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

Title	経験的モード分解と変調スペクトルを用いたロバストな音声区間検出法に関する研究
Author(s)	金井，康昭
Citation	
Issue Date	2012-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/10440
Rights	
Description	Supervisor:鵜木祐史，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修 士 論 文

経験的モード分解と変調スペクトルを用いたロバストな音声区間検出法に関する研究

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

金井 康昭

2012 年 3 月

修 士 論 文

経験的モード分解と変調スペクトルを用いたロバストな音声区間検出法に関する研究

指導教員　鵜木 祐史 准教授

審査委員主査　鵜木 祐史 准教授
審査委員　赤木 正人 教授
審査委員　党 建武 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

0910013 金井 康昭

提出年月: 2012 年 2 月

Copyright © 2012 by kanai Yasuaki

概 要

音声区間検出（Voice Activity Detection, VAD）とは，解析信号から音声の存在する区間（音声区間）と音声の存在しない区間（非音声区間）を自動的に検出する技術である [1]．VAD 自体は要素技術であるが，その応用範囲は，音声符号化技術や雑音抑圧技術，音声認識技術への応用など多種多様である [2]．そのため，実用的な VAD を考えると，対象音声以外の音の混入や，背景雑音がある状況でも，正確に解析信号から音声区間を検出できる必要がある．

古典的な VAD 法は，信号のパワーの閾値処理やゼロ交差数を利用して行われる [1]．この手法は，観測信号が対象音声のみをであれば，ほぼ完全に音声区間を検出できるが，対象音声以外の成分（雑音成分）が混入すると検出精度の著しい低下を招くという問題がある．実際問題として，VAD を活用する様々なアプリケーションは雑音のある環境や対象音声以外の非音声信号が混在する環境で利用されるため，実環境で VAD を活用するためには利用環境や対象信号に依存せずに正確に音声区間を検出できなければならない．

近年，これらの問題を解決するために，閾値の決定に統計的手法を用いる Otsu の閾値法 [3] や複数のパラメータ（スペクトル歪み，全帯域のエネルギー，低帯域のエネルギー，ゼロ交差数の４種類）を音声／非音声の判別に利用する G.729 Annex B の VAD 法 [4] や音声の非線形振動特性に着目した Exponentioal 自己回帰モデルを用いた手法 [6] など，実環境を想定した VAD 法が提案されている．これらの方法は，雑音環境でも高精度な音声区間検出が行えるが，音声と非音声を区別して検出する能力は弱い．そのため，非音声信号が存在する環境では，高精度な VAD を行うことはできない．

本研究では，実環境で正確な音声区間検出を行うために，背景雑音の混入，および，目的音声以外の非音声信号の混入に対して，ロバストな VAD 法の実現を目的とする．そのために，経験的モード分解（Empirical Mode Decomposition, EMD）[7] を利用して定常雑音の除去を行い，変調スペクトル分析（Modulation Spectrum Analysis, MSA）[8] を用いて得られた信号区間の音声／非音声判別を行う，二段階処理からなるロバスト VAD 法を提案する．

評価の結果，提案法は，すべての実験条件において，約 90 % 程度の高い正答率を得た．クリーンな環境において，提案法は従来法に劣らぬ精度で音声区間を正確に検出できた．また，背景雑音環境，非音声混入環境，そして，背景雑音と非音声寝具が同時に存在する環境において，クリーンな環境での VAD の正答率と同程度の正答率で音声区間を検出できることが分かった．

以上の結果から，EMD による定常雑音の除去と MSA による変調スペクトル上の特徴判別を組み合わせて利用することで，雑音環境下で非音声信号を含む場合でも高精度でロバストな音声区間検出を実現できた．

目 次

第 1 章　序論 **1**
1.1　はじめに 1
1.2　研究の背景 1
1.3　本研究で取り組むべき課題 2
1.4　本研究の目的 2
1.5　本論文の構成 3
1.6　記号の定義 4

第 2 章　従来の研究 **5**
2.1　古典的手法 5
2.2　近代的手法 5
2.2.1　Otsu の閾値法 6
2.2.2　G.729 Annex B の VAD 法 6
2.2.3　Exponentioal 自己回帰モデルを用いた手法 7
2.3　従来法の問題点 7
2.4　解決すべき課題 7

第 3 章　背景雑音と非音声信号の混入にロバストな VAD **8**
3.1　EMD と MSA を用いた VAD 8
3.2　定常／非定常 IMF の切り分け 9
3.3　音声／非音声の変調スペクトルの判別 12

第 4 章　経験的モード分解（EMD） **13**
4.1　EMD とは 13
4.2　固有モード関数 13
4.3　EMD の信号分解プロセス 13
4.4　EMD を用いた音信号処理 18
4.4.1　Taufiq らの雑音除去法 18
4.4.2　Molla & Hirose の雑音除去法 19

第 5 章　変調スペクトル分析（MSA） **20**
5.1　変調スペクトルとは 20

5.2 変調スペクトル分析 20
5.3 様々な信号の変調スペクトル 20

第 6 章 評価シミュレーション 28
6.1 評価尺度 28
6.1.1 実験１：従来法との精度の比較 29
6.1.2 実験 2：背景雑音に対する耐性評価 31
6.1.3 実験 3：非音声に対する耐性評価 33
6.1.4 実験 4：総合評価 35
6.1.5 実験 5：実環境を想定した実験刺激に対する VAD 精度の評価 . . . 37

第 7 章 まとめ 39
7.1 本研究で明らかになったこと 39
7.2 残された課題 39
7.3 今後の展望 40

図 目 次

1.1 論文の構成 3

3.1 提案法のブロックダイアグラム 9
3.2 音声信号 $x(t)$，雑音信号 $n(t)$，観測信号 $y(t)$ 10
3.3 IMF の様態 11
3.4 各信号の変調スペクトル 12

4.1 IMF の制約条件を満たす信号の例 14
4.2 EMD による信号分解プロセス 16
4.3 EMD による信号分解の事例 17
4.4 (a) ミュージカルノイズと (b) 各 IMF におけるミュージカルノイズのエネルギー分布（文献 [16] より引用） 18

5.1 音声信号の変調スペクトル 22
5.2 定常雑音の変調スペクトル 23
5.3 環境雑音の変調スペクトル 24
5.4 音楽の変調スペクトル 25
5.5 鳥の鳴声の変調スペクトル 26
5.6 音声信号の背景雑音環境（SNR=10 dB）での変調スペクトル 27

6.1 実験 1 の音声区間検出の例（クリーンな条件） 29
6.2 実験 1 の結果（クリーンな条件） 30
6.3 実験 2 の音声区間検出の例（雑音条件） 31
6.4 実験 2 の結果（雑音条件） 32
6.5 実験 3 の音声区間検出の例（非音声混入条件） 33
6.6 実験 3 の結果（非音声混入条件） 34
6.7 実験 4 の音声区間検出の例（総合的な条件） 35
6.8 実験 4 の結果（総合的な条件） 36
6.9 実験 5 の音声区間検出の例（実環境条件） 37
6.10 実験 5 の結果（実環境条件） 38

表 目 次

1.1　記号の定義 4

第1章　序論

1.1　はじめに

音声区間検出（Voice Activity Detection, VAD）とは，音声とそれ以外の信号からなる解析信号から，音声の存在する区間（音声区間）と音声の存在しない区間（非音声区間）を自動的に検出する技術である [1].

VAD 自体は単純な技術であるが，様々な音声処理技術において重要な要素技術として用いられている．例えば，音声認識分野では認識対象の自動抽出化や非音声区間での無駄な演算量の削減に，音声通信分野では音声区間に限定した能率的な音声符号化やそれに伴う帯域利用の効率化に利用されている．そのため，高精度 VAD の開発は様々な関連技術の精度向上にかかわるとして，様々な研究が行われてきた.

実際にこれらのアプリケーションが用いられるのは実環境である．すなわち，背景に大小さまざまな雑音が存在したり，音楽や雑踏，生活環境音などの音声以外の信号を含む環境である．しかし，VAD を行う上でこれら雑音の存在は非常に重大な問題となる．対象信号が背景雑音を含む場合，雑音に埋もれた音声信号を検出しそこなうため，誤棄却率（音声の区間を非音声の区間と誤認識してしまう確率）が増大する．また，対象信号が雑音以外にも非目的音を含む場合は，非音声区間を音声区間と誤検出するため，誤受理率（非音声の区間を音声の区間と誤検出してしまう確率）が増大する.

以上の要因によって，VAD の精度は著しく低下する．実際問題として，VAD を活用する様々なアプリケーションは雑音のある環境や対象音声以外の非音声信号が混在する環境で利用されるため，実環境で VAD を活用するためには利用環境や対象信号に依存せずに正確に音声区間を検出できなければならない.

そこで，経験的モード分解（Empirical Mode Decomposition, EMD）[7] を利用して定常雑音の除去を行い，変調スペクトル分析（Modulation Spectrum Analysis, MSA）[8] を用いて得られた信号区間の音声／非音声判別を行う，二段階処理からなるロバスト VAD 法を提案する.

1.2　研究の背景

古典的な VAD 法は，信号のパワーの閾値処理やゼロ交差数を利用して行われる [1]．この手法は，観測信号が対象音声のみをであれば，ほぼ完全に音声区間を検出できるが，対象音声以外の成分（雑音成分）が混入すると検出精度の著しい低下を招くという問題があ

る．特に，対象信号が背景雑音を含む場合，雑音に埋もれた音声信号を検出しそこなうため，誤棄却率（音声の区間を非音声の区間と誤認識してしまう確率）が増大する．また，対象信号が雑音以外にも非目的音を含む場合は，非音声区間を音声区間と誤検出するため，誤受理率（非音声の区間を音声の区間と誤検出してしまう確率）が増大する．

以上の要因によって，VAD の精度は著しく低下する．実際問題として，VAD を活用する様々なアプリケーションは雑音のある環境や対象音声以外の非音声信号が混在する環境で利用されるため，実環境で VAD を活用するためには利用環境や対象信号に依存せずに正確に音声区間を検出できなければならない．

近年，これらの問題を解決するために，閾値の決定に統計的手法を用いる Otsu の閾値法 [3] や複数のパラメータ（スペクトル歪み，全帯域のエネルギー，低帯域のエネルギー，ゼロ交差数の４種類）を音声／非音声の判別に利用する G.729 Annex B の VAD 法 [4]，音声特有の特徴を用いる AMR の VAD 法 [5] や音声の非線形振動特性に着目した Exponentioal 自己回帰モデルを用いた手法 [6] など，実環境を想定した VAD 法が提案されている．これらの方法は，雑音環境でも高精度な音声区間検出が行えるが，音声と非音声を区別して検出する能力は弱い．そのため，非音声信号が存在する環境では，高精度な VAD を行うことはできない．

1.3 本研究で取り組むべき課題

VAD は観測信号の音声区間を検出することで，様々な音処理分野に応用することができる．しかし，音声以外の背景雑音や非音声信号が存在すると，音声区間の検出精度は著しく低下する．

この問題を解決するために，音声の特徴や，統計的手法を用いることで，実環境を想定した VAD 法が提案されている．しかし，これらの方法は，雑音環境でも高精度な音声区間検出が行えるが，音声と非音声を区別して検出する能力は弱い．もしくは，音声／非音声の分別ができても，背景雑音が存在すると分別がうまく行えない．このように，現在の VAD は背景雑音と非音声信号が同時に存在するような環境には弱い．

しかし，実際に VAD ，および VAD を応用したアプリケーションが用いられるのは，背景雑音や音声以外にも様々な信号が混在するような実環境である．本研究ではこの問題を解決するため，実環境でも正確に音声区間の検出が可能な VAD 法の研究を行う．そのために，背景雑音の混入，および，目的音声以外の非音声信号の混入に対して，ロバストな VAD 法の検討を行う．

1.4 本研究の目的

本研究では，背景雑音の混入，および，目的音声以外の非音声信号の混入に対してロバストな VAD 法の検討を行う．そのために，VAD を行う前に対象信号に対して雑音除去

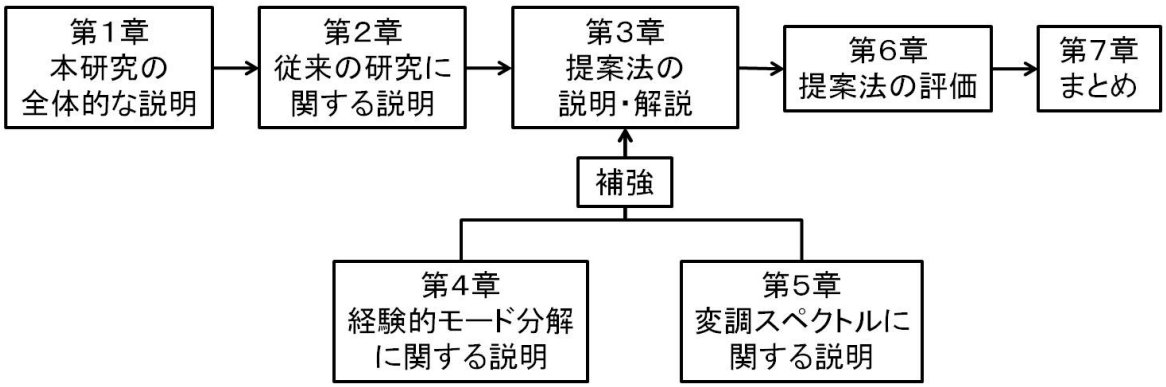

図 1.1: 論文の構成

を行い，雑音除去後の信号から音声区間の候補の検出を行う．その後，候補となる各区間が，音声と非音声のどちらの区間かを MSA によって判別することで，高精度な音声区間検出を行う VAD 法を提案する．本論文では，雑音除去に EMD を，音声／非音声の判別に MSA を用いることで目的であるロバスト VAD の検討を行う．

1.5 本論文の構成

本論文は７章で構成される．全体的な流れを図 1.1 に示す．

第１章では，本研究で対象とする研究分野の背景や問題点を述べる．そして，本研究で取り扱う問題とその目的を述べる．

第２章では，従来の研究に対して説明する．従来の研究の内容と問題点，それを解決するためにどのような VAD を設計するべきかを述べる．

第３章では，２章で説明した従来研究の欠点を克服するための提案法について説明する．目的である背景雑音と非音声信号の混入に対してロバストな VAD を，EMD，MSA のどのような特徴を，どのように組み合わせることで実現するかを述べる．EMD，MSA についての詳細な説明はそれぞれ第４章，第５章で行う．

第４章では，経験的モード分解について説明する．ここでは，信号の分解過程を示し，実際に作成した信号を分解することで，その振る舞いについて紹介する．

第５章では，変調スペクトル分析について説明する．変調スペクトルの導出方法を示し，実際にいくつかの種類の信号で，どのようなスペクトルが現れるかを紹介する．

第６章では，提案法に対して行った評価を紹介する．提案法と従来法の音声区間検出精度を比較する．その後，EMD，MSA がそれぞれ背景雑音除去と音声／非音声判別として，有効に働いているかを評価する．また，背景雑音と非音声信号が同時に存在するとき，さらに実環境において，EMD，MSA が有効に働いているかを評価する．

第７章では，本研究で得られた知見をまとめる．各評価実験の結果とそこから導かれる結果をまとめる．また，今後に残された課題を説明し，本研究の今後の展望についても説

明する.

1.6 記号の定義

ここでは，本論文で利用する記号を定義し，それらを表 1.1 に示す.

