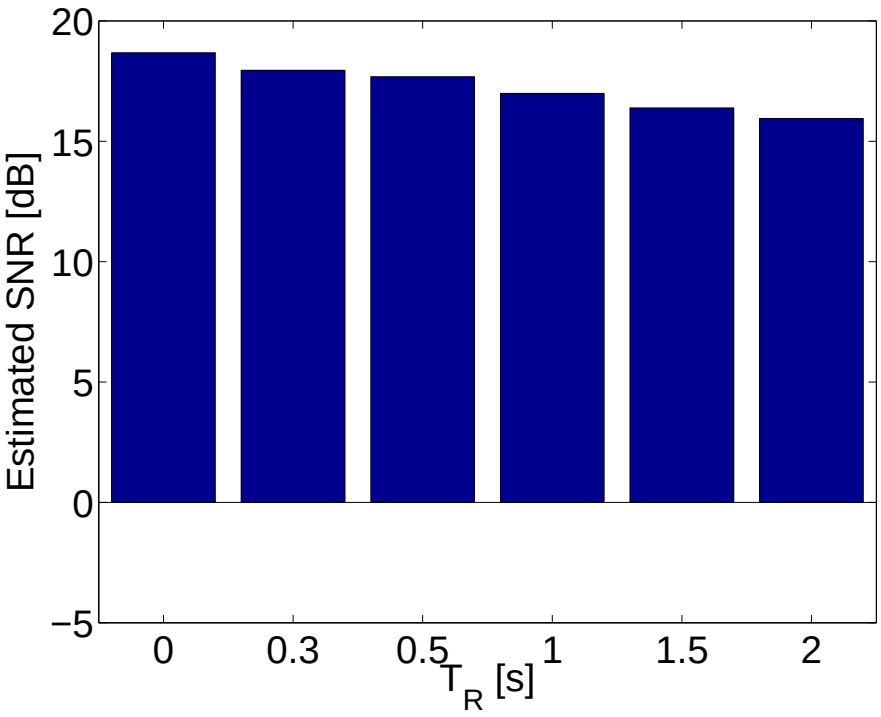


図 3.17: 雑音残響音声の SNR 推定結果 SNR = 20 dB.

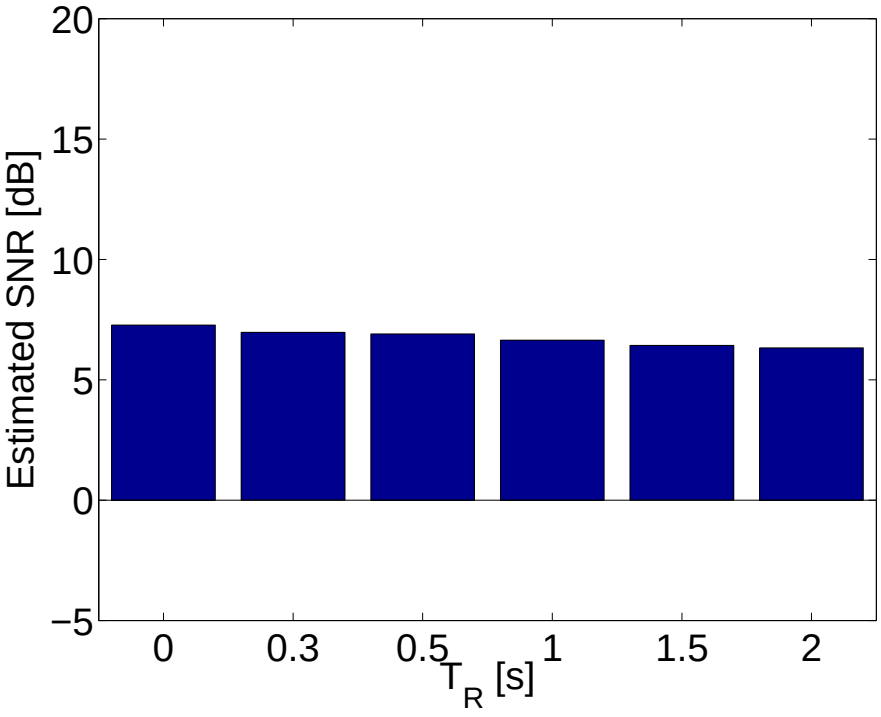


図 3.18: 雑音残響音声の SNR 推定結果 SNR = 10 dB.

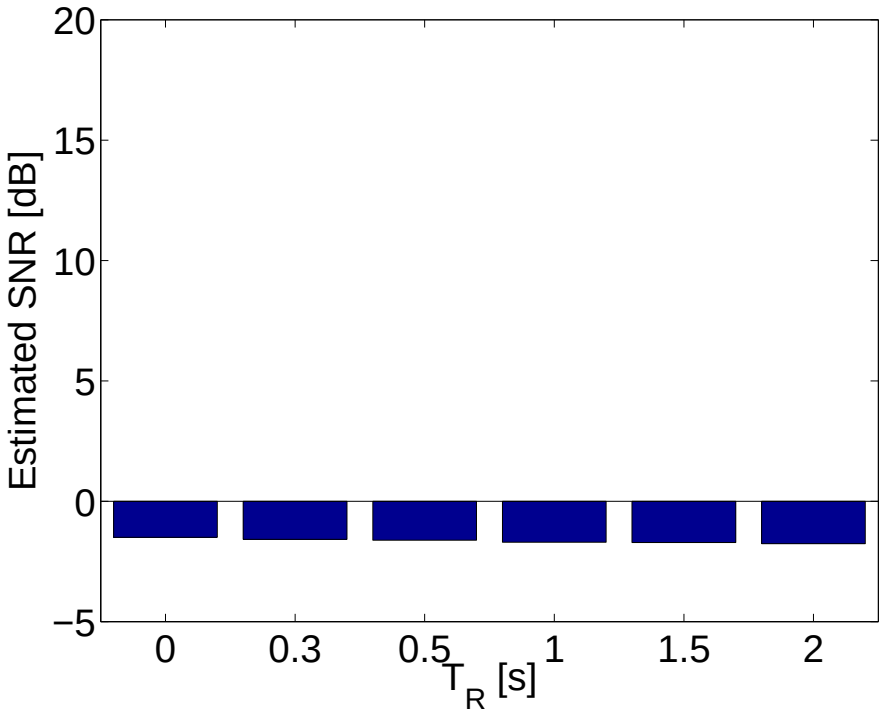


図 3.19: 雑音残響音声の SNR 推定結果 SNR = 0 dB.

SNR が低い条件（図 3.19）では残響時間が長くても SNR の推定誤差が小さいことがわかる．残響時間 0 s と 2 s の条件における，推定された SNR の差は約 0.3 dB であり，残響の影響を誤差として扱える．一方で，SNR が高い条件（図 3.17）では残響時間が長くなるにつれて SNR の推定誤差が大きくなっており，残響時間 0 s と 2 s の条件における，推定された SNR の差は約 2.7 dB である．したがって，SNR 推定では，SNR が低くなるにつれて，残響の影響を誤差として無視できる一方で，SNR が低い環境においては残響の影響を考慮する必要があることがわかった．

最後に，残響時間推定の性能評価を行う．ここでは，[37] での評価方法に従った．信号には，音声信号の代わりに 15 Hz の正弦波信号を用いた．RIR には残響時間 $T_R = 0.1, 0.2, 0.3, 0.5, 1.0, 2.0$ s の Schroeder の RIR [142] を用いた．雑音には SNR は 20, 10, 0 dB の白色ガウス雑音を利用した．各残響信号，雑音残響信号を 100 個生成して評価を行った．信号区間は既知として評価を行った．この時の評価結果を図 3.20 に示す．評価結果より，残響信号（SNR $= \infty$ dB）では Unoki らの結果 [37] と同様に，残響時間を 0.5 s までよく推定でき，残響時間が長くなるにつれて推定誤差が徐々に大きくなることがわかる．雑音残響信号では，SNR = 10

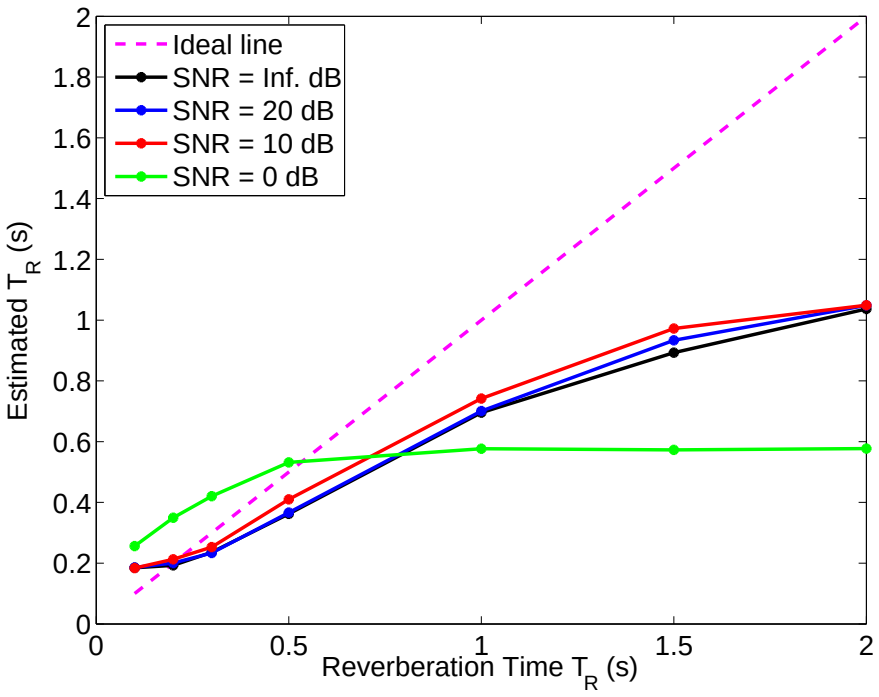


図 3.20: 残響，雑音残響環境下における残響時間の推定結果.

dB までは，パワーエンベロープ減算の効果により，残響信号と同等の推定性能が得られることがわかる．しかしながら，SNR $= 0$ dB においては，雑音の影響が大きくなり，原信号のパワーエンベロープに与える影響が大きいために推定誤差が大きく，$T_R = 0.5$ s 以降ではかなり誤差が大きく，SNR $= 0$ dB においては推定精度に関して課題が大きい結果となった．

ここでは，SNR 推定と残響時間推定という二つのパラメータ推定性能についての評価を行った．雑音のみ，残響のみの環境においては SNR と残響時間を精度よく推定できていることがわかる．しかしながら，雑音残響環境においては推定性能に誤差が生じていることがわかる．そのため，残響時間推定では雑音の影響を除去してから行うこととし，SNR $= 0$ dB での推定精度は今後の課題である．

第 4章

応用

本章では，統合的音声信号処理の応用として，帯域分割型パワーエンベロープ回復処理を前処理とした音声認識システムと，統合的音声信号処理を全面的に利用しているSTI推定について述べる．

4.1 音声認識のフロントエンド

本節では，帯域分割型パワーエンベロープ回復処理を前処理としたASRシステムについて述べる [156].

本研究におけるASRのメインシステムは，パワーエンベロープ回復を用いてASRを行うため，Lu et al.による残響音声に対するパワーエンベロープ回復を用いたASRシステムを利用する [151]．ASRでは，図4.1に示す特徴抽出に，ASRの前処理として帯域分割型パワーエンベロープ回復処理（図3.3）を組み込んだシステムを利用する．最初のK個のブロックは，パワーエンベロープからASRの特徴量に変換する処理である．まず，各帯域で（1）回復されたパワーエンベロープに対し，

$$\overline{e}_{x,k}[w] = \lambda \overline{e}_{x,k}[w-1] \times (1-\lambda)\hat{e}_{x,k}[w] \tag{4.1}$$

で平滑化処理（忘却係数$\lambda = 0.99$）を行い,（2）Hanning窓を利用したフレーム処理（32 msのフレーム長，16 msのフレームシフト）を適用し,（3）対数圧縮を行う．ここで，$\hat{e}_{x,k}[w]$は，回復された帯域分割パワーエンベロープ，$\overline{e}_{x,k}[w]$は平滑化処理後のパワーエンベロープ（wは時間フレーム番号）である．次に，ある

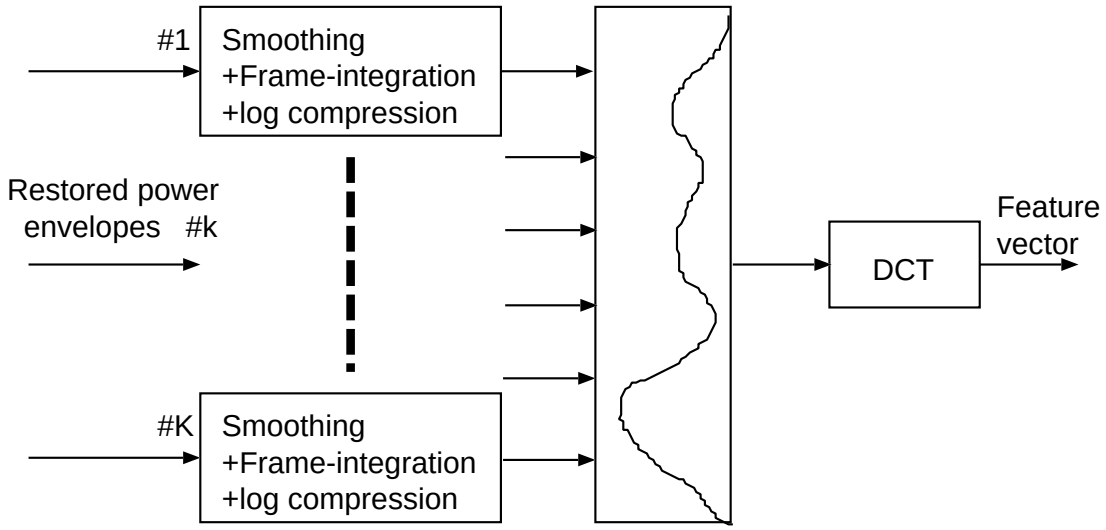


図 4.1: 帯域分割した回復パワーエンベロープに基づく音響特徴の抽出方法.