表 1.1: 記号の定義

記号	定義
θ	閾値
I_N	推定された非音声区間のデータ数
μ_N	推定された非音声区間の平均値
σ_N^2	推定された非音声区間の分散
I_S	推定された音声区間のデータ数
μ_S	推定された音声区間の平均値
σ_S^2	推定された音声区間の分散
I_θ	全体のデータ数
μ_θ	データ全体の平均値
σ_θ^2	データ全体の分散
σ_w^2	クラス内分散
σ_b^2	クラス間分散
σ_θ^2	全分散
$y(t)$	観測信号
$x(t)$	音声信号
$n(t)$	雑音信号
$\hat{x}(t)$	再合成信号
C_m	固有モード関数
M	IMF の総数
m	チャンネル数
$l(t)$	上側エンベロープ
$u(t)$	下側エンベロープ
$a(t)$	エンベロープの平均
$e_y^2(t)$	パワーエンベロープ

第2章　従来の研究

2.1　古典的手法

もっとも初期の VAD の研究は無雑音環境を想定したものである．この古典的な VAD 法は，信号のパワーの閾値処理とゼロ交差数を利用して行われる [1]．信号のパワーとゼロ交差点数は，それぞれ，有声音と無声子音などを検出するために用いられる．これは，有声音が存在すれば信号のパワーは増大し，無声子音が存在すればゼロ交差点数は多くなる性質を利用し，事前に設定した閾値に基づいいてこれらを処理することによって VAD を行っている．

これらの音響的特徴は少ない計算量で抽出できるため，近年でもよく用いられる．しかしこの手法は，観測信号が対象音声のみをであれば，ほぼ完全に音声区間を検出できるが，対象音声以外の成分（雑音成分）が混入すると検出精度の著しい低下を招くという問題がある．そのため，雑音環境を対象とした VAD の研究が行われるようになった．

2.2　近代的手法

初期の VAD は単純に無雑音環境での音声区間検出を対象にしていたが，その後，対象環境を定常的な雑音環境，非定常な雑音環境と，より実現的な環境を想定するようになった．例えば，音響的特徴に関しては音声とその他の音をよりよく区別できる特徴をとらえる研究が行われた．また，音声／非音声の識別に関しては，閾値を環境に応じて動的に変更したり，統計的な基準を用いて判断を行う手法などの研究が行われている．

音声の性質を利用する方法として，低周波数域のパワーを利用する手法（音声のパワーが 1〜2 Hz 以下の周波数帯域に集中していることから）や周波数スペクトルの概形，また，声帯の振動数に対応した基本周波数（F_0）とその倍音に当たる周波数帯域（調波成分）を利用したものなどがある．他にも，雑音情報を利用して観測信号の SNR を用いる手法や，時間情報から尤度比を抽出する手法，尖度や歪度どなどの高次統計量を用いる手法，また，これまでに紹介したような，音響的特徴を複数組み合わせて用いる手法など非常に多くの手法が研究されている．

以下にいくつかの具体的な手法を紹介する．

2.2.1 Otsu の閾値法

Otsu の閾値法は，解析する信号ごとに最適な閾値を自動的に計算して用いる手法である．閾値は分離度（separationmetrics）という値が最大となるように，クラス間分散（between-class variance）とクラス内分散（within-class variance）との比から求められる．

解析信号のパワースペクトルに対し，ある閾値 θ で音声区間検出を行う際，非音声区間と判断したデータ数を I_N，平均を μ_N，分散を σ_N^2 とする．同様に，音声区間と判断したとき，そのデータ数を I_S，平均を μ_S，分散を σ_S^2 とし，また，全体のデータ数を I_θ，平均を μ_θ，分散を σ_θ^2 とする．このときのクラス内分散 σ_w^2 は,

$$\sigma_w^2 = \frac{I_N \sigma_N^2 + I_S \sigma_S^2}{I_N + I_S} \tag{2.1}$$

クラス間分散 σ_b^2 は,

$$\begin{aligned} \sigma_b^2 &= \frac{I_N(\mu_N - \mu_\theta)^2 + I_S(\mu_S - \mu_\theta)^2}{I_N + I_S} \\ &= \frac{I_N I_S(\mu_N - \mu_S)^2}{(I_N + I_S)^2} \end{aligned} \tag{2.2}$$

として表すことができる．このとき，全分散（total variance）σ_θ^2 は,

$$\sigma_\theta^2 = \sigma_w^2 + \sigma_b^2 \tag{2.3}$$

となり，クラス内分散 σ_w^2 とクラス間分散 σ_b^2 の比である分離度は,

$$\frac{\sigma_b^2}{\sigma_w^2} = \frac{\sigma_b^2}{\sigma_\theta^2 - \sigma_b^2} \tag{2.4}$$

と求めることができる．この分離度が最も高くなるような θ を閾値として使用する．

2.2.2 G.729 Annex B の VAD 法

G.729 とは主に VoIP（Voice over Internet Protocol）で用いられているコーデックである．G.729 には，いくつかの拡張規格が存在し，その中の一つに，無音区間圧縮等によってビットレートを低減するためのアルゴリズムを規定した Aneex B という拡張規格がある．この Annex B の規格では，無音区間圧縮を行うため音声／非音声を判別するための機能が備わっている．

Annex B の VAD では，入力信号から全帯域および低帯域フレームエネルギー，線スペクトル周波数 (LSP) および零交差率の 4 つのパラメータ特性を抽出し音声区間検出に用いる．このように，Annex B の VAD は複数種類のパラメータを用いて音声区間検出を行う VAD である．

2.2.3 Exponentioal 自己回帰モデルを用いた手法

石塚らの提案した Exponentioal 自己回帰（ExpAR）モデルを用いた VAD は，音声信号固有の性質に基づいて音声区間検出を行う手法であり，音声の非線形振動特性に着目した手法である．ExpAR モデルは非線形振動機構を組み込んだ非ガウス的確率過程をモデル化したものであり，振幅や位相か確率的に変動する時系列データに対して有効といわれる手法である．

音声信号は，一定の音程・音高を意図して発話を行ったとしても，その母音波形には，基本周期や各周期ごとの振幅に微細な変動，揺らぎが発生してしまう．この非線形な振動特性を考慮したモデルが，ExpAR モデルであり，石塚らは非線形振動の度合いに対応するパラメータを設定することで，雑音環境中でも高い精度で音声区間検出を行える VAD 法を実現した．

2.3 従来法の問題点

これらの方法は，背景雑音が存在する環境でも高精度な音声区間検出を行えるが，検出された音声区間が音声か非音声を判定する能力は弱い．そのため，非音声信号が存在する環境では，高精度な VAD を行うことはできない．特に，背景雑音と非音声信号が同時に存在するような環境では，音声の区間のみを正確に区別して検出することは非常に困難である．このことは，実環境で音声区間検出を行う上で，非常に重大な問題となる．

2.4 解決すべき課題

実際に VAD が活用されるアプリケーションは，背景雑音や音声以外の信号の存在する実環境で用いられる．しかし，現在提案されている手法では背景雑音に頑健な音声区間検出を行うことはできるが，非音声信号の混入には弱い．実環境での利用を考えるのならば，背景雑音だけでなく，非音声信号に対しても頑健である必要がある．さらに，背景雑音と非音声信号が同時に存在するような環境であっても，頑健に音声区間検出ができる必要がある．

第3章　背景雑音と非音声信号の混入にロバストなVAD

3.1　EMD とMSA を用いたVAD

EMD は，解析信号を定常性の高い成分から順に固有モード関数（Intrinsic Mode Function, IMF）に分解する特徴をもつ．そのため，解析信号の成分を定常雑音の成分と非定常な音声信号の成分に容易に分解できるが，その境界を IMF 上で正確に区別することが難しい．

一方，MSA は解析信号の変調スペクトルを分析するものである．変調スペクトルは音声や環境雑音など，信号の種類によって異なった特徴を示すことが知られており，この特徴を比較することで解析信号が音声か非音声かを区別することができる．しかし，解析信号に背景雑音が含まれる場合，それぞれの変調スペクトルの特徴が混ざってしまい，目的音声の特徴を正確に判別することが難くなってしまう．

しかし，EMD により背景雑音を除去し，MSA により音声特有の特徴を正確に抽出することで，互いの欠点を補いつつ，背景雑音と非音声信号の混入に頑健な VAD を実現できるものと考えられる．

図 3.1 に，提案法のブロックダイアグラムを示す．

提案法は，雑音除去と音声／非音声判別の二段階で行われる．第一段階では，まず，観測信号 $y(t)$ を EMD により IMF に分解する．各 IMF の MSA から定常成分の IMF（$n(t)$ に対応）を判別し，それを除いた IMF ($x(t)$ に対応) を再合成することで，定常雑音除去を行う．第二段階として，得られた再合成信号 $\hat{x}(t)$ のパワーエンベロープを求め，閾値処理を行い信号の存在する区間を求める．そして得られた信号区間の変調スペクトルを MSA により求め，音声特有の特徴の有無を確認することで，その区間が音声の区間か，非音声の区間かを判別する．判別の基準は，事前調査の結果から，変調スペクトルのピークが -15 dB 以上で 2〜3 Hz にあること，そのピークから 3 dB 減少するまでの先鋭度（フィルタの Q 値に相当）が 1.00〜1.44 であること，4 Hz 以降が平坦な形状になることの 4 点を条件とした．これらの設定については，再度，3.3 節で詳細に述べる．

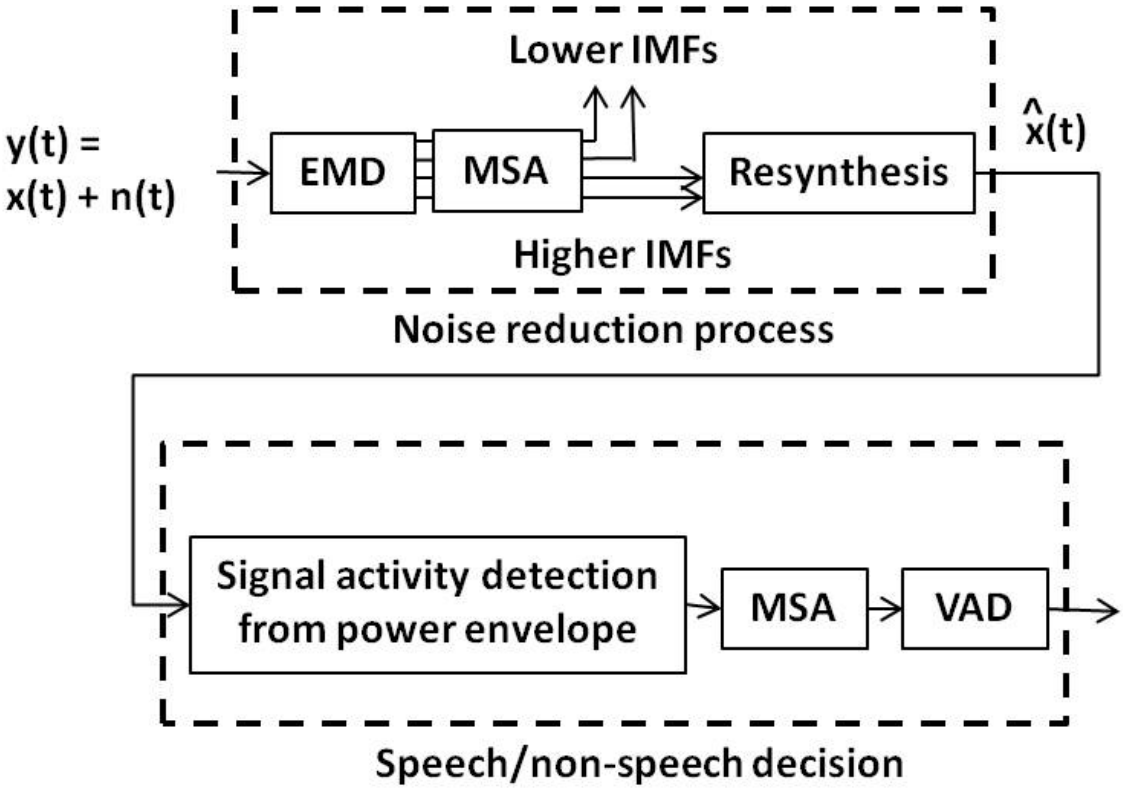

図 3.1: 提案法のブロックダイアグラム

3.2 定常／非定常 IMF の切り分け

EMD は解析信号を定常性の高い成分から順に IMF へと分解していく．その様子を図に示す．音声信号 $x(t)$ と雑音信号 $n(t)$ からなる解析信号 $y(t)$ を図 3.2 に，その信号の EMD から得られる IMF を図 3.3 に示す．全体的に一定な振幅包絡を持つ IMF から，部分的なピークを持つ IMF へと順に分解されていることが分かる．

そのため，IMF を順に調べていくと，定常成分の IMF と非定常成分の IMF が切り替わる境界がある．その境界を特定し，それ以前と以降の IMF をそれぞれ再合成すれば，元信号の定常成分と非定常成分をそれぞれ再合成できる．提案法ではこの性質を利用し，定常雑音成分を除去して，音声の成分を含む非定常成分の IMF のみを合成するが，そのためには IMF の定常成分と非定常成分を切り分ける境界を正しく決定できなければならない．

IMF を切り分ける基準を決定するために，あらかじめ音声区間の分かっている信号に対して，定常な背景雑音を付加した信号を用い，いくつかの方法を基準に IMF 切り分けを行い，その結果を比較した．用いた方法は，正解のデータを用いて切り分けを行う方法（LSD， SNR を基準とした切り分け），あらかじめ定められた IMF で切り分ける方法 (Taufiq の方法，Molla & Hirose の方法)，そして，正解のデータを用いずに切り分けを行う方法（MSA を用いて非定常成分の特徴を持つ IMF を調べる手法，MSA を用いて音

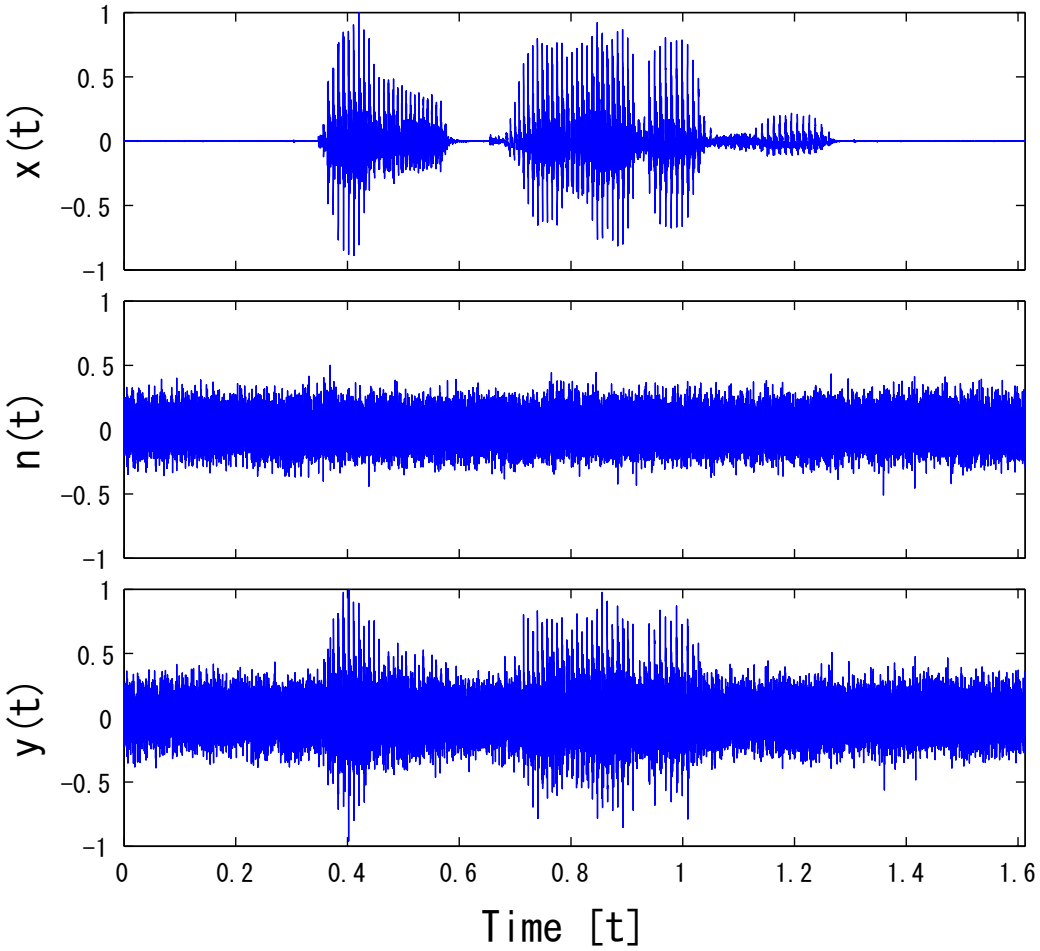

図 3.2: 音声信号 $x(t)$，雑音信号 $n(t)$，観測信号 $y(t)$

声の特徴を持つ IMF を調べる手法，MSA と変調スペクトルの Q 値を用いて音声の特徴を持つ IMF を調べる手法，音声特徴のある IMF の切りだす手法）である．

得られた再合成信号と元のクリーンな信号の SNR， LSD を比較した．

これらの方法の中では，SNR や LSD を用いた手法がもっとも正確に元のクリーンな信号を再生できるはずである．しかし，SNR と LSD の計算には元のクリーンな信号や正解の音声区間の情報が必要である．しかし，実際の環境ではこれらの情報を得ること不可能である．つまり，実環境で用いる VAD に用いることを考えるなら，正解の情報を用いずに正確な切り分けを行える方法を用いる必要がある．そこで，今回 SNR，LSD を用いる方法を目標とし，事前情報を必要とせずに正確な切り分けを行える方法を模索する．

その結果，正解のデータを用いずに切り分けた再合成信号は，正解のデータを用いて切り分けた再合成信号とほとんど同程度の SNR， LSD を示した．また，正解のデータを用いない方法で，尖鋭度を用いるか用いないかで切り分けに対する影響はなかった．以上の

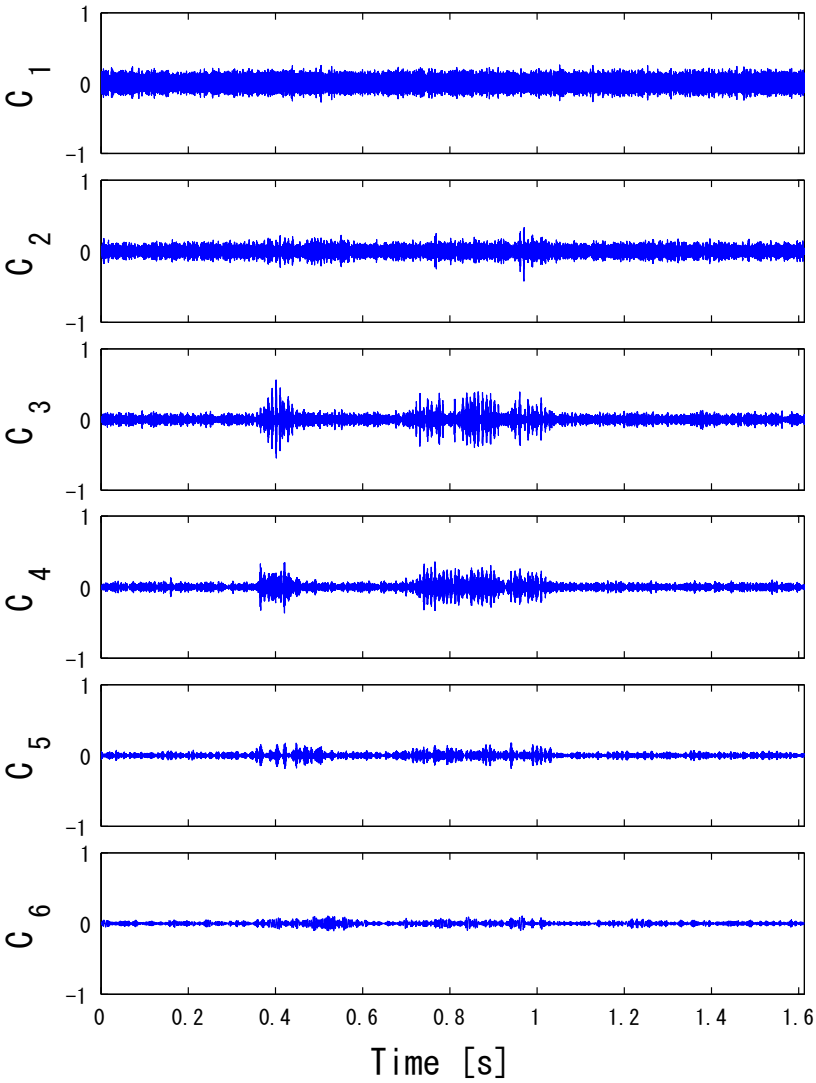

図 3.3: IMF の様態

結果から，変調スペクトルを基準に切り分けを行うことで，正解データを用いなくとも，用いた場合とほとんど変わらない精度で非定常成分を再合成できることが分かった．

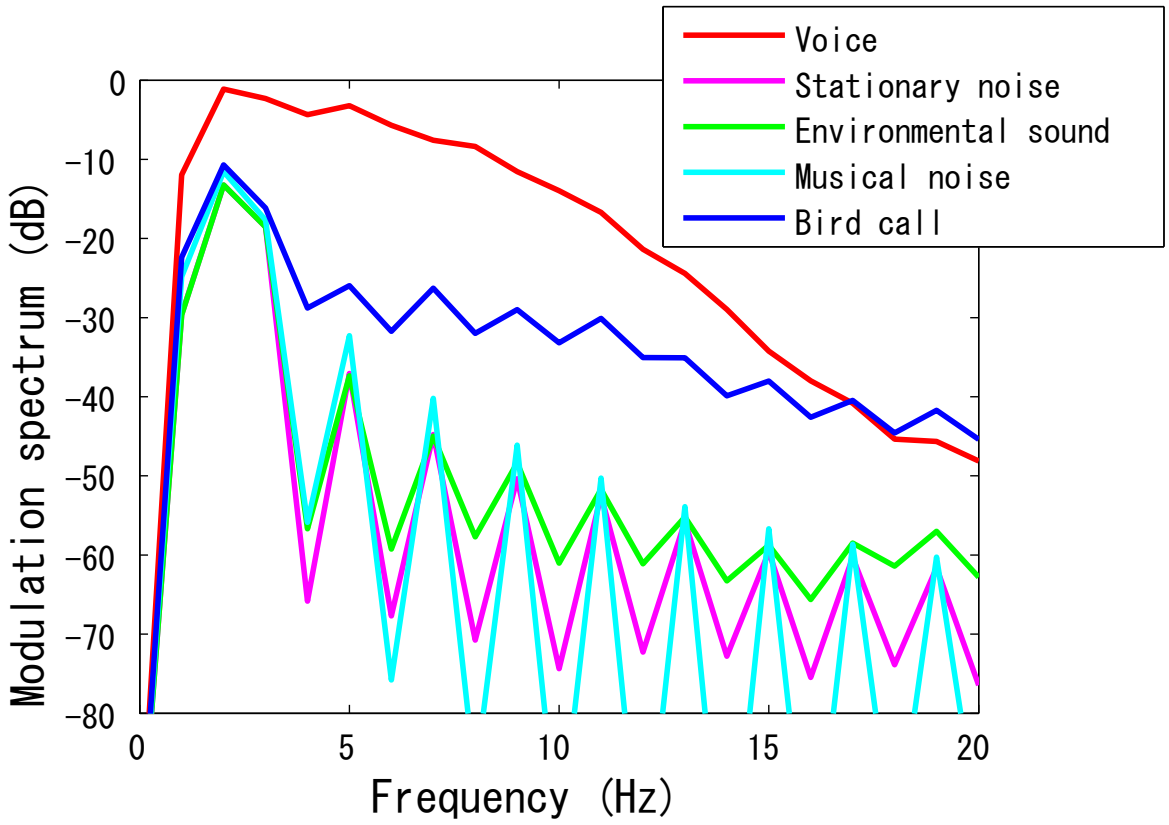

図 3.4: 各信号の変調スペクトル

3.3 音声／非音声の変調スペクトルの判別

第 3.2 節で述べたようにして求めた変調スペクトルから，音声の特徴の有無を調べることで，その変調スペクトルが音声のものか音声以外のものかを判別する．判別の基準を定めるために，音声と定常雑音，実環境での背景雑音として環境音，そして，音声に音響的特徴が近いため従来の VAD では判別が難しいとされている楽器・演奏音と鳥の鳴声の変調スペクトルを調べた．

各種類ごとに 10 個の実験刺激を用い，それぞれの変調スペクトルを求めた．求めた変調スペクトルの信号がある区間での平均を求め，それらを信号の種類ごとに平均をとったものを図 3.4 に示す．得られた結果から，定常雑音の変調スペクトルは，低く，尖鋭度の大きなピークを持ち，高周波数域のピークは平坦なピークを示すことが分かった．また，音声の変調スペクトルは， 他の信号に比べ高く，尖鋭度が小さなピークを持ち，高周波数域ではなだらかなピークを示すことが分かった．

第4章　経験的モード分解（EMD）

4.1　EMD とは

EMD は非定常信号の分析に用いられる手法であり，Huang らによって提案された．次式に示すように，信号 $y(t)$ を IMF の総和 $\sum C_m(t)$ と残差 $r(t)$ に分解する．

$$y(t) = \sum_{m=1}^{M} C_m(t) + r(t) \tag{4.1}$$

ここで，m はチャンネル数，M は IMF の総数である．EMD は時間的な振幅包絡の変動成分に基づき，定常・非定常の度合いを元に，定常性の高いものから順に解析信号を AM-FM 成分に分解する．さらに，可逆的な信号処理であるため，分解された全 IMF を再合成すると，分解前の信号をほとんど損なうことなく得ることができる．音声は非定常信号であることから，非定常成分の IMF の中に分解されると考えられる．提案法ではこの特徴を利用して，非定常成分の IMF のみを再合成して用いることで，元の信号から定常成分である定常雑音を除き，音声の存在する非定常成分のみを用いて VAD を行う．

4.2　固有モード関数

EMD によって分解された IMF は，次に示す二つの制約条件を満たす必要がある．

1. 信号の極地の数と信号のゼロ交差点数は同じか一つ違いであること．
2. 信号の任意の点において，極大点と極小点から作られるエンベロープ（上側エンベロープ $l(t)$ と下側エンベロープ $u(t)$）の平均値 $a(t)$ がゼロであること．

この二つの条件を満たす信号の例を図 4.1 に示す．
ここで，IMF の総数 M は，解析信号 $y(t)$ の性質によって決まる．

4.3　EMD の信号分解プロセス

EMD の信号分解プロセスについて説明する．

まず，入力信号 $y(t)$ に極値が存在するかどうかを確認する．極値が存在する場合，元信号である $y(t)$ を $h(t)$ と置き換え，$h(t)$ が IMF の二つの制約条件を満たしているか確認す

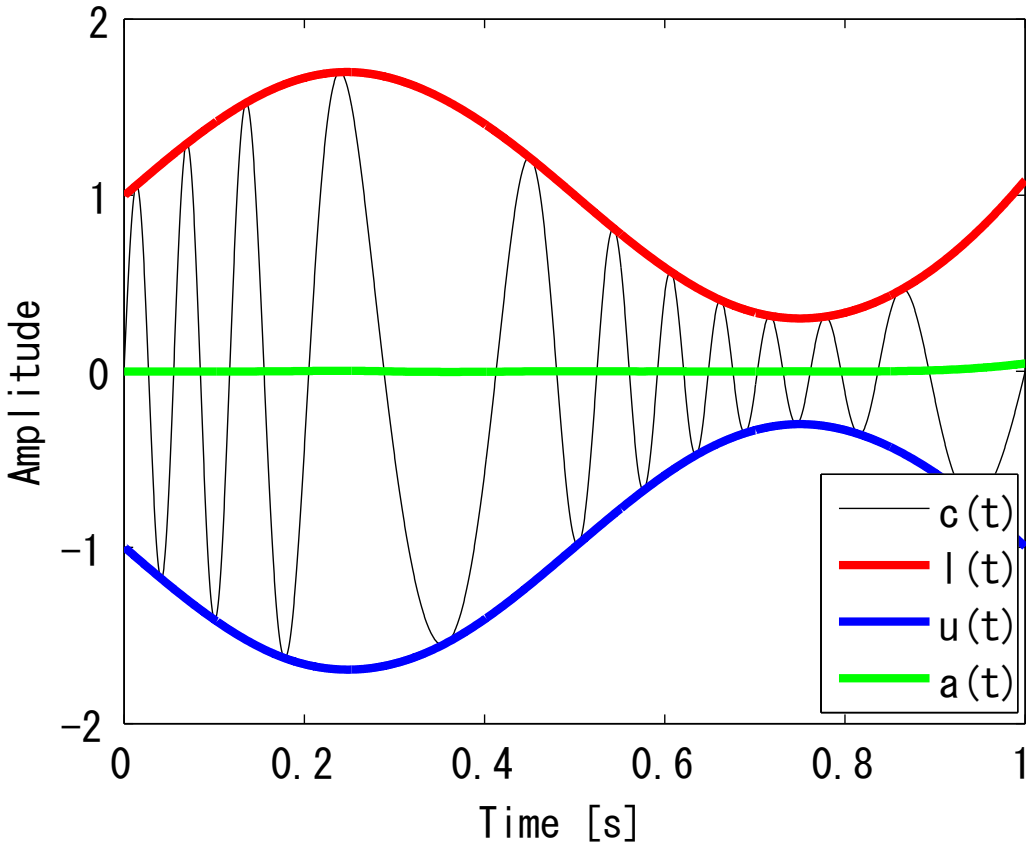

図 4.1: IMF の制約条件を満たす信号の例

る．満たしていない場合，$h(t)$ の極大点・極小点を４次スプライン曲線で補完を行い，上側のエンベロープ $u(t)$ と下側のエンベロープ $l(t)$ を求める．上側エンベロープと下側エンベロープから平均のエンベロープ $a(t)$ を求める（図 4.2 上図 ）．その式を以下に示す．

$$a(t) = \frac{1}{2}(u(t) + l(t)) \tag{4.2}$$

$h(t)$ からエンベロープの平均値を減算したものを新しい $h(t)$ とし（図 4.2 下図 ），この新しい $h(t)$ が IMF の制約条件を満たしているかを確認する．条件を満たしている場合，$h(t)$ を C_{M+1} として，新しい IMF として IMF の集合に加える（IMF の総数 M は $M+1$ となる）．満たしていない場合，新しい $h(t)$ に対して，上側・下側エンベロープの平均値を計算して減算する操作を，$h(t)$ が IMF の制約条件を満たすようになるまで繰り返す．