時刻での全帯域にわたる（1）～（3）の処理が施された特徴に対し，離散コサイン変換（DCT: Discrete Cosine Transform）を適用することで，一種のケプストラム情報を得る．ここで，最初の 12 次のケプストラム係数と対数パワー項を合わせた 13 次元の静的な特徴ベクトルとし，この静的な特徴の 1 次と 2 次の Δ ケプストラムを動的な特徴として取り扱う．そのため，これらを組み合わせた合計 39 次元の特徴ベクトルを利用することになる．HMM（Hidden Markov Model）の音響モデルは，AURORA-2J [146] で利用されたものと同じ構成とし，音響モデルの学習には，HTK3.2 [157] を利用した．そして，この ASR の前処理に 3.3.4 節の帯域分割型パワーエンベロープ回復処理を適用する．

帯域分割型パワーエンベロープ回復処理を前処理とした ASR の評価実験を行った．提案法の音響特徴は，帯域分割型パワーエンベロープ回復処理（図 3.3）と 特徴抽出（図 4.1）により抽出した（CBFB_IMTF），ケプストラム特徴である．比較のため，一般的な MFCC (Mel Frequency Cepstral Coefficient) 特徴についても同条件で ASR 評価し，MFCC の結果を今回の評価の基準とした．また， 定帯域フィルタバンクにより帯域分割されたパワーエンベロープに基づくケプストラム（CBFB）[151] と CBFB 上での RASTA フィルタ処理（CBFB_RASTA）も評価に用いた．音響モデルの学習には，AURORA-2J [146] の学習用音声 8440 発話を利用した．認識評価には，学習で利用していないテスト用音声 1001 発話を利用した．

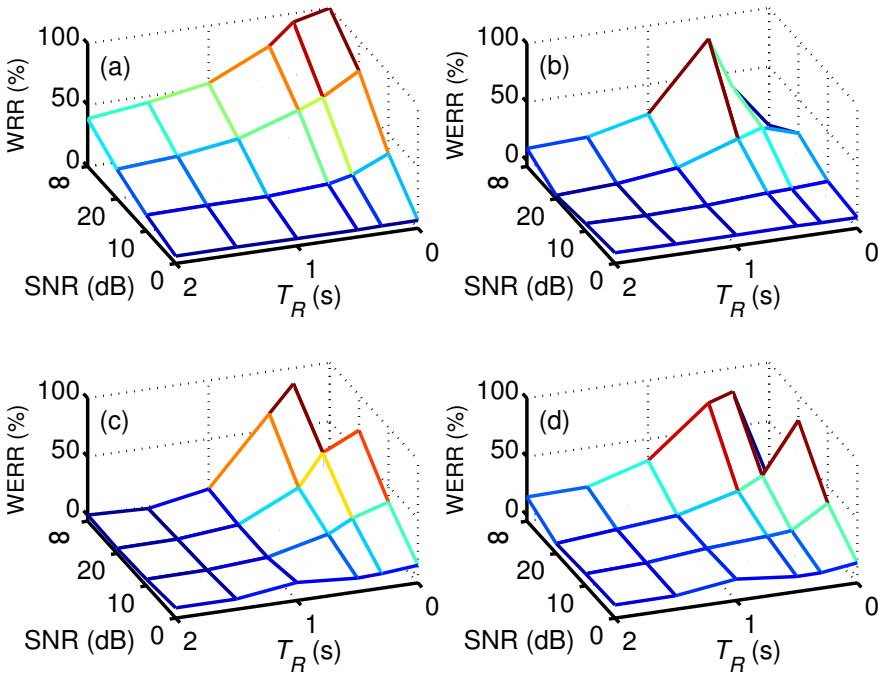


図 4.2: 人工的な雑音・残響環境における比較結果：(a)MFCC の word recognition rate (WRR)，(b)CBFB の word error reduction rate (WERR) ，(c) CBFB_RASTA の WERR，(d) CBFB_IMTF の WERR.

雑音残響音声は，1001 発話から各条件で生成した．雑音・残響条件は，3.3.4 節と同じである．ここでは，筆者らが提案した雑音残響に頑健な音声区間検出法を用いず，音声区間を既知とした．評価尺度には，MFCC の単語認識率（WRR: Word Recognition Rate）を基準として，CBFB では MFCC の単語認識率を基準に求めた単語誤り減少率（WERR: Word Error Reduction Rate）を用いた．人工的な雑音・残響環境における認識結果を図 4.2 に示す．図中（a）は MFCC の WRR，図中（a）以外は MFCC の単語認識率を基準に求めた WERR である．

WERR においては，正の値は基準とする MFCC より改善していることを示しており，負の値は MFCC より改悪になっていることを示す．どの手法も MFCC より改善していることが確認された．ほとんどの条件で，図 4.2 からは各手法の改善量に差があまりないように見える．ここで，雑音残響環境における劣悪な条件を拡大したものを図 4.3 に示す．この結果より，提案法の結果は，すべての結果で認識率が向上しており，他の手法と比較しても WERR が大きいことがわかる．帯域分割型パワーエンベロープ回復処理が音声認識率に寄与することが確認された．

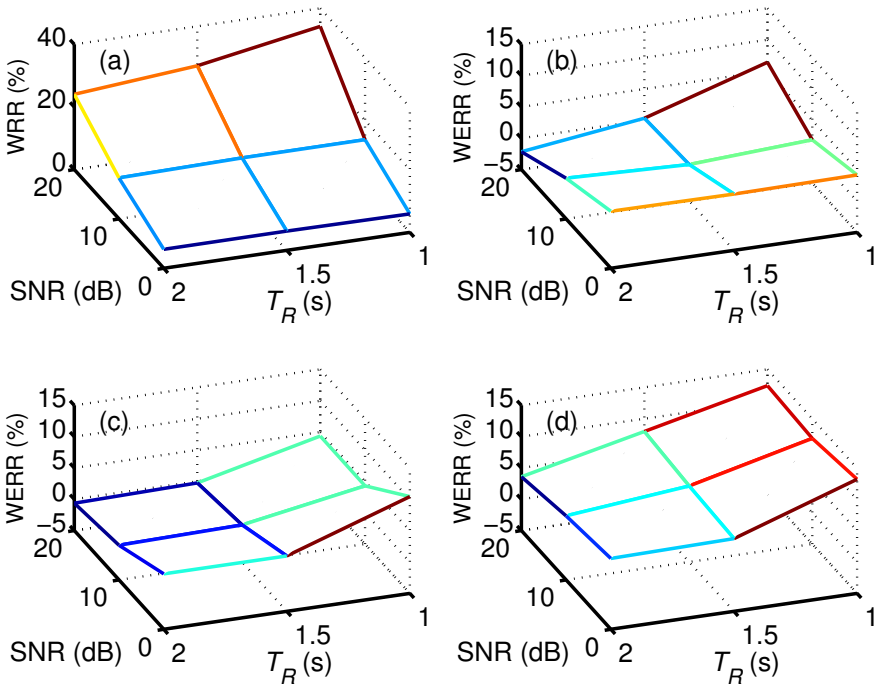


図 4.3: 図 4.2 の雑音残響環境の拡大.

続いて，実環境を想定した雑音残響環境において ASR の評価実験を行った．音声信号 $x(t)$ として，AURORA-2J 音声データベース [146] のテスト用の 1001 個のクリーン音声を利用した．室内インパルス応答 $h(t)$ として，SMILE2004 [153] の実環境で集音された 8 個の RIR と，背景雑音 $n(t)$ として NOISEX-92 [155] の白色雑音，ピンク雑音，バブル雑音，工場雑音 2 種類を利用した．

MFCC についても同条件で ASR 評価を行い，評価の基準とした．CBFB と ETSI による denoted AFE (Advanced Front-End) [158] も利用した．また，CMN [159] を CBFB 上で用いた特徴（CBFB_CMN），RASTA [160] を CBFB 上で用いた特徴（CBFB_RASTA）を残響抑圧法として比較に用いた．雑音除去法として SS 法 [50] を CBFB 上で用いた特徴（CBFB_SS）も用いた． 雑音残響に頑健な手法として SS 法と RASTA を組み合わせて CBFB 上で用いた特徴（CBFB_SS_RASTA）も比較に利用した．

各種雑音条件での評価結果を，それぞれ表 4.1，4.2，4.3，4.4，4.5 に示す．ただし，MFCC は単語認識率（WRR）であり，MFCC 以外の特徴は MFCC の単語認識率を基準に求めた単語誤り減少率（WERR）である．RIR の番号は，SMILE2004 [153] のファイル番号に該当しており，室に関する情報は表 3.1 と 3.2 に記してあ

る．また，太字は最も認識性能が高かった値を示している．

結果より，白色雑音，ピンク雑音，工場雑音 1 の多くの条件下においては，提案法である CBFB_IMTF が最もよい認識性能であった．バブル雑音や工場雑音 2 の残響時間の短い環境では，ケプストラム平均正規化である CBFB_CMN の認識性能が高い場合が確認された．雑音と残響の影響の大きい劣悪な雑音残響環境においては，ほとんどの条件下で提案法が最も優れた結果となり．雑音残響の影響が大きい条件ほど提案法が有効であることが示された．これらの結果より，パワーエンベロープ回復処理が ASR の前処理として機能することで，雑音残響環境において音声認識率を向上できることを示した．したがって，統合的音声信号処理を ASR の前処理として利用することにより，劣悪な雑音残響環境における人と機械の音環境バリアフリーに貢献できる一例を示すことができた．

4.2 STI 推定

統合的音声信号処理の応用の二例目として，統合的音声信号処理を用いた STI 推定について述べる [161, 162]．雑音残響環境において AM 信号もしくは音声信号を用いて STI を推定する方法である．ここでは，AM 信号を用いた STI 推定法について述べる．図 4.4 に STI 推定法のブロックダイアグラムを示す．本推定法では，雑音の MTF と残響の MTF を推定するという，大きく分けて二つの処理を中心として STI を推定している．雑音の MTF は，統合的音声信号処理で用いられる雑音残響に頑健な VAD と SNR 推定によって推定される．残響の MTF 推定では，まず，推定された SNR を用いてパワーエンベロープ減算処理を行い，雑音のパワーを減算することで雑音の影響を取り除いた．その後，変調スペクトル上で変調度 1 を仮定して，逆フィルタ処理により MTF を推定し，そこから Schroeder の RIR を改良した一般化 RIR を再生成する．そして推定した一般化 RIR と雑音の MTF を用いて，IEC 60268-16 [28] に準拠して STI を推定している．

原信号には，基本変調周波数 5 Hz の AM 信号を用いた．RIR には SMILE2004 [153] に収録された実環境で集音された 43 個の RIR，雑音には SNR = 20, 5 dB のガウス白色雑音を用いて，雑音残響 AM 信号を生成した．