IMF を得たら，元信号 $y(t)$ から，IMF の集合を減算し，これを次の $h(t)$ とする．その式を下に示す．

$$h(t) = y(t) - \sum_{m=1}^{M} C_m(t) + r(t) \tag{4.3}$$

このとき，$h(t)$ が極値を持つかを調べる．極致を持つ場合，再び $h(t)$ から新しい IMF

を計算する．この操作を，$h(t)$ が極値を持たなくなるまで繰り返す．また，極致を持たない場合，この元信号から全ての IMF を減算した $h(t)$ を残差 $r(t)$ とする．

最後に EMD によって得られた IMF を解析信号をあわせて 図 4.3 に示す解析信号は，$y(t) = \sin(2\pi t) + \sin(200\pi t)$ である．周波数性の高い成分から先に分解され，最後に IMF の制約条件を満たさない残差が生成されていることが分かる．

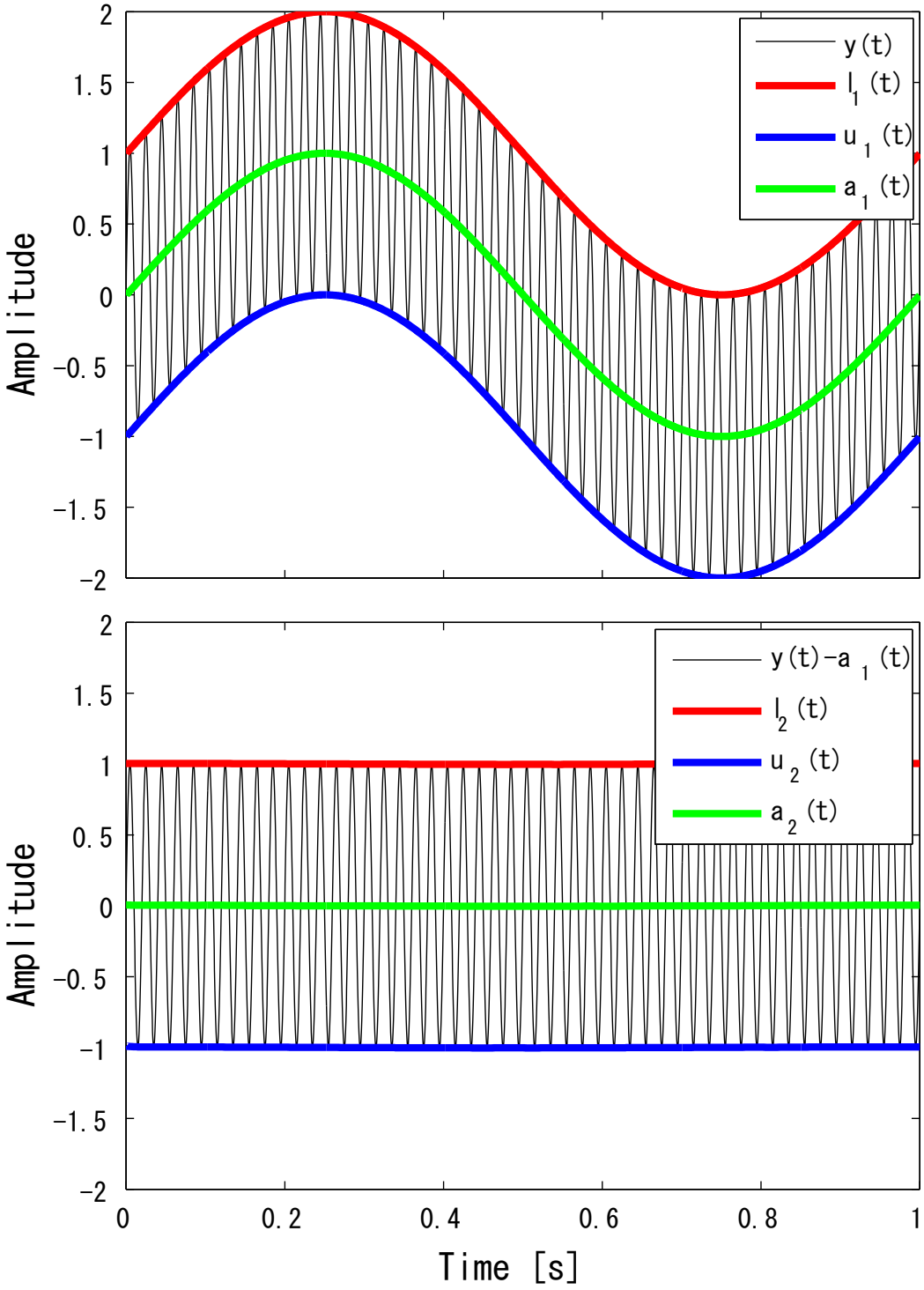

図 4.2: EMD による信号分解プロセス

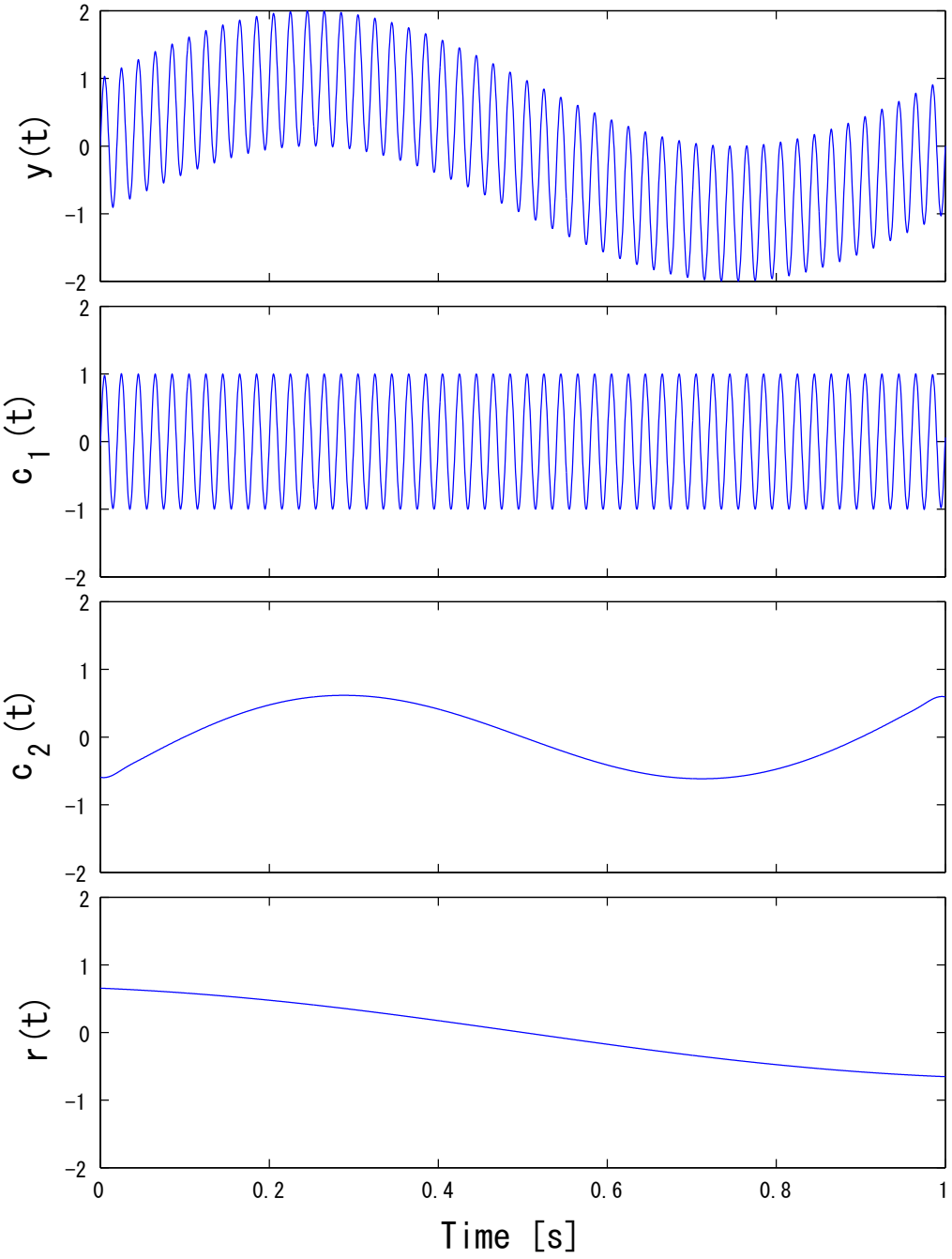

図 4.3: EMD による信号分解の事例

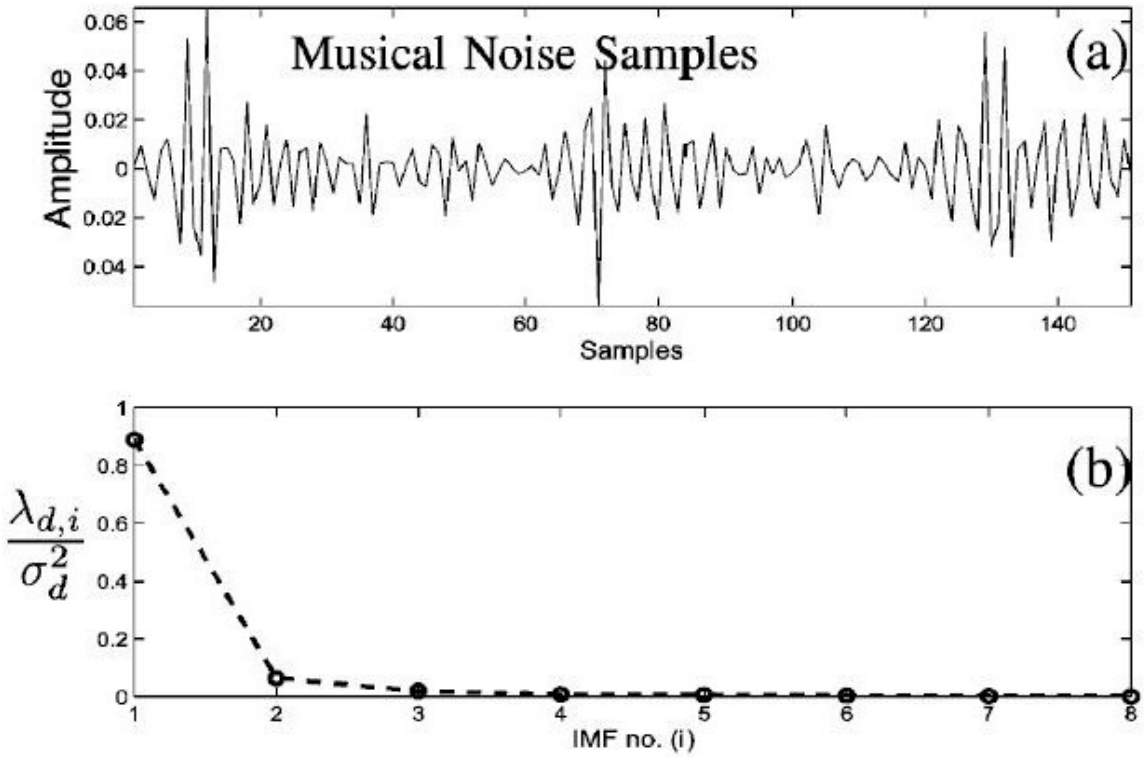

図 4.4: (a) ミュージカルノイズと (b) 各 IMF におけるミュージカルノイズのエネルギー分布（文献 [16] より引用）

4.4 EMD を用いた音信号処理

EMD は非定常信号を分析する手法であり，主に脳波解析や心電図（ECG）波形解析，地震の反射波解析，天文学，画像工学，コンクリート工学，信号処理，金融時系列解析，情報ハイディング，手ぶれ検出などの研究分野で利用されてきた．これらの分析は，どれも非定常な変化を扱わねばならない研究分野である．

近年，非定常信号である音声の解析にも EMD が利用されはじめた．例えば，音源分離や，基本周波数推定そして雑音除去などにも用いられている．今回 EMD を雑音除去に用いるにあたり，過去に提案された雑音除去法について紹介する．

4.4.1 Taufiq らの雑音除去法

Taufiq らは，EMD を用いたミュージカルノイズの抑圧法を提案している．スペクトルサブトラクション法などで雑音除去を行った場合，ミュージカルノイズと呼ばれるトーン性の雑音が発生する問題がある．Taufiq らは，ミュージカルノイズが付加された音声信号を EMD により IMF に分解すると，ミュージカルノイズのエネルギーの大半ははじめ

のいくつかの IMF に集中すると述べている．図 4.4 にミュージカルノイズのサンプル (a) と，各 IMF におけるミュージカルノイズのエネルギー分布を示す（文献 [16] より引用．$\lambda_{d,i}$ は i 個の IMF の分散で，σ_d^2 はミュージカルノイズが付加された音声信号の分散である．図 4.4 より，最初の 1 個の IMF にのミュージカルノイズのエネルギーの 88.94[%] が集中していることが分かる．これにより Taufiq らは，ミュージカルノイズが付加された信号に対して EMD を行い，得られた最初の IMF を取り除くことによって，ミュージカルノイズの抑制を実現した．

4.4.2 Molla & Hirose の雑音除去法

Molla & Hirose は，有声・無声判別の前処理として，EMD を用いた雑音除去法を提案している [15]．この手法は，音声と雑音が持つスペクトルの違いに着目し，解析信号を EMD で IMF に分解した後，高周波成分を含む IMF を取り除くことで雑音除去を行うものである．Molla & Hirose は背景雑音が白色雑音の場合，最初の 2 つの IMF を取り除くことで，雑音除去を行っている．

$$y(t) = \sum_{m=3}^{M} C_m(t) + r(t) \tag{4.4}$$

信号の周波数は，音声の場合低周波数域に分布していることが知られている．それに対して，雑音音声のスペクトルは，広い帯域にわたって存在することが知られている．この性質を利用し，Molla & Hirose は高周波数成分の IMF を除去することで，雑音除去を行う方法を提案した．前述の通り，EMD は振動の速い高周波成分から遅い低周波成分へ順番に IMF 分解していく．Molla & Hirose は各 IMF の瞬時周波数成分を調べるた．その結果，二番目までの IMF 高周波数成分の IMF を取り除くことで，高周波成分を除去し，雑音除去を行う手法を実現した．

第5章　変調スペクトル分析（MSA）

5.1　変調スペクトルとは

変調スペクトルは，振幅の時間的変動を表すものである．音声の変調スペクトルの場合，変調周波数 2〜5 Hz に特有のピークが現れることが知られている．そのため，対象信号に目的音声以外が混入したとしても，この特徴を活用することで，得られた音声区間が本当に音声のものか，それとも音声以外の信号のものかを容易に判別できる可能性がある．

5.2　変調スペクトル分析

観測信号 $y(t)$ の変調スペクトルは，$y(t)$ のパワーエンベロープ $e_y^2(t)$ の Fourier 変換から求めることができる．また，パワーエンベロープ $e_y^2(t)$ は，次式のように Hilbert 変換を用いて求めることができる．

$$e_y^2(t) = \mathrm{LPF}\left[\left|y(t) + j\mathrm{Hilbert}(y(t))\right|^2\right] \tag{5.1}$$

ただし，Hilbert$(\cdot)$ は Hilbert 変換，LPF$[\cdot]$ は低域通過フィルタ（カットオフ周波数 20 Hz ）である．こうして得られたパワーエンベロープに対してフレーム処理（フレーム長 1000 ms, フレームシフト 10 ms）を行い，Hanning 窓をかけて Fourier 変換を行うことで，その振幅スペクトル，つまり変調スペクトルを得る．

5.3　様々な信号の変調スペクトル

MSA の例として，本研究で扱ういくつかの種類の信号に対して MSA を行い，得られた変調スペクトルを示す．扱う信号は，音声，定常雑音，環境音，楽器・演奏音，鳥の鳴き声であり，解析に用いた信号と，それぞれの変調スペクトルを，図 5.1 ，図 5.2，図 5.3 ，図 5.4 ，図 5.5 に示す．