SNR = 20 dB の条件での評価結果を図 4.5 に示す．推定値と計算値が一致した

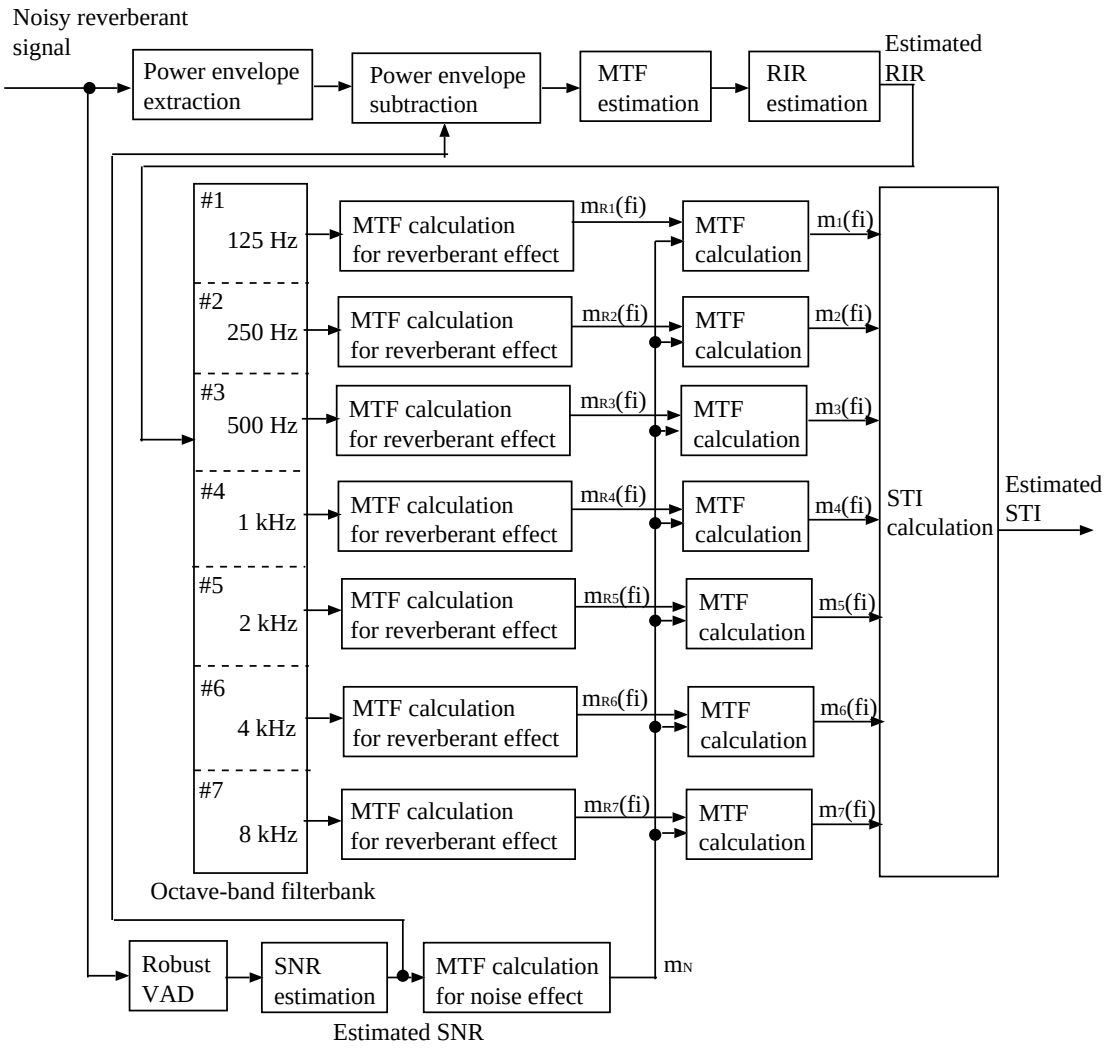


図 4.4: STI 推定の概要図.

場合，破線上に結果が布置されるため，破線上に結果が布置されるほど推定精度が高いことを意味している．図 4.5 において，結果が破線上に布置されていることから精度よく推定できており，RMS も 0.04 という結果になった．雑音の影響がさらに大きい SNR = 5 dB における評価結果を図 4.6 に示す．SNR = 20 dB の時とほぼ同等の推定結果を得ることができ，RMS も 0.05 という結果となった．この結果より，統合的音声信号処理を用いることにより STI を精度よく推定でき，このような音環境の特徴を用いる様々な音声信号処理において利用できることを示した．

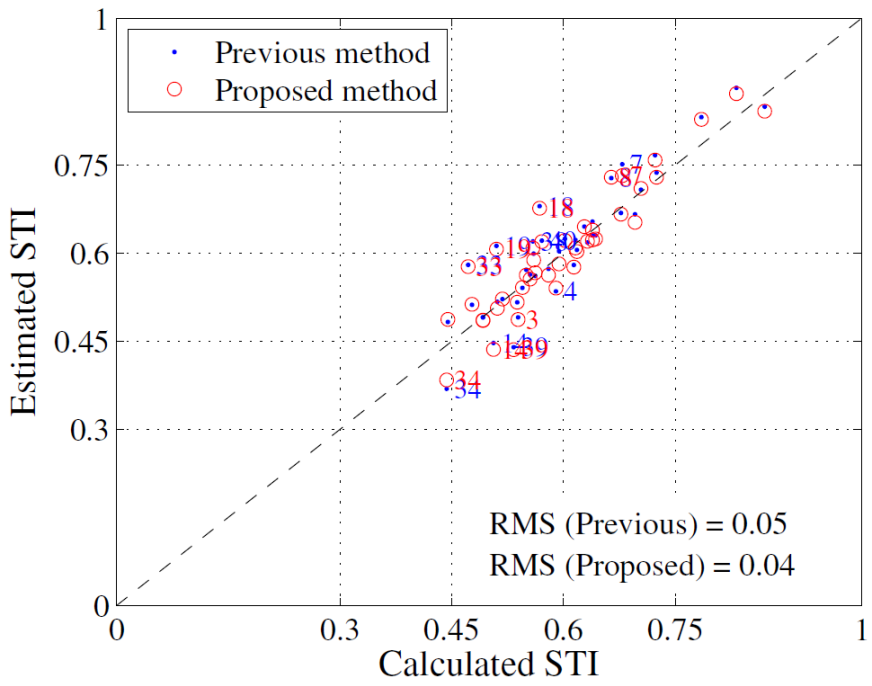


図 4.5: SNR=20 dB の雑音残響 AM 信号による STI 推定結果 [161].

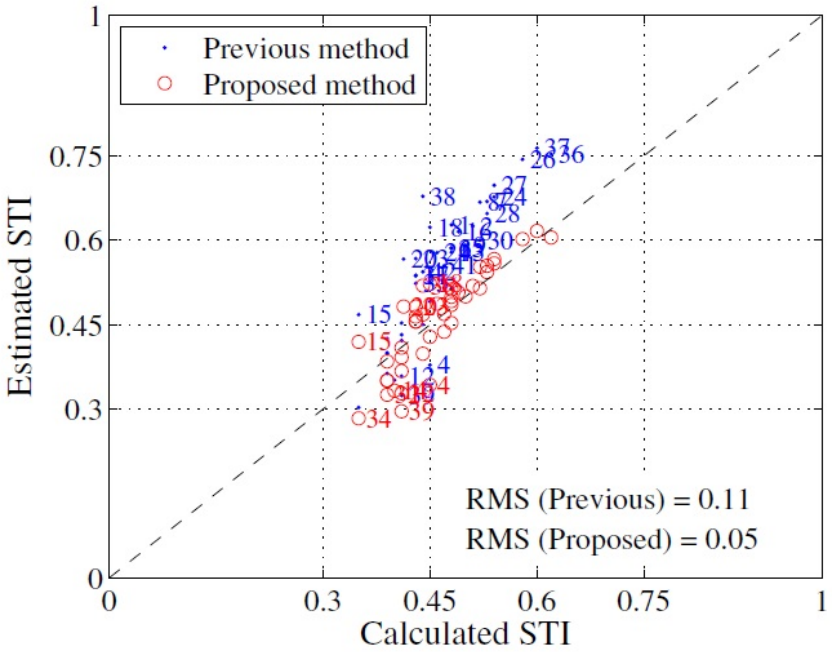


図 4.6: SNR=5 dB の雑音残響 AM 信号による STI 推定結果 [161].

表 4.1: MFCC を基準とした WERR の比較結果（白色雑音）.

RIRs No.	411	405	301	404	305	309	407	408
T_R (s)	0.36	0.71	1.09	1.54	1.93	2.35	3.03	3.62
				White noise 20 dB				
MFCC (WRR)	36.94	24.35	21.58	18.58	16.79	12.00	15.39	15.77
AFE	65.52	28.29	16.77	20.24	7.86	16.19	7.41	8.60
CBFB	67.24	29.09	12.18	8.97	1.12	4.85	0.19	0.46
CBFB_SS	68.11	31.67	28.16	32.92	17.94	24.05	18.91	21.76
CBFB_CMN	76.85	37.94	16.64	9.65	0.64	4.30	*-0.58*	*-1.40*
CBFB_RASTA	61.43	22.23	9.02	11.00	*-1.50*	5.17	1.42	2.95
CBFB_SS_RASTA	66.21	45.97	31.47	33.14	17.13	25.13	19.05	**22.30**
CBFB_IMTF	**78.81**	**58.40**	**44.29**	**37.14**	**24.88**	**26.52**	**20.14**	21.47
				White noise 10 dB				
MFCC (WRR)	14.12	10.87	12.93	13.36	12.56	11.91	10.44	11.24
AFE	34.79	14.94	8.53	10.27	1.90	4.43	0.71	1.42
CBFB	6.35	1.31	*-5.96*	*-9.33*	*-8.43*	*-7.91*	*-5.66*	*-6.23*
CBFB_SS	37.76	24.22	14.84	13.25	10.14	11.40	12.20	12.38
CBFB_CMN	32.89	14.81	4.62	1.41	*-0.29*	*-1.50*	*-0.03*	*-1.22*
CBFB_RASTA	20.41	8.75	1.44	2.02	*-1.20*	*-1.99*	*-0.03*	-0.48
CBFB_SS_RASTA	39.58	28.45	16.18	16.54	9.26	10.42	9.32	10.51
CBFB_IMTF	**55.71**	**41.16**	**33.92**	**28.81**	**20.96**	**19.53**	**17.07**	**17.29**
				White noise 0 dB				
MFCC (WRR)	6.42	6.39	6.02	7.31	7.25	8.44	9.30	9.06
AFE	*-2.87*	*-5.61*	*-4.64*	*-19.7*	1.19	*-1.10*	*-1.72*	*-1.45*
CBFB	5.61	4.88	4.83	3.31	2.88	2.15	*-0.74*	*-0.31*
CBFB_SS	4.72	2.68	0.35	0.26	*-0.63*	*-1.77*	*-1.25*	*-1.77*
CBFB_CMN	5.21	3.80	3.14	1.06	1.02	*-0.60*	*-1.65*	*-1.48*
CBFB_RASTA	5.58	4.58	3.69	2.35	2.38	*-0.23*	*-1.15*	*-0.85*
CBFB_SS_RASTA	5.58	3.28	0.49	*-1.58*	*-2.44*	*-3.61*	*-3.31*	*-2.04*
CBFB_IMTF	**24.18**	**21.94**	**20.36**	**18.02**	**15.78**	**12.81**	**9.89**	**9.89**

表 4.2: MFCC を基準とした WERR の比較結果（ピンク雑音）.