音声信号の変調スペクトルは，音声信号がある部分で，2, 3 Hz にピークを持ち，以降なだらかに減少していく．それ以外の，定常雑音，環境音，楽器・演奏音は，2 Hz あたりにピークを示した後，急激に減少し，以降の中・高周波数域では平坦なプロットになっていることが見て取れる．鳥の鳴声は，音声に近い緩やかなピークを示すことがあるが，

ピークの高さが音声に比べて低いため，この特徴から音声信号の識別が行えると期待できる．

しかし，MSA は複数の信号成分が混在する信号を解析する際，得られる変調スペクトルの特徴が混じってしまい，目的成分の特徴の区別が難しくなる．また，音声信号に白色雑音を SNR = 10 dB で付加した信号と，その変調スペクトルを図 5.6 に示す．ピークの後の変調スペクトルが，白色雑音や環境雑音，楽器・演奏音の変調スペクトルのように，急激に減少し，以降平坦なプロットになっていることが分かる．このため，実環境で変調スペクトルを用いる際，目的以外の信号成分の除去・抑制が重要となる．

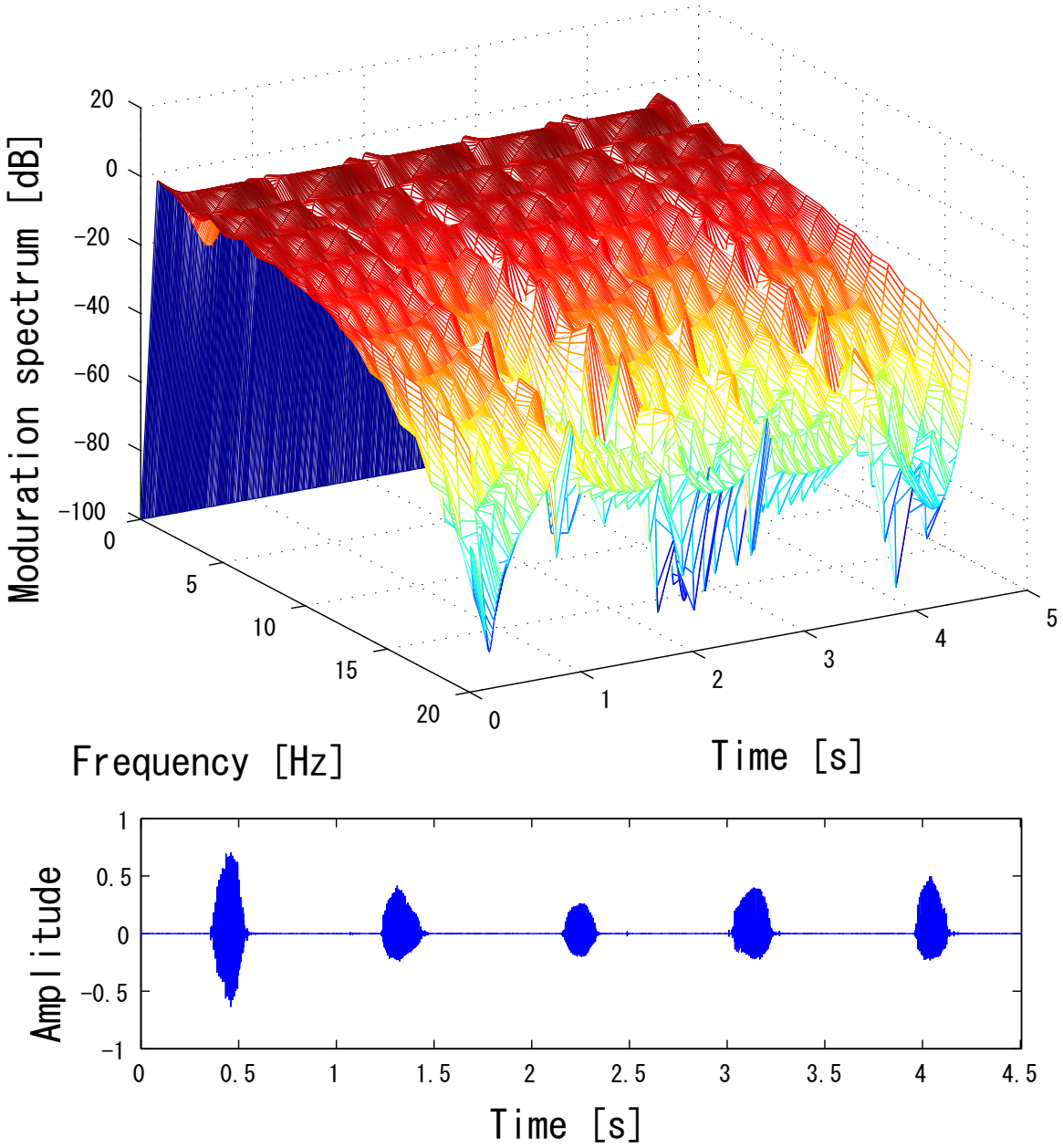

図 5.1: 音声信号の変調スペクトル

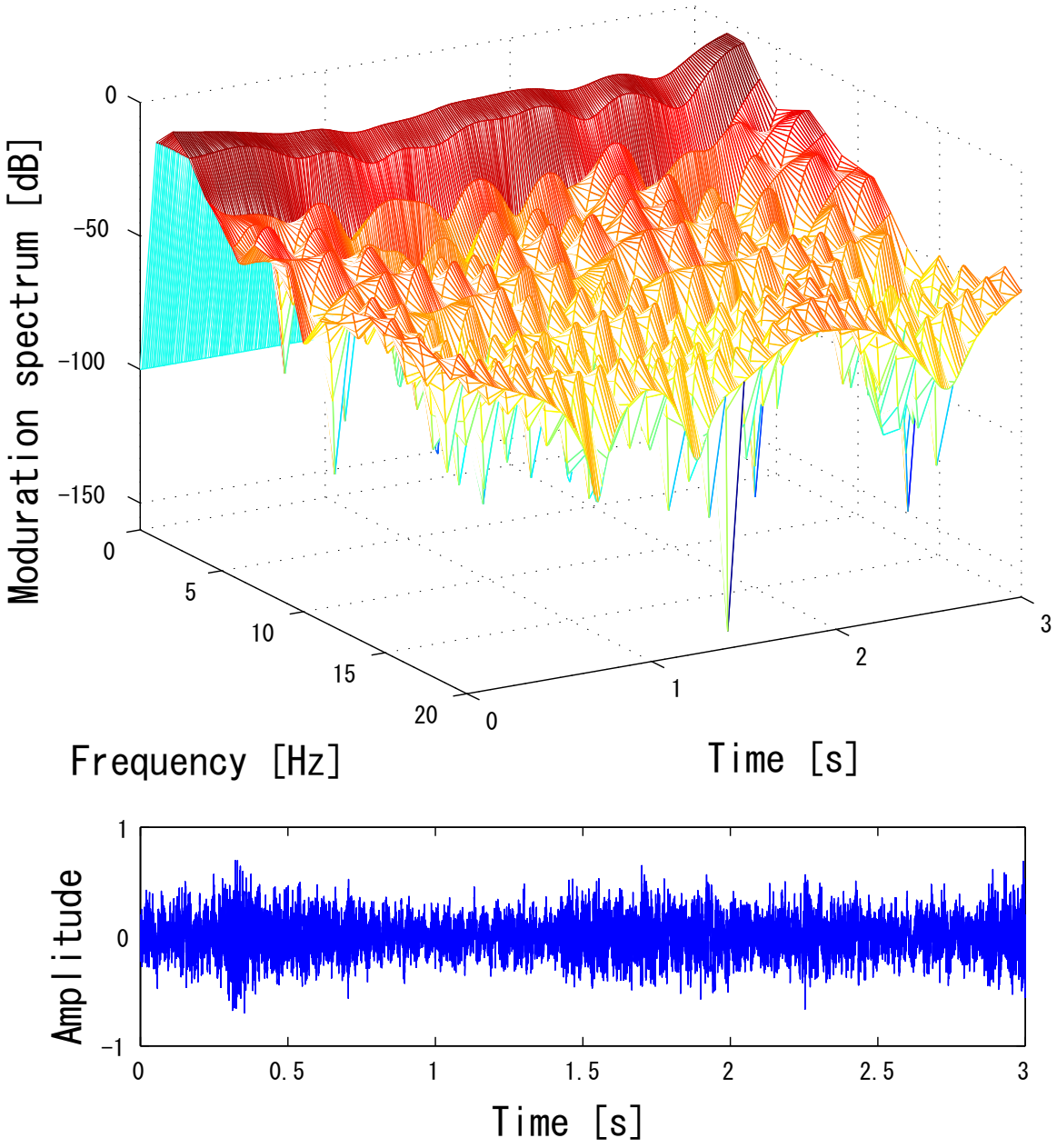

図 5.2: 定常雑音の変調スペクトル

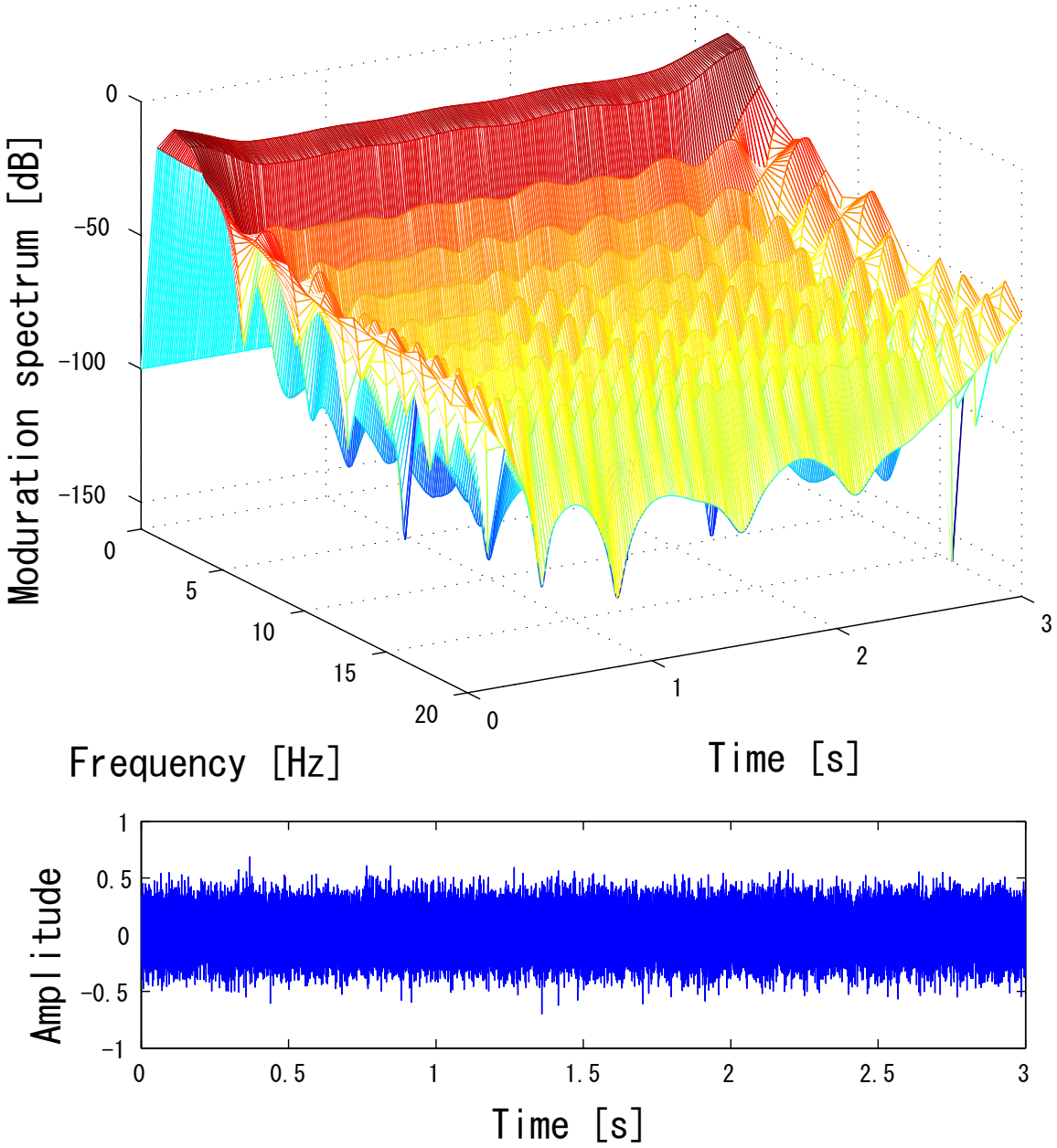

図 5.3: 環境雑音の変調スペクトル

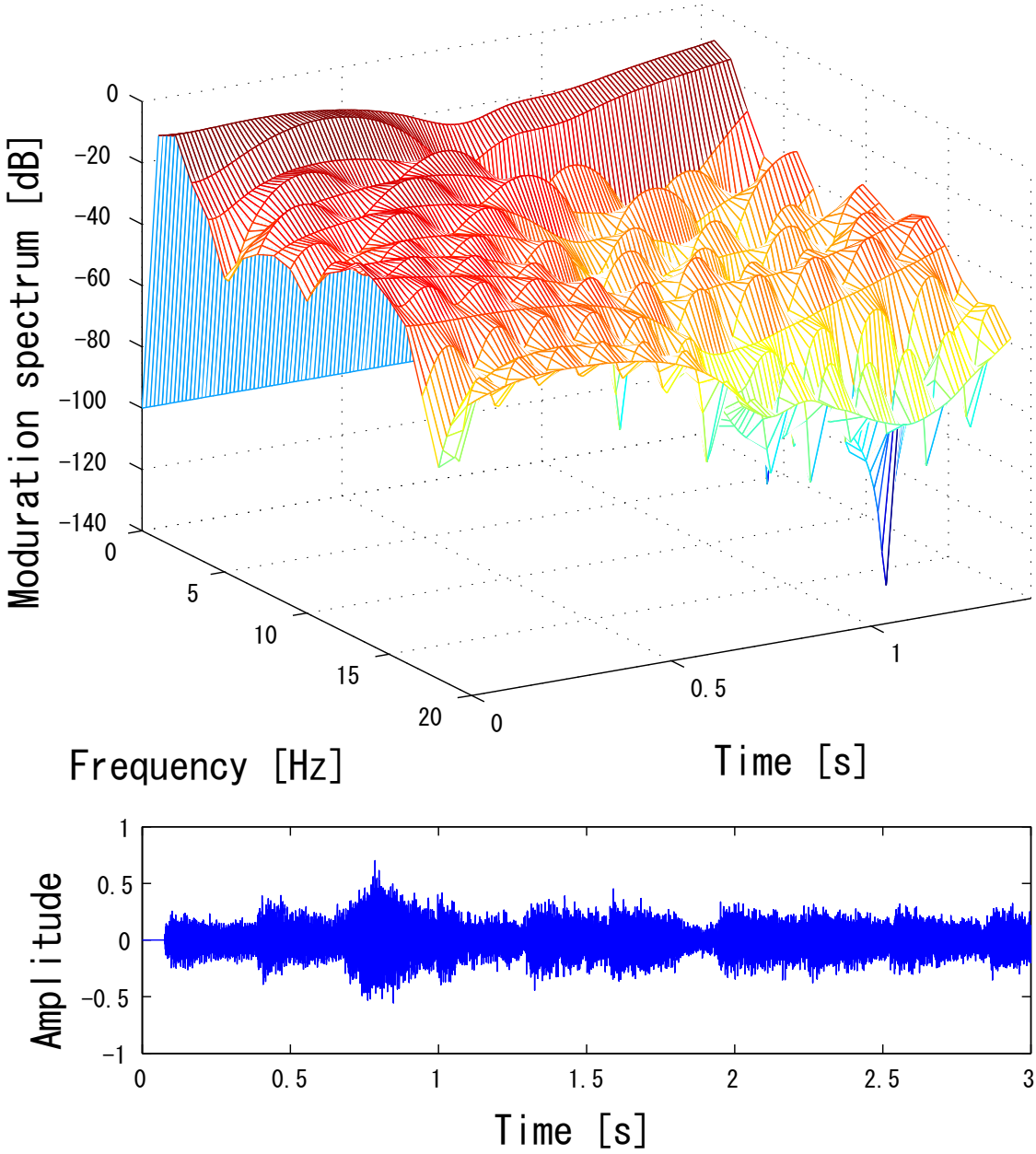

図 5.4: 音楽の変調スペクトル

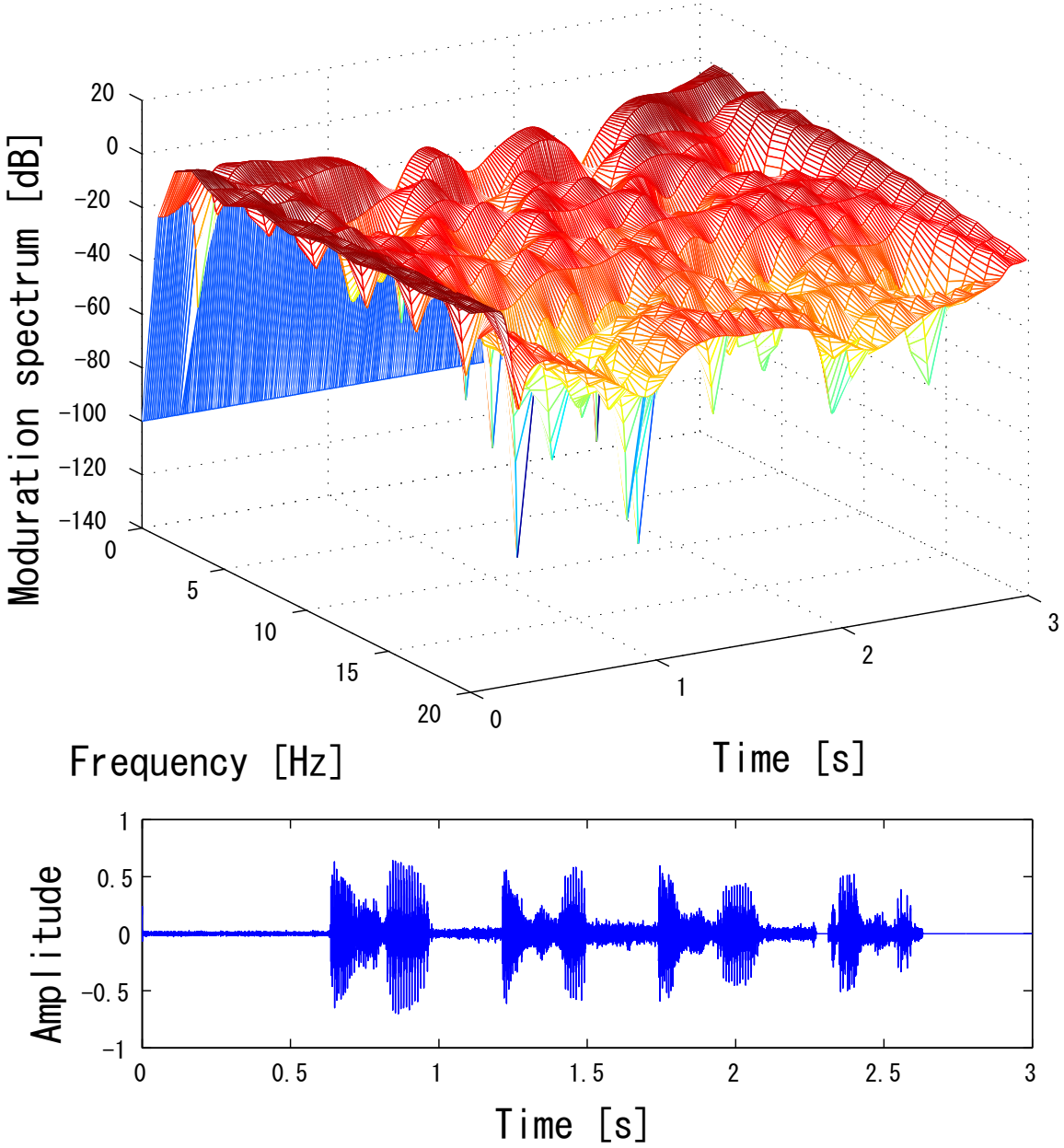

図 5.5: 鳥の鳴声の変調スペクトル

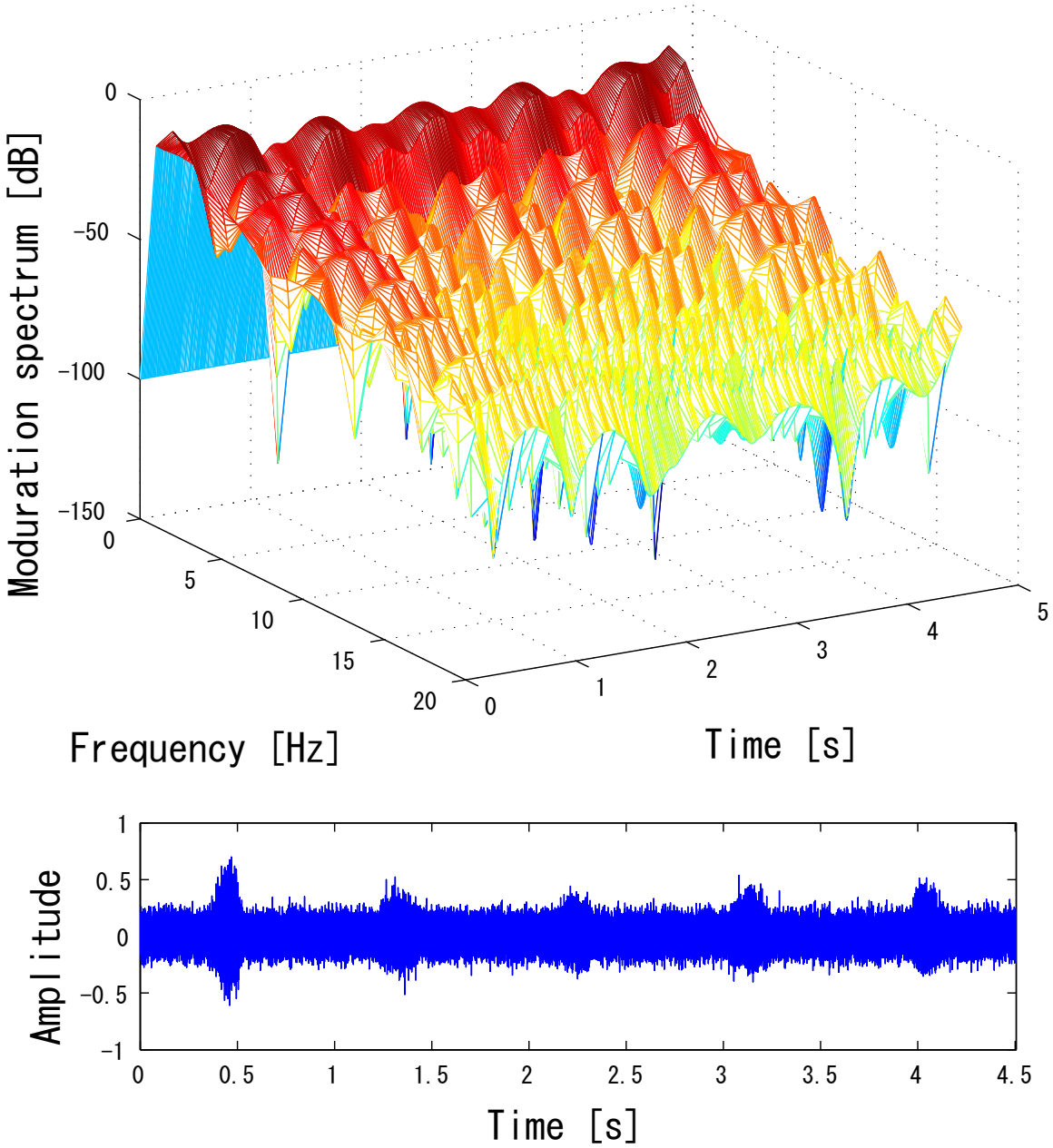

図 5.6: 音声信号の背景雑音環境（SNR=10 dB）での変調スペクトル

第6章　評価シミュレーション

5 種類の環境を想定した実験刺激を作成し，その実験刺激に対して音声区間検出を行うことによって，各環境における提案法の性能評価を行った．また，従来法として Otsu の閾値法 [3] や G.729 Annex B の方法 [4]，パワーエンベロープの閾値処理による VAD の正答率も調べ，比較を行った． 想定した環境は，音声信号のみのクリーンな環境，背景に定常雑音が存在する環境，音声以外に非音声信号も存在する環境，非音声信号と背景雑音（定常）が同時に存在する環境，そして，現実的な環境（音声以外にも他の非音声信号を含み，背景に非定常雑音が存在する環境）である．これらの環境に対する評価を，それぞれ実験 1 から実験 5 として行った．

6.1　評価尺度