RIRs No.	411	405	301	404	305	309	407	408
T_R (s)	0.36	0.71	1.09	1.54	1.93	2.35	3.03	3.62
				Pink noise 20 dB				
MFCC (WRR)	50.11	27.76	24.13	23.18	19.50	16.24	14.77	15.44
AFE	78.59	40.03	25.41	25.46	14.57	18.52	15.78	16.89
CBFB	**90.28**	50.28	27.15	19.42	7.63	11.15	12.85	14.89
CBFB_SS	81.60	56.99	36.02	29.34	14.98	20.61	22.22	25.09
CBFB_CMN	90.10	**60.17**	34.32	18.86	5.75	5.10	7.90	8.50
CBFB_RASTA	82.96	46.32	26.80	19.94	7.50	9.49	13.43	17.18
CBFB_SS_RASTA	84.19	57.63	34.77	27.86	14.76	21.63	**23.13**	**25.82**
CBFB_IMTF	79.51	58.18	**44.52**	**33.38**	**24.37**	**25.33**	22.19	24.88
				Pink noise 10 dB				
MFCC (WRR)	31.19	17.10	14.95	15.04	14.09	11.91	11.08	11.21
AFE	28.08	12.22	1.79	8.13	*-3.39*	1.23	*-0.03*	0.45
CBFB	12.32	3.00	*-2.26*	*-5.81*	*-7.11*	*-6.23*	*-2.94*	*-3.29*
CBFB_SS	50.02	32.48	20.14	16.98	9.25	13.97	13.44	15.70
CBFB_CMN	**70.99**	**41.86**	22.26	12.36	3.97	4.64	5.53	6.70
CBFB_RASTA	60.73	32.86	17.41	13.96	17.32	6.94	9.44	11.48
CBFB_SS_RASTA	57.17	38.75	23.18	16.44	6.90	15.20	16.44	18.56
CBFB_IMTF	49.48	38.82	**31.00**	**28.26**	**19.90**	**21.30**	**19.41**	**19.98**
				Pink noise 0 dB				
MFCC (WRR)	13.45	9.82	9.76	10.10	9.64	9.15	9.54	9.18
AFE	*-0.92*	*-5.74*	*-9.25*	*-4.92*	*-10.13*	*-8.03*	*-7.39*	*-7.54*
CBFB	*-25.65*	*-18.93*	*-15.51*	*-15.74*	*-11.31*	*-11.88*	*-9.52*	*-13.08*
CBFB_SS	1.06	*-0.17*	*-1.22*	*-4.24*	*-3.26*	*-2.62*	*-4.73*	*-3.95*
CBFB_CMN	17.74	8.48	5.08	2.70	0.34	0.73	0.90	0.85
CBFB_RASTA	17.60	9.06	5.82	4.75	2.99	2.01	3.08	2.47
CBFB_SS_RASTA	4.44	1.98	0.79	*-1.50*	*-2.79*	*-1.02*	*-1.72*	*-2.57*
CBFB_IMTF	**17.89**	**13.86**	**13.18**	**12.19**	**8.29**	**11.45**	**7.75**	**7.84**

表 4.3: MFCC を基準とした WERR の比較結果（バブル雑音）.

RIRs No.	411	405	301	404	305	309	407	408
T_R (s)	0.36	0.71	1.09	1.54	1.93	2.35	3.03	3.62
				Babble noise 20 dB				
MFCC (WRR)	90.11	54.81	42.40	39.27	31.84	30.40	29.75	30.24
AFE	*-1.52*	4.01	1.88	6.01	2.44	5.07	1.54	1.32
CBFB	42.26	9.98	2.34	*-1.12*	*-7.63*	*-3.62*	*-4.37*	*-3.83*
CBFB_SS	69.57	*-8.43*	13.44	7.69	*-0.28*	2.07	3.54	6.34
CBFB_CMN	**69.26**	**53.99**	25.49	13.55	1.76	0.13	*-0.44*	1.98
CBFB_RASTA	52.48	33.08	15.19	9.20	0.72	1.90	3.63	8.59
CBFB_SS_RASTA	68.96	35.32	14.93	10.41	3.65	9.74	6.73	9.29
CBFB_IMTF	40.44	44.21	**31.51**	**16.93**	**9.77**	**10.76**	**7.56**	**9.86**
				Babble noise 10 dB				
MFCC (WRR)	90.05	53.39	41.30	37.89	30.83	29.60	29.38	30.21
AFE	*-2.11*	6.99	2.30	7.62	3.15	5.31	1.83	1.23
CBFB	*-20.00*	*-11.20*	*-9.57*	*-8.21*	*-12.09*	*-9.33*	*-9.83*	*-9.80*
CBFB_SS	59.60	24.44	10.24	5.43	*-2.53*	0.30	1.66	3.30
CBFB_CMN	**71.26**	**56.85**	27.14	16.02	2.13	0.99	0.65	1.72
CBFB_RASTA	55.88	34.13	15.32	9.98	0.04	0.30	2.35	6.78
CBFB_SS_RASTA	61.71	28.66	9.73	6.92	*-0.10*	5.88	4.74	4.97
CBFB_IMTF	33.07	41.90	**29.81**	**17.05**	**10.64**	**10.72**	**6.40**	**8.50**
				Babble noise 0 dB				
MFCC (WRR)	87.69	48.20	38.87	35.86	28.55	27.57	27.76	28.46
AFE	*-6.01*	12.45	4.22	9.43	4.34	7.12	1.05	1.41
CBFB	*-121.20*	*-26.08*	*-19.43*	*-16.81*	*-14.39*	*-13.99*	*-12.71*	*-12.75*
CBFB_SS	42.89	17.55	6.43	1.68	*-3.47*	0.30	*-0.26*	0.08
CBFB_CMN	**76.52**	**60.95**	**31.44**	**18.01**	3.92	1.53	1.40	2.80
CBFB_SS-RASTA	49.63	24.19	5.87	3.21	*-2.31*	1.95	0.72	1.47
CBFB_IMTF	16.41	37.88	28.68	15.56	**10.62**	**12.34**	**6.51**	**7.46**

表 4.4: MFCC を基準とした WERR の比較結果（工場雑音 1）.

RIRs No.	411	405	301	404	305	309	407	408
T_R (s)	0.36	0.71	1.09	1.54	1.93	2.35	3.03	3.62
				Factory1 noise 20 dB				
MFCC (WRR)	66.33	38.16	31.75	30.61	24.95	23.64	23.00	23.58
AFE	71.81	34.31	18.26	18.19	12.29	12.99	9.36	10.89
CBFB	82.12	29.59	13.36	7.03	*-0.25*	0.76	1.19	3.94
CBFB_SS	textbf90.23	55.85	31.71	24.07	10.35	13.55	15.06	18.05
CBFB_CMN	88.33	**60.24**	31.94	14.34	2.76	*-1.96*	0.64	2.77
CBFB_RASTA	81.68	42.76	21.36	13.36	3.28	1.45	6.38	10.29
CBFB_SS_RASTA	89.69	57.10	31.53	22.97	11.84	**17.77**	**15.99**	**18.52**
CBFB_IMTF	78.02	57.89	**40.53**	**24.38**	**17.04**	16.04	13.91	16.88
				Factory1 noise 10 dB				
MFCC (WRR)	38.44	21.19	18.91	19.01	16.36	15.26	15.38	16.33
AFE	79.45	45.81	30.30	28.53	16.24	19.64	14.33	15.82
CBFB	54.27	17.45	8.90	6.67	*-0.47*	0.98	0.73	0.55
CBFB_SS	87.13	55.27	35.44	26.76	16.08	20.86	20.98	23.09
CBFB_CMN	84.49	53.25	29.84	15.38	4.75	3.08	3.34	3.24
CBFB_RASTA	71.91	38.37	22.27	15.09	6.05	4.24	7.62	9.04
CBFB_SS_RASTA	**87.83**	**60.11**	36.50	26.57	16.19	19.93	20.07	22.21
CBFB_IMTF	80.51	58.94	**43.29**	**31.84**	**23.35**	**23.12**	**20.12**	**21.29**
				Factory1 noise 0 dB				
MFCC (WRR)	15.54	12.13	10.93	11.05	10.47	8.69	10.16	9.52
AFE	55.21	29.28	21.26	20.67	9.23	11.53	6.67	8.95
CBFB	2.79	*-4.27*	*-3.10*	*-7.32*	*-5.04*	*-3.57*	*-4.31*	*-3.87*
CBFB_SS	29.22	10.94	7.76	4.87	*-0.61*	4.61	3.01	5.02
CBFB_CMN	36.93	15.71	8.48	5.18	2.33	1.68	0.38	1.25
CBFB_SS-RASTA	38.79	19.56	13.82	7.12	2.19	5.18	2.67	4.41
CBFB_IMTF	**57.47**	**40.42**	**33.88**	**28.99**	**22.18**	**22.83**	**18.32**	**20.93**

表 4.5: MFCC を基準とした WERR の比較結果（工場雑音 2）.

RIRs No.	411	405	301	404	305	309	407	408
T_R (s)	0.36	0.71	1.09	1.54	1.93	2.35	3.03	3.62
				Factory2 noise 20 dB				
MFCC (WRR)	90.11	54.81	42.40	39.27	31.84	30.40	29.75	30.24
AFE	*-1.52*	4.01	1.88	6.01	2.44	5.07	1.54	1.32
CBFB	42.26	9.98	2.34	*-1.12*	*-7.63*	*-3.62*	*-4.37*	*-3.83*
CBFB_SS	69.57	*-8.43*	13.44	7.69	*-0.28*	2.07	3.54	6.34
CBFB_CMN	**69.26**	**53.99**	25.49	13.55	1.76	0.13	*-0.44*	1.98
CBFB_RASTA	52.48	33.08	15.19	9.20	0.72	1.90	3.63	8.59
CBFB_SS_RASTA	68.96	35.32	14.93	10.41	3.65	9.74	6.73	9.29
CBFB_IMTF	40.44	44.21	**31.51**	**16.93**	**9.77**	**10.76**	**7.56**	**9.86**
				Factory2 noise 10 dB				
MFCC (WRR)	90.05	53.39	41.30	37.89	30.83	29.60	29.38	30.21
AFE	*-2.11*	6.99	2.30	7.62	3.15	5.31	1.83	1.23
CBFB	*-20.00*	*-11.20*	*-9.57*	*-8.21*	*-12.09*	*-9.33*	*-9.83*	*-9.80*
CBFB_SS	59.60	24.44	10.24	5.43	*-2.53*	0.30	1.66	3.30
CBFB_CMN	**71.26**	**56.85**	27.14	16.02	2.13	0.99	0.65	1.72
CBFB_RASTA	55.88	34.13	15.32	9.98	0.04	0.30	2.35	6.78
CBFB_SS_RASTA	61.71	28.66	9.73	6.92	*-0.10*	5.88	4.74	4.97
CBFB_IMTF	33.07	41.90	**29.81**	**17.05**	**10.64**	**10.72**	**6.40**	**8.50**
				Factory2 noise 0 dB				
MFCC (WRR)	87.69	48.20	38.87	35.86	28.55	27.57	27.76	28.46
AFE	*-6.01*	12.45	4.22	9.43	4.34	7.12	1.05	1.41
CBFB	*-121.20*	*-26.08*	*-19.43*	*-16.81*	*-14.39*	*-13.99*	*-12.71*	*-12.75*
CBFB_SS	42.89	17.55	6.43	1.68	*-3.47*	0.30	*-0.26*	0.08
CBFB_CMN	**76.52**	**60.95**	**31.44**	**18.01**	3.92	1.53	1.40	2.80
CBFB_RASTA	60.36	37.10	15.38	10.20	1.76	2.13	3.70	7.13
CBFB_SS-RASTA	49.63	24.19	5.87	3.21	*-2.31*	1.95	0.72	1.47
CBFB_IMTF	16.41	37.88	28.68	15.56	**10.62**	**12.34**	**6.51**	**7.46**

第 5章

結論

5.1 本論文で明らかにしたこと

本研究では，ユビキタス音声コミュニケーションの “どこでも” に着目したときに，実環境では雑音や残響の影響が音環境のバリアとなるという問題に対して，音環境バリアフリーを実現することを大きな目標とした．本論文では，この音環境バリアフリーの問題を加法性の定常雑音および拡散音場を仮定した雑音残響環境とし，この音場においてコミュニケーションの円滑化を図るべく，音環境バリアフリーのためのパワーエンベロープ処理体系を示し，統合的音声信号処理によって音環境バリアフリーを実現できることを示した．

音環境バリアフリーのためのパワーエンベロープ処理体系では，変調伝達関数の概念に基づき変調度 1 を規範とし，最適化による逆フィルタ処理によって音環境と人の調和が取れた処理になることを示した．最適化による逆フィルタ処理では，変調度 1 を規範と処理にすることで，従来手法で起こる過少・過剰回復を低減できる理論・アプローチであることを示した．この変調度 1 を規範とする処理の実現にあたり，VAD ではパワーエンベロープの回復とパワー閾値の最適化が重要であり，変調度 1 を規範とした音声/非音声判別の重要性について述べた．

音環境バリアフリーのためのパワーエンベロープ処理体系を実現するために，統合的音声信号処理を提案した．統合的音声信号処理では，雑音残響に頑健な VAD，パワーエンベロープ回復処理，SNR と残響時間のパラメータ推定により実現した．雑音残響に頑健な VAD では，パワーエンベロープ回復と最適化したパワー閾値を

用いることで，変調度 1 のパワーエンベロープに対するパワー閾値と同等の性能で音声区間検出をできるアプローチを提案した．パワーエンベロープ回復処理は，変調度 1 に着目したパワーエンベロープ減算処理とパワーエンベロープ逆フィルタ処理による方法について述べ，SNR からパワーエンベロープ減算を行う方法を示した．SNR 推定法では，頑健な VAD で用いたパワー閾値の最適化の考えと帯域分割処理による方法を提案した．残響時間推定法は，従来の残響信号のみに対応した理論を，雑音残響信号に対応した理論へと拡張した．統合的音声信号処理の各要素技術について評価を行った結果，まず，人工的な雑音残響環境ならびに実環境を想定した雑音残響環境において提案した頑健な VAD 法は，他の比較手法に比べて音声/非音声の検出性能が高い結果となった．また，SNR や残響時間の推定評価では，雑音残響環境において概ね精度よく推定できることを示した．

統合的音声信号処理の応用として，ASR システムと STI 推定法について述べた．ASR では，統合的音声信号処理の帯域分割型パワーエンベロープ回復を ASR の前処理とした ASR システムとした．ASR の音響特徴量には，回復した帯域分割パワーエンベロープから一種のケプストラムを求めて利用した．雑音残響環境における評価の結果，雑音・残響の影響が大きい環境では，比較手法よりも提案法が優れた結果となった．室内の音声伝達性能の物理指標である STI の推定法は，統合的音声信号処理の頑健な VAD と SNR 推定，パワーエンベロープ減算処理が利用されており，AM 信号や音声信号を観測信号として推定する．残響の MTF は変調スペクトル上で残響時間を推定して統計的な RIR を生成してから求めた．そして，雑音の MTF と残響の MTF から STI を計算した．評価の結果，この STI 推定法においても，精度よく STI を推定できることを示した．これらの統合的音声信号処理の応用結果より，パワーエンベロープ処理体系が一アプリケーションとして音環境バリアフリーに貢献できることを示した．

本論文では，加法性の定常雑音および拡散音場を仮定した雑音残響環境ではあるが，パワーエンベロープを用いて，統合的音声信号処理により音環境バリアフリーを実現できることを示し，人と機械の音声コミュニケーションに貢献できることを示した．

5.2 今後の展望

本研究では，ユビキタス音声コミュニケーションの “どこでも” に着目して，ASR を通じて人と機械の音声コミュニケーションにおいて音環境バリアフリーを実現した．この成果により，ユビキタス音声コミュニケーションの “どこでも” の一部の問題を解決できたものと考えるが，本論文では，雑音に定常性を仮定し，拡散音場を仮定してパワーエンベロープに対する処理で実現した．より多くの音環境において音バリアフリーを実現するためには，本論文で扱っていない非定常性の雑音や突発性雑音，拡散音場ではない屋外などの音場やマイクロフォン距離などの問題についても，今後取り組む必要がある．

パワーエンベロープ回復処理においては，MTF の概念では時間変化に一定な雑音を仮定しており，雑音のパワーエンベロープは時間変化に一定であることを仮定しているが，実際に時間変化に行っていな雑音は無く時間変動を有している．同様のことは，RIR のパワーエンベロープにも言えることである．このようなパワーエンベロープの微細な時間変動は，パワーエンベロープ減算処理やパワーエンベロープ逆フィルタ処理，残響時間推定などにおいて誤差の原因となる．今後，このような微細な時間変動の問題に対処することができれば，現在よりも高精度に回復・推定が行えるようになるものと考えている．パワーエンベロープ回復の更なる発展として考えると，現在，原信号のパワーエンベロープの変調度が 1 であると仮定した処理体系としているが，個人の持つ発話能力の差などによって，原信号の変調度が常に 1 であるとは考えにくい．そこで，原信号のパワーエンベロープの変調度が 1 以下であるような場合，変調度が 1 となるようにパワーエンベロープを回復することで，明瞭性を向上させるようなパワーエンベロープ回復処理の検討が必要である．

本論文では，パワーエンベロープ回復ならびにパラメータ推定によって音環境バリアフリーを実現する統合的音声信号処理を提案した．パワーエンベロープ処理体系で述べた，最適化による MTF の逆フィルタ処理というアプローチによる音環境バリアフリーを実現でき，SNR，残響時間，音声区間，回復パワーエンベロープを同時に求めることができれば，処理の単純化ができるだけでなく，それぞれが協調した処理となることから回復・推定性能の向上も期待できる．しかし，

実現にあたっては，音声信号の変調スペクトルでの振る舞いと扱い方についての十分な検討・考察を要するものと考えられる．

雑音残響に頑健なVADは，音声区間を精度よく推定することを目的としてFARとFRRのRMSが最小となるような手法を本論文で提案した．しかしながら，ASRの前処理に利用することを考えると，従来の手法のように，音声区間を確実に検出する必要があるという問題に陥る．そこで，音声/非音声として判別された区間に対して，ケプストラムやスペクトル，変調スペクトルを用いた判別を行うことで，音声区間検出性能を向上させる必要があるものと考えている．また，音声信号に対する処理であるため，帯域分割処理を組み合わせた処理とすることで検出性能の向上も期待できるが，各帯域で検出された音声/非音声情報をどのように統合して，最終的な音声/非音声区間として判別するかといった問題に陥るため，さらなる検討・考察が必要となる．

本研究で用いたパワーエンベロープは，Drullmanが述べているように音声知覚に重要な特徴を含んでいる．音声信号はエンベロープと微細構造に分けることができ，本研究で扱ったのは時間包絡線であるエンベロープのみである．人工内耳で発話者の識別を目標とした研究があり，Zhu et al.はフィルタバンクにより帯域分割したエンベロープの情報に着目し，このエンベロープに感情情報が含まれている可能性を示唆している [163, 164]．帯域分割したエンベロープに対して各帯域で雑音を付加して再合成すると，雑音駆動音声になる．雑音駆動音声は，難聴者が人工内耳を装着した時に聴こえる音声と言われており，発話内容をある程度認識することができる．例えば，人工内耳の前処理として統合的音声信号処理の帯域分割型パワーエンベロープ処理を適用させることで，音環境バリアフリーを実現する人工内耳として利用できる可能性がある．また，変調度1を規範とした処理であることから，過少・過剰回復が軽減され，どのような音環境でも快適に音声を聴き取ることに発展できるかもしれない．もし実現できれば，難聴者の音環境バリアフリーへ貢献できることから，ユビキタス音声コミュニケーションの “誰とでも” の難聴者とでも，として貢献できる可能性がある．

この他，雑音残響信号から回復音声を再生性できる可能性もある．Unoki et al.の研究によって，残響音声から音声の微細構造を再生成し，回復したエンベロープと合成して回復音声を得る研究もある [165]．本研究は，統合的音声信号処理に

よって雑音残響信号のパワーエンベロープを回復するに至っている．したがって，雑音残響信号から微細構造を精度よく再生成できれば音声を回復し，人と人との音声コミュニケーションを円滑に行うことにつながることから，ユビキタス音声コミュニケーションの “誰とでも” に貢献することができる．しかしながら，現状では要素技術が全く整っていないために実現にはかなりの時間を要すると考えられる．これは，主に微細構造を生成するのに必要な情報である基本周波数（F0）や有声/無声を雑音残響信号から精度よく推定・検出できないためである．もし，雑音残響に頑健なこれらの要素技術が実現すれば，統合的音声信号処理を用いた音声回復への道は大きく前進して発展するものと考えられる．これらの問題にも少し取り組んでいるので，簡単に問題点を列挙したいと思う．雑音音声の F0 推定は櫛形フィルタを用いるなどして現状の手法で対応できるものの，残響音声の影響により音声が歪んでいるため，1, 2 秒程度での F0 は精度よく推定できるが，逐次変化する F0 に対応することができない．残響の問題が解決できないために，雑音残響における F0 推定の目処は立っていない．一方で，有声/無声判別は更に難しい問題となる．これは，雑音・残響それぞれの環境でも解決されていないためである．雑音によって母音や子音，特に子音が簡単に雑音に埋もれる，残響の影響によって母音や子音の区間がずれるといった問題がある．現状これらの解決策は見出されていない．

今後の音環境バリアフリーの研究が発展することで，ユビキタス音声コミュニケーションとして完結する日が来ることを願っている．

謝辞

本研究を遂行するにあたり，修士課程・博士課程で約 8 年間にわたり，終始多大なる御指導ならびに御鞭撻を賜りました北陸先端科学技術大学院大学情報科学研究科　鵜木祐史 教授に深甚な謝意を表します．鵜木先生の熱血な御指導のおかげで，本論文を執筆するに至ることができました．

本研究を遂行するにあたり，熱心な御指導を賜りました北陸先端科学技術大学院大学　赤木正人 教授に深甚な感謝の意を表します．赤木先生からは，多くの温かな御助言をいただき，研究を発展することができました．

本研究を遂行するにあたり，熱心な御指導ならびに御助言を賜りました国立研究開発法人情報通信研究機構　Lu Xugang 博士に深甚な感謝の意を表します．

本論文のまとめ，ならびに副テーマの遂行にあたり，貴重な御助言ならびに御指導を賜りました北陸先端科学技術大学院大学 党建武 教授に心より感謝いたします．

本論文をまとめるにあたり，貴重な御助言ならびに御指導を賜りました北陸先端科学技術大学院大学　田中宏和 准教授に心より感謝いたします．

本研究を遂行するにあたり，日ごろから熱心に討論頂き，有益な御助言を賜りました金沢大学　三好正人 教授に心より感謝いたします．

筆者がドレスデン工科大学（ドイツ）に研究員として留学したとき，そして本研究を遂行するにたり，熱心な御指導ならびに御助言を賜りましたドレスデン工科大学 Ruediger Hoffmann 教授に心より感謝いたします．

筆者が米子工業高等専門学校在学中から，今日に至るまで終始ひとかたならぬ御指導と御教授を賜ったとともに，多大なる激励をいただきました米子工業高等専門学校電気情報工学科　浅倉邦彦 准教授に深甚な感謝の意を表します．

本研究を遂行するにあたり，日ごろから熱心に御討論いただき、また有益なる御助言を賜りました北陸先端科学技術大学院大学　宮内良太 助教，北陸先端科学

技術大学院大学　森川大輔 助教，札幌保健医療大学（前北陸先端科学技術大学院大学　助教）末光厚夫 准教授，群馬工業高等専門学校電子情報工学科（前北陸先端科学技術大学院大学　助教）川本真一 講師に心より感謝いたします.

筆者が多大なる励ましをいただきました米子工業高等専門学校電気情報工学科松原孝史 教授，北陸大学未来創造学部 木谷俊介 助教ならびに国立研究開発法人情報通信研究機構 岡本琢磨 博士に心より感謝いたします.

日頃より多大なる議論と激励をいただきました北陸先端科学技術大学院大学の諸先生方，赤木・鵜木研究室時代の諸先輩方，並びに赤木・鵜木研究室の諸氏に厚く御礼申し上げます.

最後に，私の研究生活を温かく見守ってくれた両親，祖母，妹に心より感謝いたします.

参考文献

[1] 松本希，“補聴器と人人工内耳：”聞くこと”を手に入れる，” 電子情報通信学会誌，Vol. 98，No. 4，pp. 272–278，2015.

[2] S. Nakagawa， T. Hotehama, and T. Kagomiya, “重度難聴者のための骨導超音波補聴器の開発：明瞭度および異聴解析による変調方式評価，” *Proc. Symposium on Ultrasonic Electronics*，pp. 5–7，2015.

[3] 植松 道冶，曽根 敏夫，二村 忠元，“ランダム変動騒音下の音声明瞭度と了解度に関する基礎実験：変動騒音の言語聴取妨害に関する研究　その 1，” 日本音響学会誌，Vol. 34，No. 9，pp. 516–521，Sept. 1978.

[4] 近藤 和弘，泉 良，藤森 雅也，加賀 類，中川 清司，“二者択一型日本語音声了解度試験方法の検討，” 日本音響学会誌，Vol. 63，No. 4，pp. 196–205，2007.

[5] M. Morimoto, H. Sato, and M. Kobayashi, “Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces，” *J. Acoust. Soc. Am.*, Vol. 116, No. 3, pp. 1607–1613, Sept. 2004.

[6] J.-C. Junqua and J.P. Haton, Robustness in automatic speech recognition, fundamentals and applications, Kluwer Academic Publishers, 1996.

[7] 飯田 茂隆，“明瞭度試験法について，” 日本音響学会誌，Vol. 43，No. 7，pp. 532–536，1987.

[8] 橋本 修，木村 翔，宇津木 淳一，“会話音声の発声レートを考慮した三連音節明瞭度試験洋音源による室内音場の明瞭度評価について，” 日本建築学会計画系論文，Vol. 456，pp. 1–8，Feb. 1994.

[9] 佐藤 洋，佐藤 逸人，吉野 博，鈴木 陽一，天野 成昭，近藤 公久，長友宗重，“単語親密度と加齢による聴力損失が残響及び雑音下における単語了解度に及ぼす影響,” 日本音響学会誌，Vol. 58，No. 6，pp. 346–354，2002.

[10] 佐藤 逸人，森本 政之，佐藤 洋，“聴き取りにくさによる音声伝達性能の評価,” 日本音響学会誌，Vol. 63，No. 5，pp. 275–280，2007.