評価には次式で示される正答率（%）と誤受理率（False acceptance rate, FAR），誤棄却率（False rejection rate, FRR）を用いた．

正答率は，全データ数のうち，音声／非音声の判別が正しく行われたデータ数の割合から求めた．

$$\text{正答率} = \frac{\text{音声・非音声区間検出正解数}}{\text{フレームの総数}} \times 100 \tag{6.1}$$

FAR は非音声区間を音声区間として誤検出した割合であり，FRR は音声区間を非音声区間として誤棄却した割合である．この二つの値はトレードオフの関係にある．

$$\text{FAR} = \frac{\text{音声区間と検出した中で非音声だったフレーム数}}{\text{音声区間と検出したフレーム数}} \times 100 \tag{6.2}$$

$$\text{FRR} = \frac{\text{非音声区間と検出した中で音声だったフレーム数}}{\text{非音声区間と検出したフレーム数}} \times 100 \tag{6.3}$$

FAR，FRR の値は低い方が正確に音声区間の検出が行われていることを意味する．しかし，今回は視覚的な分かりやすさのため，FAR，FRR を $100 - \text{FAR}[\%]$，$100 - \text{FRR}[\%]$ の形で表記した．そのため，結果を集計したグラフを見る際は，グラフが高い値を示している方が良い結果と判断できる．

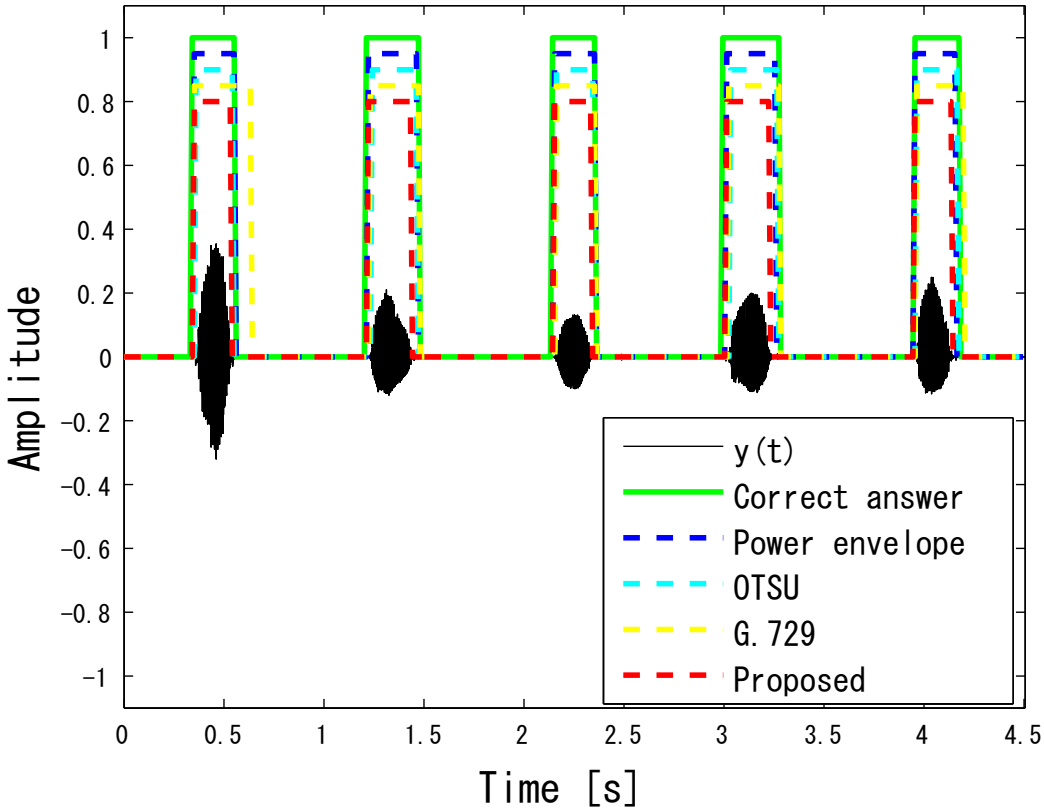

図 6.1: 実験 1 の音声区間検出の例（クリーンな条件）

6.1.1 実験１：従来法との精度の比較

クリーンな環境での音声区間検出の評価を行った．提案法が従来法と比較してどの程度の音声区間検出精度を持つか，また，クリーンな環境でそれぞれの VAD 法がどの程度の検出精度を持つかを示し，以降の実験で雑音環境での各 VAD の頑健性を測る基準とした．

解析信号は，男女五名ずつの発話者から，それぞれ単母音（/a/，/i/，/u/，/e/，/o/）と子音 (/sa/，/ga/，/pe/，/myu/) からなる二種類の実験刺激を作成して利用した．各手法ごとに得られた正答率の，手法ごとの平均値と標準偏差を求めた．検出された音声区間の例を図 6.1 に，得られた正答率，FAR ，FRR をそれぞれまとめたものを図 6.2 に示す．

図 6.1 から，クリーンな環境では，どの手法もかなり正確に音声区間の検出ができていることが分かった．また，図 6.2 でも，正答率，FAR，FRR ともに高い数値を示していた．このことから，提案法，従来法ともに，クリーンな環境では高い精度で音声区間の検出を行えることが確認できた．