[11] 佐藤 逸人，森本 政之，小吹 佳織，“住宅の居室における残響音が会話に与える影響,” 日本音響学会誌，Vol. 66，No. 11，pp. 541–551，2010.

[12] 中谷 智広，三好 正人，木下 慶介，“調波構造に基づくモノラル音声信号のブラインド 残響除去,” 電子情報通信学会論文誌，Vol. J88-D，No. 3，pp. 509–520，2005.

[13] 上羽 貞行，荒井 隆行，栗栖 清浩，倉片 憲治，坂本 真一，船場ひさお，佐藤 洋，“音バリアフリーの現状と課題,” 日本音響学会誌，Vol. 63，No. 12，pp. 723–730，2007.

[14] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, “Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments,” *Speech Communication*, Vol. 45, pp. 101–113, 2005.

[15] N. Hodoshima, T. Arai, A. Kusumoto, and K. Kinoshita, “Improving syllable identification by a preprocessing method reducing overlap-masuking in reverberant environments,” *J. Acoust. Soc. Am.,*, Vol. 119, pp. 4055–4064, 2006.

[16] 鹿島教昭，田村明弘，太田篤史，安藤祐子，鈴木和子，小澤繁之，“音声情報装置を用いた視覚障害者の歩行実験,” 横浜市環境科学研究報，Vol. 26，pp. 79–89，2002.

[17] M. Vondrasek and P. Pollak, “Methods for speech SNR estimation: evaluation tool and analysis of VAD dependency,” *Radioengineering*, Vol. 14, No. 1, pp. 6–11, 2005.

[18] R. Martin, "An efficient algorithm to estimate the instantaneous snr of speech signal," *Proc. EusroSpeech1993*, pp. 1093–1096, 1993.

[19] . Nemer, R. Goubran, and S. Mahmoud, "SNR estimation of speech signals using subbands and forth-order statistics," *IEEE Signal Processing Letter*, Vol. 6, No. 7, pp. 171–174, 1999.

[20] M. Kleinschmidt and V. Hohmann, "Sub-band SNR estimation using auditory feature processing," *Speech Communication*, Vol. 39, pp. 47–63, 2003.

[21] C. Kin and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," *Proc. Interspeech 2008*, pp. 2598–2601, 2008.

[22] A. Narayanan and D. Wang, "A CASA-based system for long-term SNR estimation," *IEEE Trans. Audio, Speech, and language processing*, Vol. 20, No. 9, pp. 2514–2527, 2012.

[23] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans.*, Vol. ASSP-26, No. 3, pp. 197–210, 1978.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans.*, Vol. ASSP-32, No. 6, pp. 1109–1121, 1984.

[25] K. Nakayama, S. Higashi, and A. Hirano, "a noise estimation method based on improved vad used in noise spectral suppression under highly non-stationary noise environments," *Proc. EUSIPCO 2009*, pp. 2494–2498, 2009.

[26] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to *a priori* SNR estimation," *IEEE Trans. on Speech and Language Processing*, Vol. 19, No. 1, pp. 186–195, 2001.

[27] S. Lee, C. Lim, and J.-H. Chang, "A new *a priori* snr estimator based on multiple linear regression technique for speech enhancement," *Digital Signal Processing*, Vol. 30, pp. 154–164, 2014.

[28] IEC 60268-16, "Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index," 2003.

[29] M. Martin, SPEECH AUDIOMETRY SECOND EDITION, London, Wwhurr Publishers Ltd, 1997.

[30] R. Ratnam, D. L. Jones, and W. D. O'Brien, "Fast algorithms for blind estimation of reverberation time," *IEEE Signal Processing Letters*, Vol. 11, No. 6, pp. 537–540, 2004.

[31] T. H. Falk, H. Yuan, and W. Chan, "Spectro-temporal processing for blind estimation of reverberation time and single-ended quality measurement of reverberant speech," *Proc. Interspeech2007*, pp. 514–517, 2007.

[32] R. Talmon and E. A. P. Habets, "Blind reverberation time estimation by intrinsic modeling of reverberant speech," *Proc. ICASSP2013*, pp. 156–160, 2013.

[33] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," *Proc. ICASSP2013*, pp. 161–165, 2013.

[34] F. F. Li and T. J. Cox, "Speech Transmission index from running speech: A neural network approach," *J. Acoust. Soc. Am.*, Vol. 113, pp. 1999–2008, 2003.

[35] P. Kendrick, T. J. Cox, Z. Yonggang, J. A. Chamber, and F. F. Li, "Room acoustic parameter extraction from music signals," *Proc. ICASSP2006*, Vol. 6, pp. 801–804, 2006.

[36] P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers, "Monaural room acoustic parameters from music and speech," *J. Acoust. Soc. Am.*, Vol. 124, No. 1, pp. 278–287, 2008.

[37] M. Unoki, K. Sakata, M. Furukawa, and M. Akagi, “A speech dereverberation method based on the MTF concept in power envelope restoration,” *Acoust. Sci. and Tech.*, Vol. 25, No. 4, pp. 243–254, 2004.

[38] M. Unoki and S. Hiramatsu, “Mtf-based method of blind estimation of reverberation time in room acoustics,” *Proc. EUSIPCO 2008*, pp. CD–ROM, 2008.

[39] M. Unoki, K. Sasaki, R. Miyauchi, M. Akagi, and N. S. Kim, “Blind method of estimating speech transmission idex from reverberant speech signals,” *Proc. EUSIPCO 2013*, pp. CD–ROM, 2013.

[40] J. Jo and M. Koyasu, “Measurement of reverberation time based on the direct-reverberant sound energy ratio in steady state,” *Proc. InterNoise75*, pp. 579–582, 1975.

[41] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, “Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model,” *IEEE Trans. Audio, Speech, and language processing*, Vol. 19, No. 8, pp. 2374–2384, 2011.

[42] T. M. Prego, A. A. Lima, R. Zambrano-Lopez, and S. L. Netto, “Blind estimators for reverberation time and direct-toreverberant energy ratio using subband speech decomposition,” *Proc. WASPAA 2015*, pp. 1–5, 2015.

[43] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, “The ACE challenge Corpus description and performance evaluation,” *Proc. WASPAA 2015*, pp. 1–5, 2015.

[44] ハインリッヒ・クットルフ，室内音響学 建築の響きとその理論，市ヶ谷出版社，2003.

[45] T. Houtgast and H. J. M. Steeneken, “The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility,” *Acustica*, Vol. 28, pp. 66–73, 1973.

[46] T. Houtgast, H. J. M. Steeneken, and R. Plomp, “Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function. I. General Room Acoustics,” *Acustica*, Vol. 46, pp. 60–72, 1980.

[47] T. Houtgast and H. J. M. Steeneken, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” *J. Acoust. Soc. Am.*, Vol. 77, pp. 1069–1077, 1985.

[48] P. Larn and V. Hongisto, “Experimental comparison between speech transmission index, rapid speech transmission index, and speech intelligibility index,” *J. Acoust. Soc. Am.*, Vol. 119, No. 2, pp. 1106–1117, 2006.

[49] 宮崎 晃和，森田 翔太，鵜木 祐史，“実環境における音声伝送指標のブラインド推定法の検討,” 音講論，pp. 785–788，March 2014.

[50] S. F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” *IEEE Trans.*, Vol. ASSP-27, No. 2, pp. 113–120, 1979.

[51] Z. Goh, K. Tan, and B. T. G. Tan, “Postprocessing Method for Suppressing Musical noise generated by Spectral Subtraction,” *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 3, pp. 287–292, 1998.

[52] 北岡 教英，赤堀 一郎，中川 聖一，“スペクトルサブトラクションと時間方向スムージングを用いた雑音環境下音声認識,” 電子情報通信学会論文誌，Vol. J83-D，No. 2，pp. 500–508，Feb. 2000.

[53] M. R. Sambur, “Adaptive Noise Canceling for Speech Signals,” *IEEE Trans.*, Vol. ASSP-26, No. 5, pp. 419–423, 1978.

[54] W. A. Harrison, J. S. LIM, and E. Singer, “A New Application of Adaptive Noise Cancellation,” *IEEE Trans.*, Vol. ASSP-34, No. 1, pp. 21–27, 1986.

[55] J. E. Greenberg, “Modified LMS Algorithms for Speech Processing with an Adaptive Noise Canceller,” *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 4, pp. 338–351, 1998.

[56] 片山 徹，応用カルマンフィルタ，朝倉書店，2000.

[57] 藤本 雅清，有木 康雄，“カルマンフィルタに基づく音声信号推定法を用いた雑音環境下での音声認識,” 電子情報通信学会論文誌，Vol. J58-D-II，No. 1，pp. 1–11，Jan. 2002.

[58] A. Nower, Y. Liu, and M. Unoki, “Restoration scheme of instantaneous amplitude and phase using Kalman filter with efficient linear prediction for speech enhancement,” *Speech Communication*, Vol. 70, pp. 13–27, 2015.

[59] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans.*, Vol. ASSP-28, No. 2, pp. 137–145, 1980.

[60] 古屋 武志，金田 圭一，五反田 博，“ブラインド信号分離による雑音除去法のSN比改善量,” 電子情報通信学会論文誌，Vol. J87-A，No. 7，pp. 1054–1058，2004.

[61] 高橋 祐，高谷 智哉，猿渡 洋，鹿野 清宏，“独立成分分析に基づく空間的サブトラクションアレーによる雑音抑圧,” 電子情報通信学会 技術研究報告 EA，Vol. 106，No. 125，pp. 13–18，2006.

[62] H. Hermansky and N. Morgan, “RASTA Processing of Speech,” *IEEE Trans. Speech Audio Process.*, Vol. 2, No. 4, pp. 578–586, 1994.

[63] S. T. Neely and J. B. Allen, “Invertibility of a room impulse response,” *J. Acoust. Soc. Am.*, Vol. 66, No. 1, pp. 165–169, 1979.

[64] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustics,” *IEEE Trans.*, Vol. ASSP-36, pp. 145–152, 1988.

[65] 古家 賢一，片岡 章俊，“チャネル間相関行列と音声の白色化フィルタを用いたSemi-blind 残響抑圧,” 電子情報通信学会論文誌，Vol. J88-A，No. 10，pp. 1089–1099，Oct. 2005.

[66] K. Furuya and A. Kataoka, "Robust Speech Dereverberation Using Multichannel Blind Deconvolution With Spectral Subtraction," *IEEE Trans. Audio, Speech, and language processing*, Vol. 15, No. 5, pp. 1579–1591, 2007.

[67] 古家 賢一，片岡 章俊，"残響抑圧処理による音声品質改善効果の要因分析，" 電子情報通信学会論文誌，Vol. J91-A，No. 8，pp. 763–771，2008.

[68] H. Wang and F. Itakura, "Realization of acoustic inverse filtering through multi-microphone sub-band processing," *IEICE Trans. Fundamanetals*, Vol. E75-A, pp. 1474–1483, 1992.

[69] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, and language processing*, Vol. 17, No. 4, pp. 534–545, 2009.

[70] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Evaluation of separation and dereverberation performance in frequency domain blind source separation," *Acoust. Sci. and Tech.*, Vol. 25, No. 2, pp. 119–126, 2004.

[71] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Harmonicity Based Dereverberation for Improving Automatic Speech Recognition Performance and Speech Intelligibility," *IEICE Trans. Fundamentals*, Vol. E88-A, No. 7, pp. 1724–1731, 2005.

[72] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-Based Blind Dereverberation for Single-Channel Speech Signals," *IEEE Trans. Audio. Speech, and Lancuage Processing*, Vol. 15, No. 1, pp. 80–95, 2007.

[73] B. E.D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Soeech Communication*, Vol. 25, No. 1-3, pp. 117–132, 1998.

[74] R. Gomez and T. Kawahara, "Optimizing spectral subtraction and wiener filtering for robust speech recognition in reverberant and noisy condition,"

Proc. ICASSP2010, pp. 4566–4569, 2010.

[75] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Multi-step linear prediction based speech enhancement in noisy reverberant environment," *Proc. Interspeech-2007*, pp. 854–857, 2007.

[76] Y. Liu, N. Nower, S. Morita, and M. Unoki, "Speech enhancement of instantaneous amplitude and phases for application in noisy reverberant environments," *Speech Communication*, Vol. 84, pp. 1–14, 2016.

[77] 水町 光徳，赤木 正人，"マイクロホン対を用いたスペクトルサブトラクションによる雑音除去法," 電子情報通信学会論文誌，Vol. J82-A，No. 4，pp. 503–512，1999.

[78] 伊藤 憲三，水島 昭英，北脇 信彦，"音声と非音声の識別処理に基づく定常雑音抑圧方式," 日本音響学会誌，Vol. 61，No. 8，pp. 431–440，2005.

[79] S. Das, E. Hamid, K. Hirose, and K. I. Molla, "Single-channel speech enhancement by NWNS and EMD," *Signal Processing: An International Journal (SPIJ)*, Vol. 4, No. 5, pp. 279–291, 2011.

[80] Y. Zhang and Y. Zhao, "Spectral subtraction on real and imaginary modulation spectra," *Proc. ICASSP2011*, pp. 4744–4747, 2011.

[81] 加藤 正徳，杉山 昭彦，芹沢 昌宏，"重み付き雑音推定と MMSE STSA 法に基づく高音質雑音抑圧," 電子情報通信学会論文誌，Vol. J87-A，No. 7，pp. 851–860，July 2004.

[82] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *IEEE Proc.*, Vol. 67, No. 12, pp. 1586–1604, Dec. 1979.

[83] 中谷 智広，奥乃 博，"音オントロジーに基づいた音環境理解システムの統合," 人工知能学会誌，Vol. 14，No. 6，p.11，1999.

[84] 鵜木 祐史，赤木 正人，"雑音が付加された波形からの信号波形の一抽出法," 電子情報通信学会論文誌，Vol. J80-A，No. 3，pp. 444–453，1997.

[85] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, Vol. 27, pp. 261–279, 1999.

[86] 鵜木 祐史，赤木 正人，"聴覚の情景解析に基づいた雑音下の調波複合音の一抽出法," 電子情報通信学会論文誌，Vol. J82-A，No. 10，pp. 1497–1507，1999.

[87] K. Kinoshita, M. Delcoix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech," *Proc. WASPAA 2013*, pp. 1–4, 2013.

[88] http://reverb2014.dereverberation.com/ 2014.

[89] 木下 慶介，デルクロア マーク，吉岡 拓也，中谷 智広，"REVERB challenge(残響下音声強調・認識チャレンジ)：企画内容と結果報告," 音講論，pp. 655–658，Sept. 2014.

[90] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, Vol. 27, pp. 621–633, 2013.

[91] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Communication*, Vol. 50, No. 3, pp. 244–263, 2008.

[92] T. Yoshioka, A. Sehr, M. Delcoix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 114–126, Nov. 2012.

[93] 石塚 健太郎，藤本 雅清，中谷 智広，"音声区間検出技術の最近の研究動向," 日本音響学会誌，Vol. 65，No. 10，pp. 537–543，2009.

[94] 藤本 雅清，“音声区間検出の基礎と世界的な研究動向，今後の展開,” 電子情報通信学会誌，Vol. 95，No. 8，pp. 754–758，2012.

[95] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, “CENSREC-1C: An evaluation framework for voice activity detection under noisy environments,” *Acoust. Sci. and Tech.*, Vol. 30, No. 5, pp. 363–371, 2009.

[96] A. Benyassine, E. Shlomot, S. Huan-yu, D. Massaloux, C. Lambin, and J. P. Petit, “ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application,” *IEEE Commun. Mag.*, Vol. 35, pp. 64–73, 1997.

[97] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, “Noise robust voice activity detection based on periodic to aperiodic component ratio,” *Speech Communication*, Vol. 52, pp. 41–60, 2010.

[98] J. Ramirez, J. C. Segura, C. Benitez, A. D. L. Torre, and A. Rubio, “Efficient voice activity detection algorithms using long-term speech information,” *Speech Communication*, Vol. 42, pp. 271–287, 2004.

[99] P. K. Ghosh, A. Tsiartas, and S. Narayanan, “Robust voice activity detection using long-term signal variability,” *IEEE Trans. Audio, Speech, and language processing*, Vol. 19, No. 3, pp. 600–613, 2011.

[100] S.-K. Kim and J.-H. Chang, “Voice activity detection based on conditional MAP criterion incorporating the spectral gradient,” *Signal processing*, Vol. 92, pp. 1699–1705, 2012.

[101] ETSI EN 301 v7.1, “Digital cellular telecommunications system: Voice Activity Detector (VAD) for adaptive multi-rate (AMR) speech traffic channels,” 1999.

[102] N. Megarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," *Proc. ICASSP'04*, Vol. 1, pp. 601–604, 2004.

[103] K. Pek, T. Arai, and N. Kanedera, "Voice activity detection in noise using modulation spectrum of speech: Investigation of speech frequency and modulation frequency ranges," *Acoust. Sci. and Tech.*, Vol. 33, No. 1, pp. 33–44, 2012.

[104] Y. Kanai, S. Morita, and M. Unoki, "Concurrent processing of voice activity detection and noise reduction using empirical mode decomposition and modulation spectrum analysis," *Proc. Interspeech 2011*, pp. 742–746, 2013.

[105] W. Shi and Y Zou, "A novel instantaneous frequency based voice activity detection for strong noisy speech," *Proc. International Conference on Information and Automation*, pp. 956–959, 2012.

[106] S.-H. Chen, H.-T. Wu, Y. Chang, and T. K. Truong, "Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator," *Pattern Recognition Letter*, Vol. 28, pp. 1327–1332, 2007.

[107] T. V. Pham, M. Stark, and E. Rank, "Performance analysis of wavelet subband based voice activity detection in cocktail party environment," *Proc. The 2010 International Conference on Advanced Technologies for Communications*, pp. 85–88, 2010.

[108] M. Eshaghi and K. Mollaei, "Voice activity detection based on using wavelet packet," *Digital Signal Processing*, Vol. 20, pp. 1102–1115, 2010.

[109] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans.*, Vol. SMC-9, No. 1, pp. 62–66, 1979.

[110] A. Torre, J. Ramirez, C. Benitez, J. C. Segura, L. Garcia, and A. J. Rubio, "Noise robust model-based voice activity detection," *Proc. Interspeech-2006*, pp. 1954–1957, 2006.

[111] D. Ying, Y. Shi, X. Lu, J. Dang, and F. Soong, "Robust voice activity detection based on noise eigenspace," *Acoust. Sci. and Tech.*, Vol. 28, No. 6, pp. 413–423, 2007.

[112] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-Term Spectro-Temporal and Static Harmonic Features for Voice Activity Detection," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, No. 5, pp. 834–844, 2010.

[113] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Trans. Audio, Speech, and language processing*, Vol. 19, No. 8, pp. 2624–2633, 2011.

[114] N. Cho and E.-K. Kim, "Enhanced voice activity detection using acoustic event detection and classification," *IEEE Trans. on consumer electronics*, Vol. 57, No. 1, pp. 196–202, 2011.

[115] D. Vlaj, Z. Kacic, and M. Kos, "Voice activity detection algorithm using nonlinear spectral weights, hangover and hangbefore criteria," *Computers and Electrical Engineering*, Vol. 38, pp. 1820–1836, 2012.

[116] B. Mak, J.-C. Junqua, and B. Reaves, "A robust speech/non-speech detection algorithm using time and frequency-based features," *Proc. ICASSP'92*, Vol. I, pp. 269–272, 1992.

[117] S. Basu, "A linked-HMM model for robust voicing and speech detection," *Proc. ICASSP'03*, Vol. 1, pp. 816–819, 2003.

[118] O. Varela, R. San-Segimdp, and L. Hernandez, "Combining pulse-based features for rejecting far-field speech in HMM-based voice activity detector," *Computers and Electrical Engineering*, Vol. 37, pp. 589–600, 2011.

[119] H. Veisi and H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement," IET Signal Processing, Vol. 6, No. 1, pp. 54–63, 2012.

[120] S. Shafiee, F. Almasganj, B. Vazirnezhad, and A. Jafari, "A two-stage speech activity detection system considering fractal aspects of prosody," *Pattern Recognition Letter*, Vol. 31, pp. 936–948, 2010.

[121] S.-H. Chen, R. C. Guido, T.-K. Truong, and Y. Chang, "Improved voice activity detection algorithm using wavelet and support vector machine," *Computer Speech and Language*, Vol. 24, pp. 531–543, 2010.

[122] J, Wu and X.-L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Processing Letter*, Vol. 18, No. 8, pp. 466–469, Aug. 2011.

[123] M. Farsinejad and M. Analoui, "A new robust voice activity detection method based on genetic algorithm," *Proc. Telecommnication network and applications conference*, pp. 80–84, 2008.

[124] J. W. Shin, J.-H. Chang, and N. S. Kim, "Voice activity detection based on a family of parametric distributions," *Pattern Recognition Letter*, Vol. 28, pp. 1295–1299, 2007.

[125] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letter*, Vol. 6, No. 1, pp. 1–3, Jan. 1999.

[126] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and adaptive threshold," *IEEE Trans. on Speech and Audio Processing*, Vol. 14, No. 2, pp. 412–424, 2006.

[127] J. W. Shin, H. J. Kwon, S. H. Jin, and N. S. Kim, "Voice activity detection based on conditional MAP criterion," *IEEE Signal Processing Letter*, Vol. 15, pp. 257–260, 2008.

[128] Y. Suh and H. Kim, "Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection," *IEEE Signal Processing Letter*, Vol. 19, No. 8, pp. 507–510, Aug. 2012.

[129] S.-W. Deng and J.-Q. Han, "Statistical voice activity detection based on sparse representation over learned dictionary," *Digital Signal Processing*, Vol. 23, pp. 1228–1232, 2013.

[130] X. Lu, M. Unoki, R. Isotani, H. Kawai, and S. Nakamura, "Adaptive regularization framework for robust voice activity detection," *Proc. Interspeech 2011*, pp. 2653–2656, Aug. 2011.

[131] K. Ishizuka and H. Kato, "A feature for voice activity detection derived from speech analysis with the exponential autoregressive model," *Proc. ICASSP'06*, Vol. I, pp. 789–792, 2006.

[132] H. S, Kato, K. Ishizuka, and M. Fujimoto, "Voice activity detection based on adjustable linear prediction and GARCH models," *Speech Communication*, Vol. 50, pp. 476–486, 2008.

[133] T.Petsatodis and C. Boukis, "Efficient voice activity detection in reverberant enclosure using far field microphones," *Proc. International conference on digital signal processing*, pp. 1–5, 2009.

[134] J. E. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani, and M. Fujimoto, "Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates," *Proc. ICASSP'07*, Vol. 4, pp. 385–388, 2007.

[135] H. Lee and D. Yook, "Space-Time Voice Activity Detection," *IEEE Trans. on Consumer Electronics*, Vol. 55, No. 3, pp. 1471–1476, 2009.

[136] M. Unoki, X. Lu, R. Petrick, S. Morita, M. Akagi, and R. Hoffmann, "Voice activity detection in MTF-based power envelope restoration," *Proc. Interspeech 2011*, pp. 2609–2612, Aug. 2011.

[137] NICT, "VoiceTra". http://voicetra.nict.go.jp/.

[138] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am*, Vol. 95, No. 2, pp. 1053–1064, Feb. 1994.

[139] 小椋 靖夫，浜田 晴夫，三浦 種敏，“音場における音声伝達品質のための MTF と STI,” 日本音響学会誌，Vol. 40，No. 3，pp. 181–191，1984.

[140] 戸井田 義徳，“小特集-音声の明瞭度と認識率-空間内における音情報伝達,” 日本音響学会誌，Vol. 51，No. 4，pp. 312–316，1995.

[141] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, “Syllable intelligibility for temporally filtered LPC cepstral trajectories,” *J. Acoust. Soc. Am.*, Vol. 105, No. 5, pp. 2783–2791, 1999.

[142] M.R. Schroeder, “Modulation transfer functions: definition and measurement,” *Acustica*, Vol. 49, pp. 179–182, 1981.

[143] A. Papouris, Probability, random variables and stochastic processes, 3 edition, New York, MacGraw-Hill, Inc., 1991.

[144] M. Unoki, Y. Yamasaki, and M. Akagi, “MTF-BASED POWER ENVELOPE RESTORATION IN NOISY REVERBERANT ENVIRONMENTS,” *Proc. EUSIPCO 2009*, pp. 228–232, 2009.

[145] S. Morita, M. Unoki, X. Lu, and M. Akagi, “Robust Voice Activity Detection Based on the Concept of Modulation Transfer Function in Noisy Reverberant Environments,” *Journal of Signal Processing Systems*, Vol. 82, No. 2, pp. 163–173, 2016.

[146] http://www.slp.cs.tut.ac.jp/CENSREC/ja/CENSREC/AURORA-2J/ 2012.

[147] 北岡 教英，中村 哲，“雑音下音声認識評価基盤 CENSREC,” 日本音響学会誌，Vol. 68，No. 16，pp. 305–310，2012.