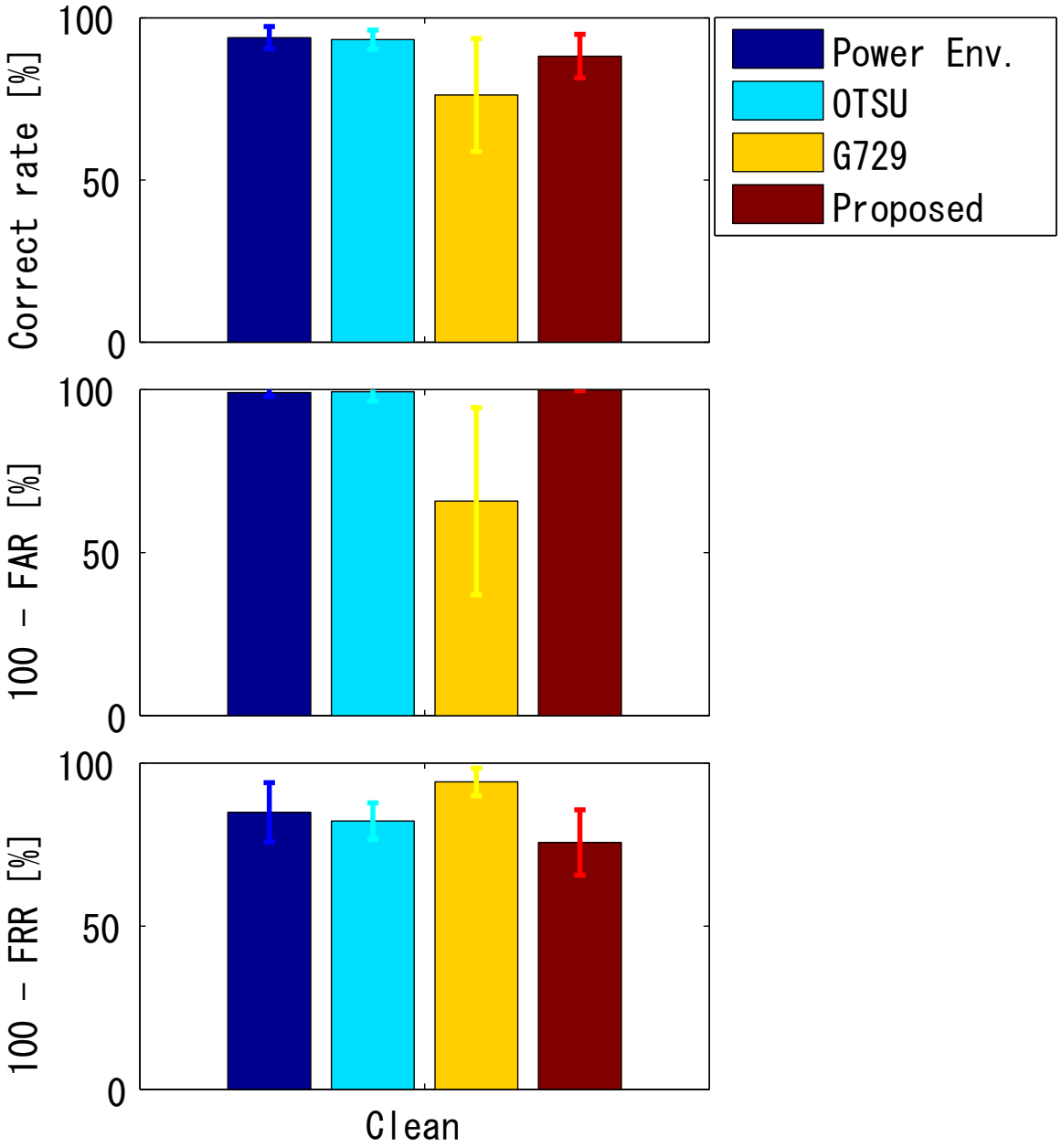

図 6.2: 実験 1 の結果（クリーンな条件）

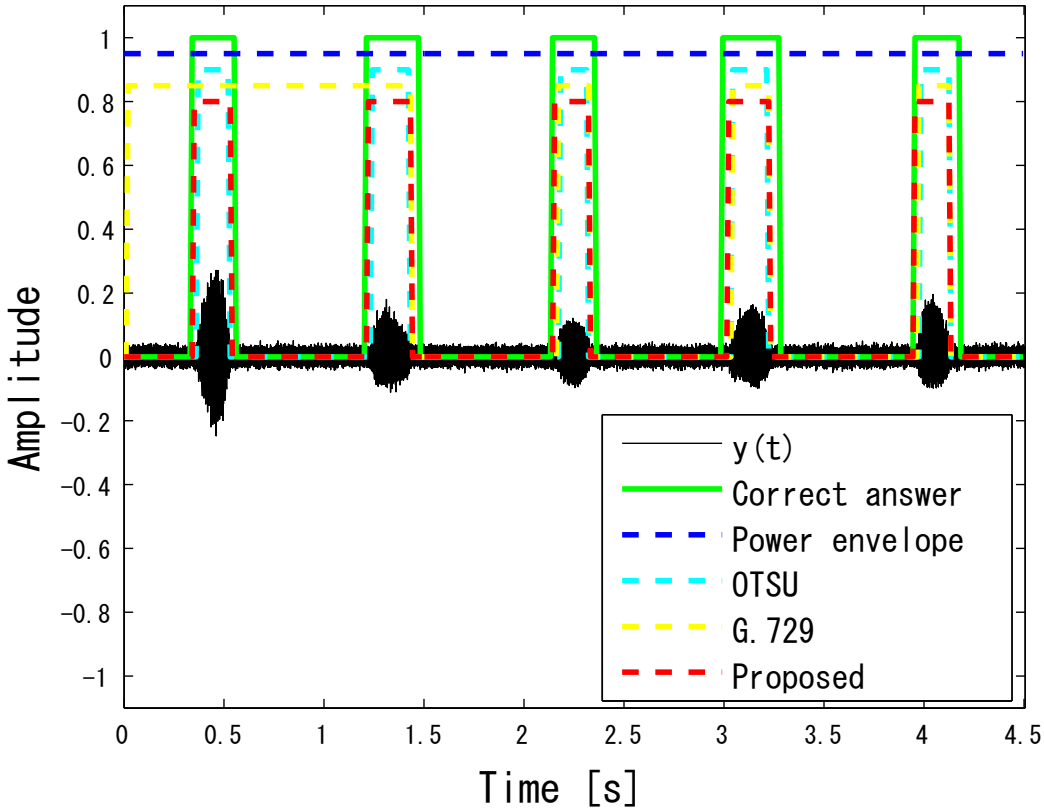

図 6.3: 実験 2 の音声区間検出の例（雑音条件）

6.1.2　実験 2：背景雑音に対する耐性評価

背景雑音として定常雑音が存在する環境での音声区間検出の正答率の評価を行った．この実験により，各 VAD の背景雑音に関する頑健性に対する比較・検討を行った．実験 1 で用いたクリーンな音声信号に定常雑音（ホワイト，ピンク，バブルノイズ）を SNR =20, 10, 0 dB となるように合成した実験刺激を作成し，評価を行った．検出された音声区間の例（白色雑音を SNR = 10 dB で付加した実験刺激）を図 6.3 に，得られた正答率，FAR，FRR をそれぞれまとめたものを図 6.4 に示す．

図 6.3 から，提案法と OTSU 法は，正確に音声区間のみを検出しているが，G.729 とパワーエンベロープの閾値処理では雑音に紛れた音声区間を正確に検出できず，非音声区間を含む広い範囲を音声区間と検出してしまっていることが確認できた．また，図 6.3 から，OTSU 法と G.729 法では SNR の低下に伴い，雑音に紛れた音声区間を検出しそこなうことで，FRR が増加してしまっていることが確認できた．また，パワーエンベロープの閾値処理では，低 SNR 環境下では，雑音の区間も音声の区間として検出してしまうことで，非音声区間が検出できていないことが分かった．しかし，提案法は SNR が低くなっても正答率の減少や FAR，FRR の増加はほとんどみられなかった．この結果から，EMD による定常雑音除去の効果を確認できた．

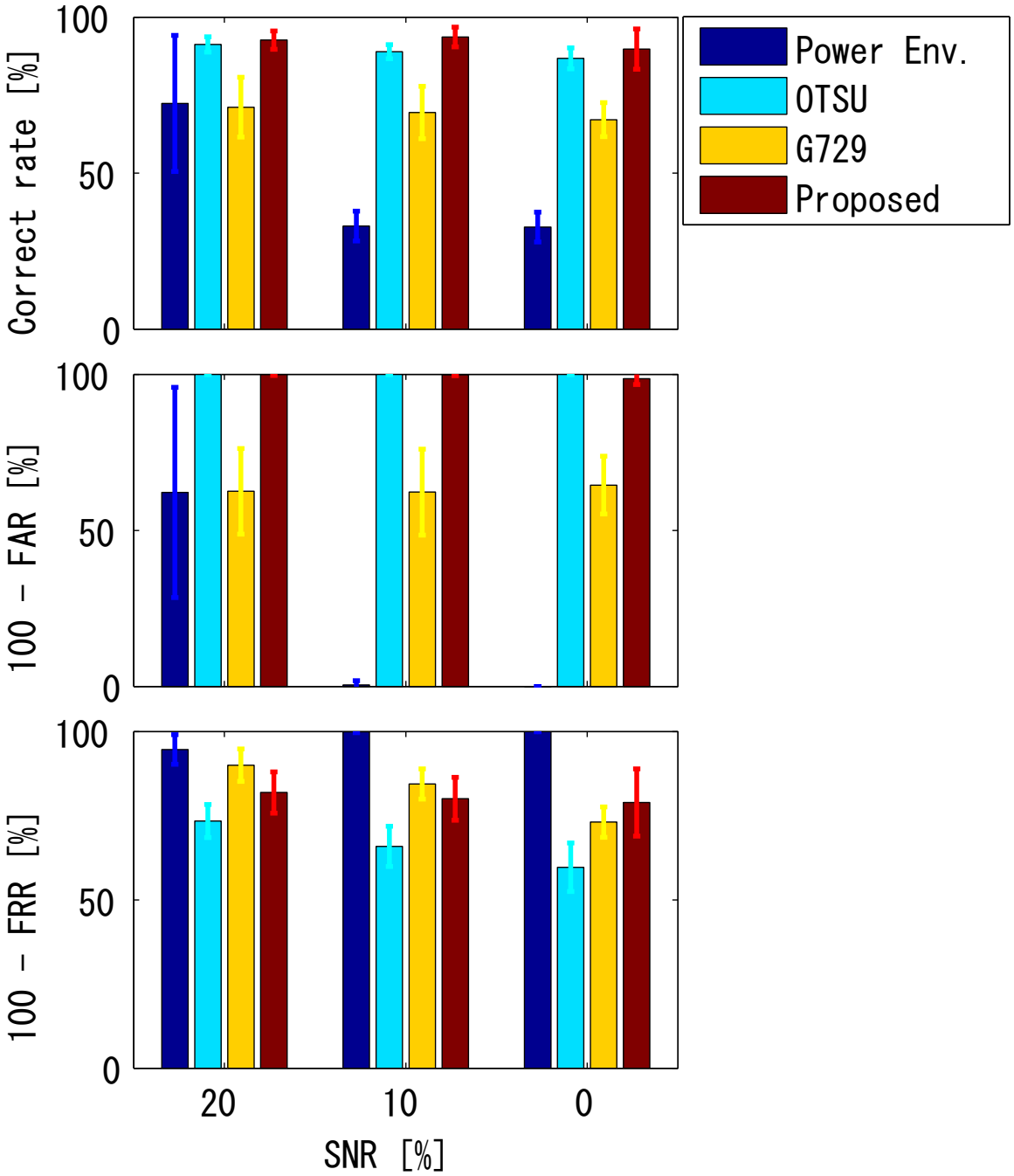

図 6.4: 実験 2 の結果（雑音条件）

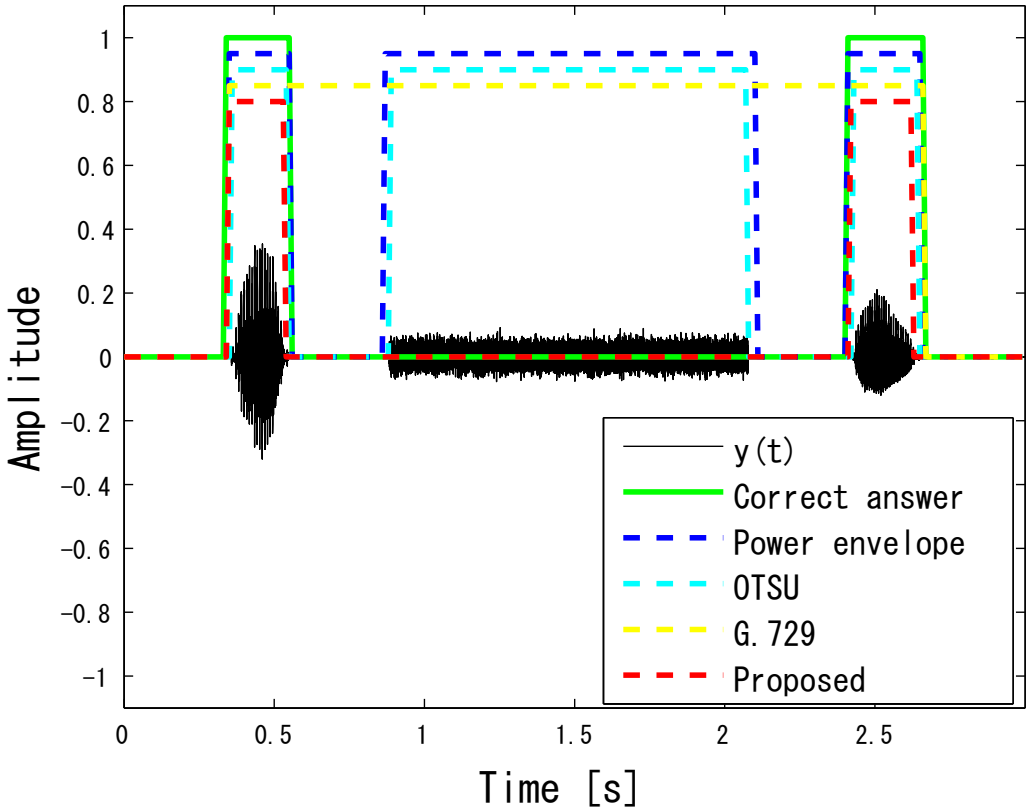

図 6.5: 実験 3 の音声区間検出の例（非音声混入条件）

6.1.3　実験 3：非音声に対する耐性評価

音声と非音声の信号が混在した環境における音声区間検出の正答率を調べた．この実験により，各 VAD の音声以外の信号の混入に対する頑健性を比較・検討する．非音声信号の混入を想定して，/a/+非音声信号+/i/ となるような実験刺激を作成した．ここで，非音声信号には，白色雑音，および，環境音，楽器・演奏音，鳥の鳴声から各 5 種類ずつを用いた．検出された音声区間の例（非音声信号として鳥の鳴声を付加した実験刺激）を図 6.5 に，得られた正答率，FAR ，FRR をそれぞれまとめたものを図 6.6 に示す．

図 6.5 から，提案法が音声区間のみを検出できていることに対して，OTSU 法とパワーエンベロープの閾値処理では，非音声の区間も音声区間として検出して今っていることが分かった．また，G.729 は音声と非音声の区間をまとめて，音声区間と検出していた．図 6.5 から，従来法は，非音声区間を音声区間と誤検出してしまうことにより，クリーンな環境での結果に比べて，FAR の値に増加がみられた．また，正答率にも同様に減少が見られた．これに対し提案法では，検出された信号が存在する区間に対し，それぞれの区間の変調スペクトルから，その区間が音声の区間かどうかの判別を行い，音声区間のみを検出することで平均的に高い正答率と FAR を維持できていた．この結果から，MSA による音声／非音声区間判別の効果を確認できた．