[148] 金寺 登，荒井 隆行，船田 哲男，“変調スペクトルの重要な成分のみを選択的に用いた雑音に強い音声認識,” 電子情報通信学会論文誌，Vol. J84-D-II, No. 7，pp. 1261–1269，2001.

[149] Y. Yamasaki and M. Unoki, “Study on a method of suppressing noise based on the MTF concept,” *J. Signal Processing*, Vol. 13, No. 4, pp. 335–338, 2009.

[150] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, “An improved method based on the MTF concept for restoring the power envelope from reverberant signal,” *Acoust. Sci. and Tech.*, Vol. 25, No. 4, pp. 232–242, 2004.

[151] X. Lu, M. Unoki, and M. Akagi, “Comparative evaluation of modulation-transfer-function-based blind restoration of sub-band power envelopes of speech as a front-end processor for automatic speech recognition systems,” *Acoust. Sci. and Tech.*, Vol. 29, No. 6, pp. 351–361, 2008.

[152] S. Morita, X. Lu, M. Unoki, and M. Akagi, “Signal to noise ratio estimation based on an optimal design of subband voice activity detection,” *Proc. ISCSLP2014*, pp. 560–564, 2014.

[153] 日本建築学会，建築と環境のサウンドライブラリ，技報堂出版，2004.

[154] K. Kawai, K. Fujimoto, T. Iwase, H. Yasuoka, T. Sakuma, and Y. Hidaka, “Developement of a sound source database for environmental/architectual acoustics: Instroduction of SMILE 2004 (Sound Material in Living Environment 2004),” *Proc. ICA*, pp. 1561–1564, 2004.

[155] A. Varga and H.J.M. Steeneken, “Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, Vol. 12, No. 13, pp. 247–251, 1993.

[156] S. Morita, S. Lu, M. Unoki, M. Akagi, and R. Hoffmann, “MTF-based subband power-envelope restoration for robust speech recognition in noisy reverberant environments,” *Proc. APSIPA2011*, pp. CD–ROM, 2011.

[157] Cambridge University Engineering Department, “The HTK book (version 3.2),” 2002.

[158] ETSI EN 202 050 v1.1.5, "Speech processing, transmission and quality aspects (STQ), distributed speech recognition, advanced front-end feature extraction algorithm; compression algorithms, ETSI standard," 2007.

[159] F.H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient Cepstral normalization for robust speech recognition," *Proc. ARPA speech and natural language workshop*, pp. 69–74, 1993.

[160] H. Hermansky, N. Morgan, and H.G. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *Proc. ICASSP'93*, Vol. I, pp. 83–86, 1993.

[161] A. Miyazaki, S. Morita, and M. Unoki, "Study on blind method of estimating speech transmission index from noisy reverberant amplitude-modulated-signal," *Journal of Signal Processing*, Vol. 18, No. 4, pp. 201–204, 2014.

[162] M. Unoki, S. Morita, A. Miyazaki, and M. Akagi, "Preliminary study on blind estimation of room acoustic parameters in noisy reverberant environments," *Proc. WESPAC2015*, pp. 428–435, 2015.

[163] Z. Zhu, Y. Nishino, R. Miyauchi, and M. Unoki, "Study on linguistic information and speaker individuality contained in temporal envelope of speech," *Acoustical Science and Technology*, Vol. 37, No. 5, pp. 258–261, 2016.

[164] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants," *Proc. Interspeech2016*, pp. 262–266, 2016.

[165] M. Unoki, M. Toi, and M. Akagi, "Refinement of an MTF-based speech dereverberation method using an optimal inverse-MTF filter," *Proc. SPECOM'06*, Vol. 1, pp. 323–326, 2006.

本研究に関する研究業績

論文

[1] Shota Morita, Masashi Unoki, and Masato Akagi, “A study on the IMTF-based filtering on the modulation spectrum of reverberant signal,” Journal of Signal Processing, Vol. 14, No. 4, pp. 269–272, July 2010.

[2] Akikazu Miyazaki, Shota Morita, and Masashi Unoki, “Study on blind method of estimating speech transmission index from noisy reverberant amplitude-modulated-signals,” Journal of Signal Processing, Vol. 18, No. 4, pp. 201–204, 2014.

[3] Shota Morita, Masashi Unoki, Xugang Lu, and Masato Akagi, “Robust Voice Activity Detection Based on the Concept of Modulation Transfer Function in Noisy Reverberant Environments,” Journal of Signal Processing Systems, Vol. 82(2), pp. 163–173, February 2016.

国際会議

[4] Shota Morita, Masashi Unoki, and Masato Akagi, “A study on the IMTF-based filtering on the modulation spectrum of reverberant signal,” in Proc. NCSP’10, pp. 265–268, Hawaii, USA, March 2010 (CD-ROM).

[5] Shota Morita, Xugang Lu, Masashi Unoki, and Masato Akagi, “Study on MTF-based power envelope restoration in noisy reverberant environments,” in Proc. NCSP’11, pp. 247–250, Tianjin, China, March 2011 (CD-ROM).

[6] Shota Morita, Xugang Lu, Masashi Unoki, Masato Akagi, Ruediger Hoffmann, “MTF-based sub-band power-envelope restoration for robust speech recognition in noisy reverberant environments,” in Proc. APSIPA2011, Xi’an, October 2011 (CD-ROM).

[7] Shota Morita, Xugang Lu, Masashi Unoki, Masato Akagi, Ruediger Hoffmann, “A modulation-transfer-function-based method for restoring sub-band power envelope from noisy reverberant speech,” in Proc. Acoustics2012, Hong-Kong, China, May 2012.

[8] Shota Morita, Masashi Unoki, Xugang Lu, and Masato Akagi, “Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments,” in Proc. ISCSLP2014, pp. 108–112, Singapore, September 2014.

[9] Shota Morita, Xugang Lu, Masashi Unoki, and Masato Akagi, “Signal to noise ratio estimation based on an optimal design of subband voice activity detection,” in Proc. ISCSLP2014, pp. 560–564, Singapore, September 2014.

研究会

[10] 森田 翔太, 山崎 悠, 鵜木 祐史, 赤木 正人, “雑音残響環境下における MTF に基づくパワーエンベロープ回復処理の検討,” 信学技報 EA, Vol. 110, No. 239, pp. 121–126, October 2010.

[11] 森田 翔太, Lu Xugang, 鵜木 祐史, 赤木 正人, Hoffmann Ruediger, “雑音残響環境下での変調伝達関数に基づくパワーエンベロープ回復処理と音声認識への応用,” 信学技報 EA, Vol. 111, No. 26, pp. 37–42, May 2011.

[12] 森田 翔太, 鵜木 祐史, ルー シュガン, 赤木 正人, “変調伝達関数に基づく雑音残響に頑健な音声区間検出法～帯域分割 SNR 推定法の利用～,” 信学技報 SP, Vol. 114, No. 52, pp. 383–388, May 2014.

[13] 森田 翔太, ルー シュガン, 鵜木 祐史, “帯域分割型音声区間検出法の最適設計に基づく信号対雑音比推定法の検討,” 信学技報 EA, Vol. 114, No. 358, pp. 37–42, December 2014.

口頭発表

[14] 森田 翔太, 鵜木 祐史, 赤木 正人, “変調伝達関数に基づいた変調スペクトル逆フィルタ処理の検討,” 音講論, 2-P-21, pp. 837–840, March 2010.

[15] 森田 翔太, Lu Xugang, 鵜木 祐史, 赤木 正人, “雑音残響にロバストな音声認識のための帯域分割型パワーエンベロープ回復処理の検討,” 音講論, 2-P-26, pp. 163–166, March 2011.

[16] 森田 翔太, 鵜木 祐史, ル シュガン, “MTF に基づくパワーエンベロープ回復における統合的な雑音残響除去の検討,” 音講論, 3-P-4, pp. 123–126, September 2012.

[17] 森田 翔太, 鵜木 祐史, 赤木 正人, “雑音残響にロバストな音声区間検出法の検討,” 音講論, 1-P-22b, pp. 155–158, September 2013.

[18] 森田 翔太, 鵜木 祐史, ルー シュガン, 赤木 正人, “変調伝達関数に基づく雑音残響に頑健な音声区間検出法の検討,” 第 28 回信号処理シンポジウム, pp. 614–619, November 2013.

[19] 森田 翔太, 鵜木 祐史, ルー シュガン, 赤木 正人, “PRISM の総合評価：PRISM をフロントエンドとした音声認識性能,” 音講論, 3-Q5-25, pp. 249–252, March 2014.

その他の研究業績

論文

[1] Yang Liu, Shota Morita, and Masashi Unoki, “MTF-based Kalman filtering with linear prediction for power envelope restoration in noisy reverberant environments,” IEICE Transaction Fundamentals, A, Vol. E99-A, No. 2, pp. 560–569, February 2016.

[2] Yang Liu, Naushin Nower, Shota Morita, and Masashi Unoki, “Speech enhancement of instantneous amplitude and phases for application in noisy reverberant environments,” Speech Communication, Vol. 84, pp. 1–14, November 2016.

国際会議

[3] Masashi Unoki, Xugang Lu, Rico Petrick, Shota Morita, Masato Akagi, and Ruediger Hoffmann, “Voice activity detection in MTF-based power envelope restoration,” in Proc. Interspeech2011, pp. 2609–2612, Florence, Italy, August 2011.

[4] Yasuaki Kanai, Shota Morita, and Masashi Unoki, “Concurrent Processing of voice activity detection and noise reduction using empirical mode decomposition and modulation spectrum analysis,” in Proc. Interspeech2013, pp. 742–746, Lyon, France, August 2013.

[5] Akikazu Miyazaki, Shota Morita, and Masashi Unoki, “Study on blind method of estimating speech transmission index from noisy reverberant amplitude-modulated-signal,” in Proc. NCSP’14, pp. 105–108, Hawaii, USA, March 2014.

[6] Masashi Unoki, Shota Morita, Akikazu Miyazaki, and Masato Akagi, “Preliminary study on blind estimation of room acoustic parameters in noisy reverberant environments,” in Proc. WESPAC2015, pp. 428–435, December 2015.

[7] Yang Liu, Naushin Nower, Shota Morita, and Masashi Unoki, “Robust front-end for speech recognition by human and machine in noisy reverberant environments: the effect of phase information,” in Proc. ISCSLP 2016, pp. 17–20, October 2016.

研究会

[8] 鵜木 祐史, ル シュガン, ペトリック リコ, 森田 翔太, 赤木 正人, ホフマン ルディィガー, “変調伝達関数に基づいたパワーエンベロープ回復処理における音声区間検出の検討,” 信学技報 EA, Vol. 112, No. 47, pp. 7–12, May 2012.

[9] 鵜木 祐史, ル シュガン, 森田 翔太, “MTF に基づくパワーエンベロープ回復処理における統合的な残響除去法,” 信学技報 EA, Vol. 112, No. 292, pp. 29–34, November 2012.

[10] 金井 康昭, 森田 翔太, 鵜木 祐史, “経験的モード分解と変調スペクトルを用いた音声区間検出と雑音残響の同時処理,” 信学技報 EA, Vol. 112, No. 27, pp. 145–150, May 2013.

[11] 宮崎 晃和, 森田 翔太, 鵜木 祐史, “背景雑音を考慮した音声伝送指標のブラインド推定法の検討,” 信学技報 EA, Vol. 113, No. 349, pp. 1–6, December 2013.

[12] 鵜木 祐史，森田 翔太，宮崎 晃和，赤木 正人，“雑音残響環境における音声伝送指標の推定と音声回復処理,” 建築研究会資料，Vol. 45，No. 5，pp. 449–454, July 2015.

口頭発表

[13] 鵜木 祐史, 森田 翔太, 澤口 知希, 赤木 正人, “MTF に基づいたパワーエンベロープ回復処理による音声区間検出の検討,” 音講論, 2-P-16, pp. 187–190, March 2011.

[14] 鵜木 祐史, 森田 翔太, ルー シュガン, 赤木 正人, “残響にロバストな音声区間検出法とその比較評価,” 音講論, 3-P-1, pp. 143–146, September 2011.

[15] 金井 康昭, 森田 翔太, 鵜木 祐史, “経験的モード分解と変調スペクトルを用いた音声区間検出と雑音除去の同時処理の検討,” 音講論, 1-8-6, pp.19–22, September 2013.

[16] 宮崎 晃和, 森田 翔太, 鵜木 祐史, “実環境における音声伝送指標のブラインド推定法の検討,” 音講論, 1-P4-18, pp. 785–788, March 2014.

[17] 鵜木 祐史，森田 翔太，宮崎 晃和，赤木 正人，“雑音残響環境における室内音響パラメータのブラインド推定法の検討,” 音講論，1-6-11，pp. 539–542, September 2015.