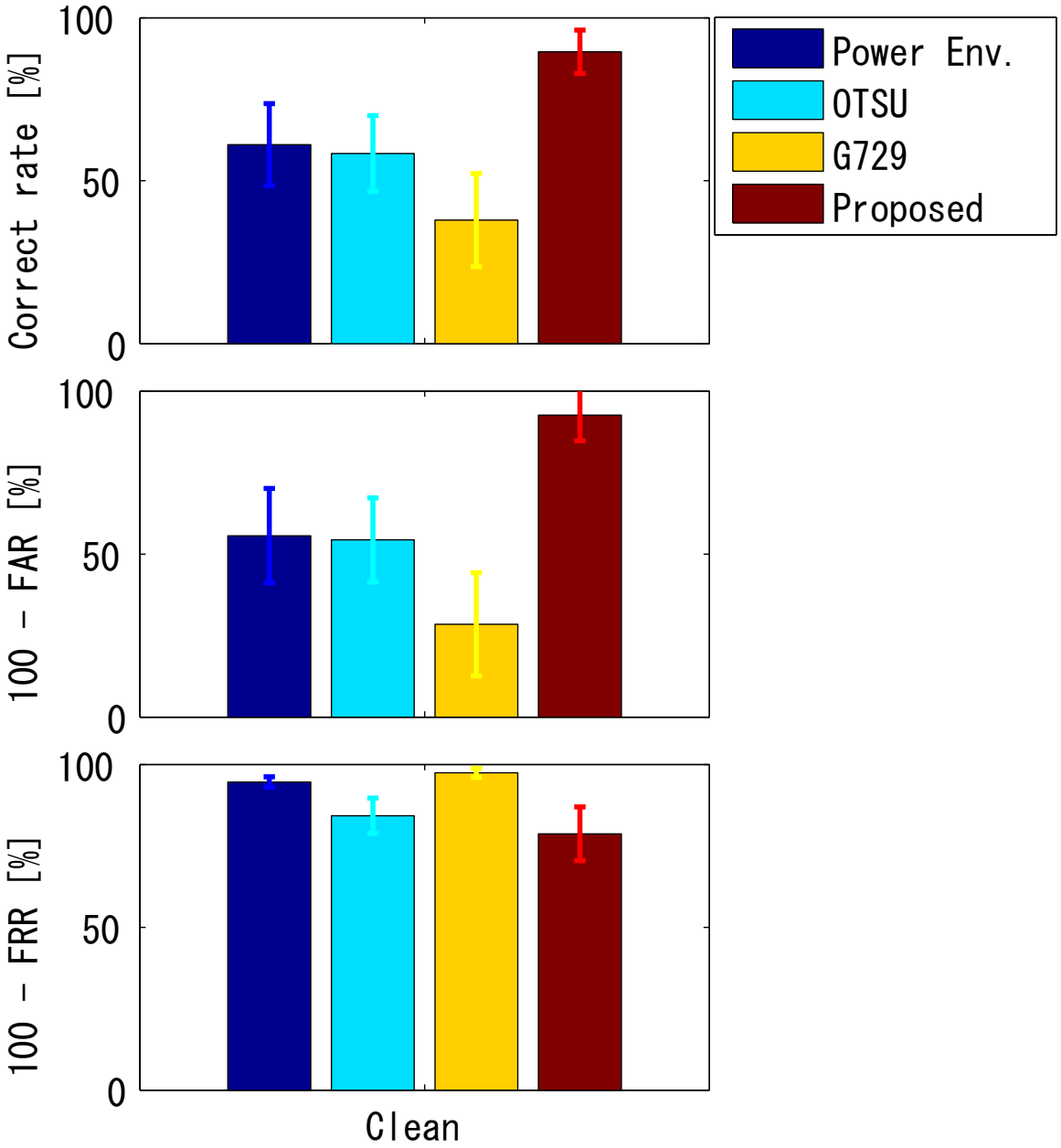

図 6.6: 実験 3 の結果（非音声混入条件）

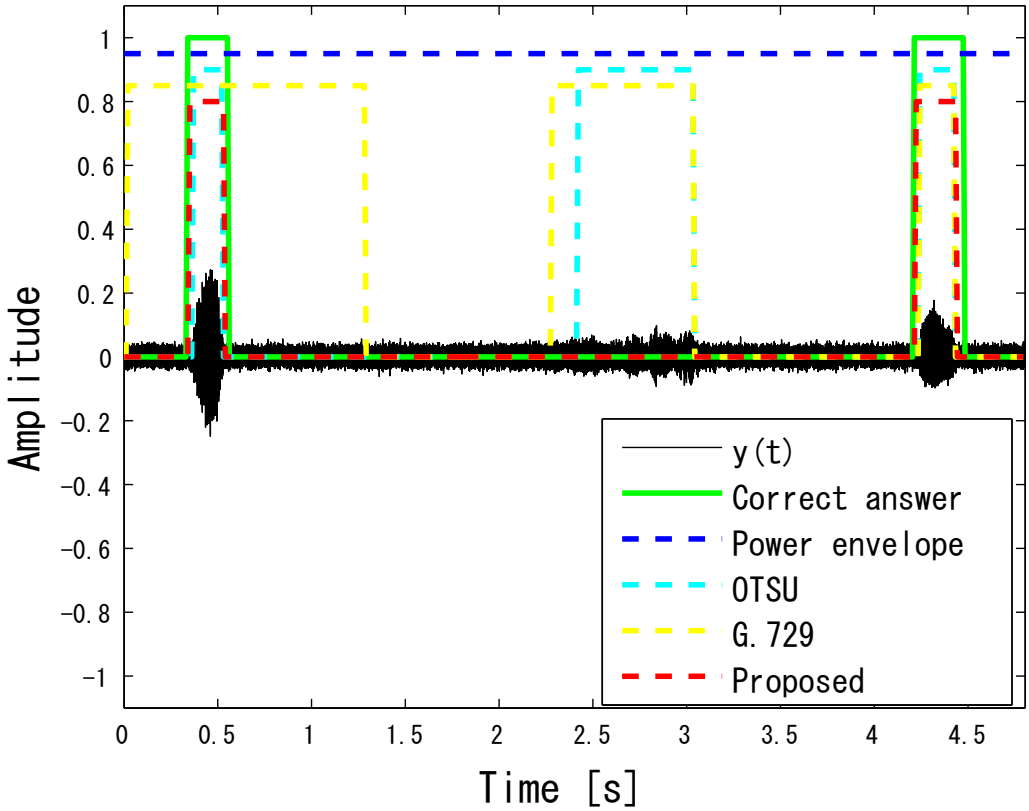

図 6.7: 実験 4 の音声区間検出の例（総合的な条件）

6.1.4　実験 4：総合評価

音声信号と非音声信号からなる信号に，背景雑音として定常雑音を付加した際の VAD の性能評価を行った．この実験により，各 VAD の背景雑音と非音声信号が同時に存在する環境に対する頑健性について比較・検討する．実験 3 で用いた，音声／非音声信号からなる実験刺激に，背景雑音として白色雑音を SNR =20, 10, 0 dB となるように合成し，得られた解析信号から音声区間検出を行い，正答率を調べた．検出された音声区間の例（鳥の鳴声を含む信号に白色雑音を SNR = 10 dB で付加した実験刺激）を図 6.7 に，得られた正答率，FAR ，FRR をそれぞれまとめたものを図 6.8 に示す．

図 6.7 から，パワーエンベロープの閾値処理では，低 SNR 環境下では，実験 2 と同様に，背景雑音の影響で非音声区間の判別ができていなかった．OTSU 法と G.729 法では，非音声区間を音声区間と検出するエラーがみられた．提案法は，音声区間のみをうまく検出できていた．また，図 6.8 より，SNR の増加に伴い，OTSU 法は FAR が増加し FRR が減少し，逆に G.729 法は FRR が増加し FAR が減少することが確認できた．しかし，提案法には，正答率の減少や FAR，FRR の増加は見られなかった．この結果から，提案法は，非音声信号を含み雑音のある環境下でも，高い精度で音声区間検出を行えることが確認できた．

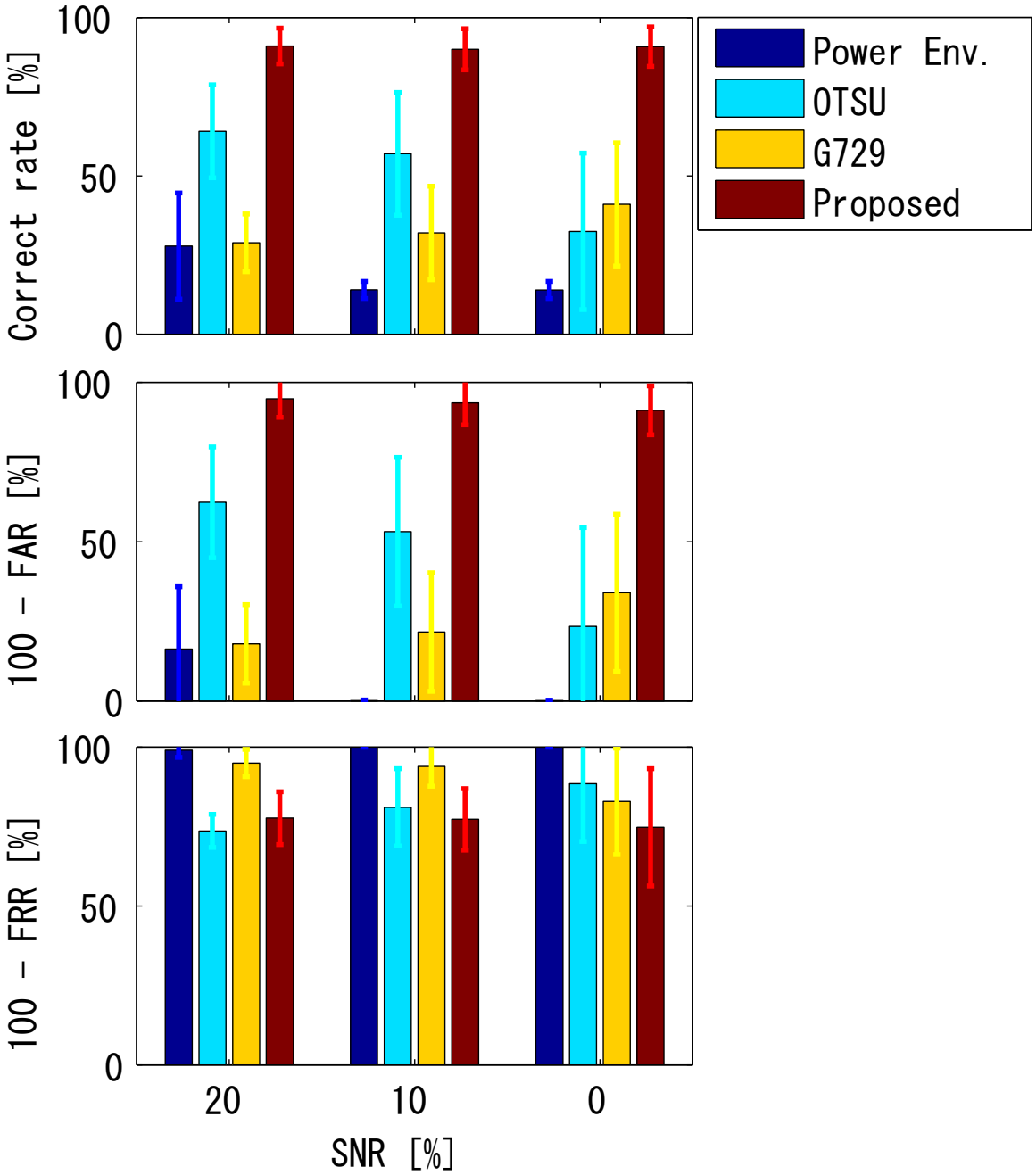

図 6.8: 実験 4 の結果（総合的な条件）

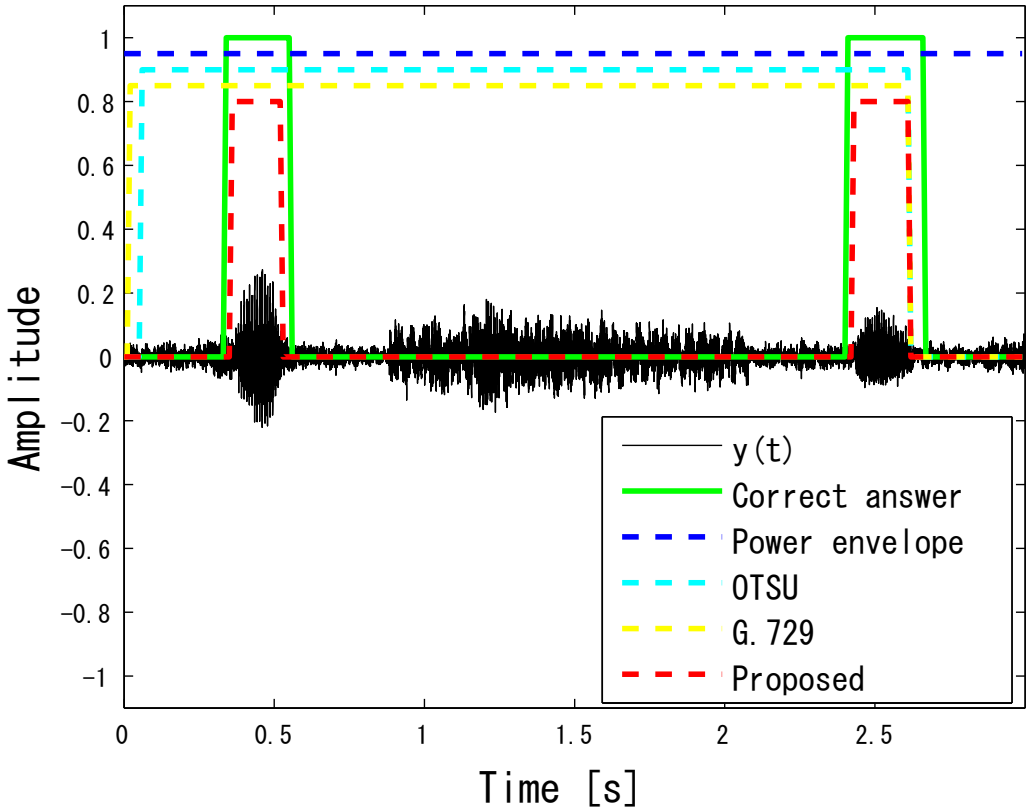

図 6.9: 実験 5 の音声区間検出の例（実環境条件）

6.1.5 実験 5：実環境を想定した実験刺激に対する VAD 精度の評価

音声信号と非音声信号からなる音信号に，背景雑音として非定常信号を付加した際の VAD の性能評価を行った．ここでは，特に実環境を想定した音信号を作成し，その環境で VAD がどの程度正確に音声区間を検出できるかを評価する．

実験 3 で用いた，音声／非音声信号からなる実験刺激に，背景雑音として環境雑音（工場雑音），音楽・演奏音（ロック），鳥の鳴声（ウグイス）を SNR =20, 10, 0 dB となるように合成し，得られた合成信号から音声区間検出を行い，正答率を調べた．検出された音声区間の例（楽器音が混入した信号に工場雑音を SNR = 10 dB で付加した実験刺激）を図 6.9 に，得られた正答率，FAR ，FRR をそれぞれまとめたものを図 6.10 に示す．

図 6.9 から，提案法が，音声区間のみを検出できていることに対して，従来法は非音声区間を含む，ほとんどの区間を音声区間として検出してしまっていることが分かった．また，図 6.10 従来法の FAR が，高 SNR 環境下を含めて高い値になっており，非音声区間をほとんど区別できていないことが分かった．提案法は，SNR = 0 dB のときに，FRR の増加が見られたが，それ以外では，正答率，FAR，FRR ともに，SNR の低い環境であってもほとんど悪化がみられなかった．

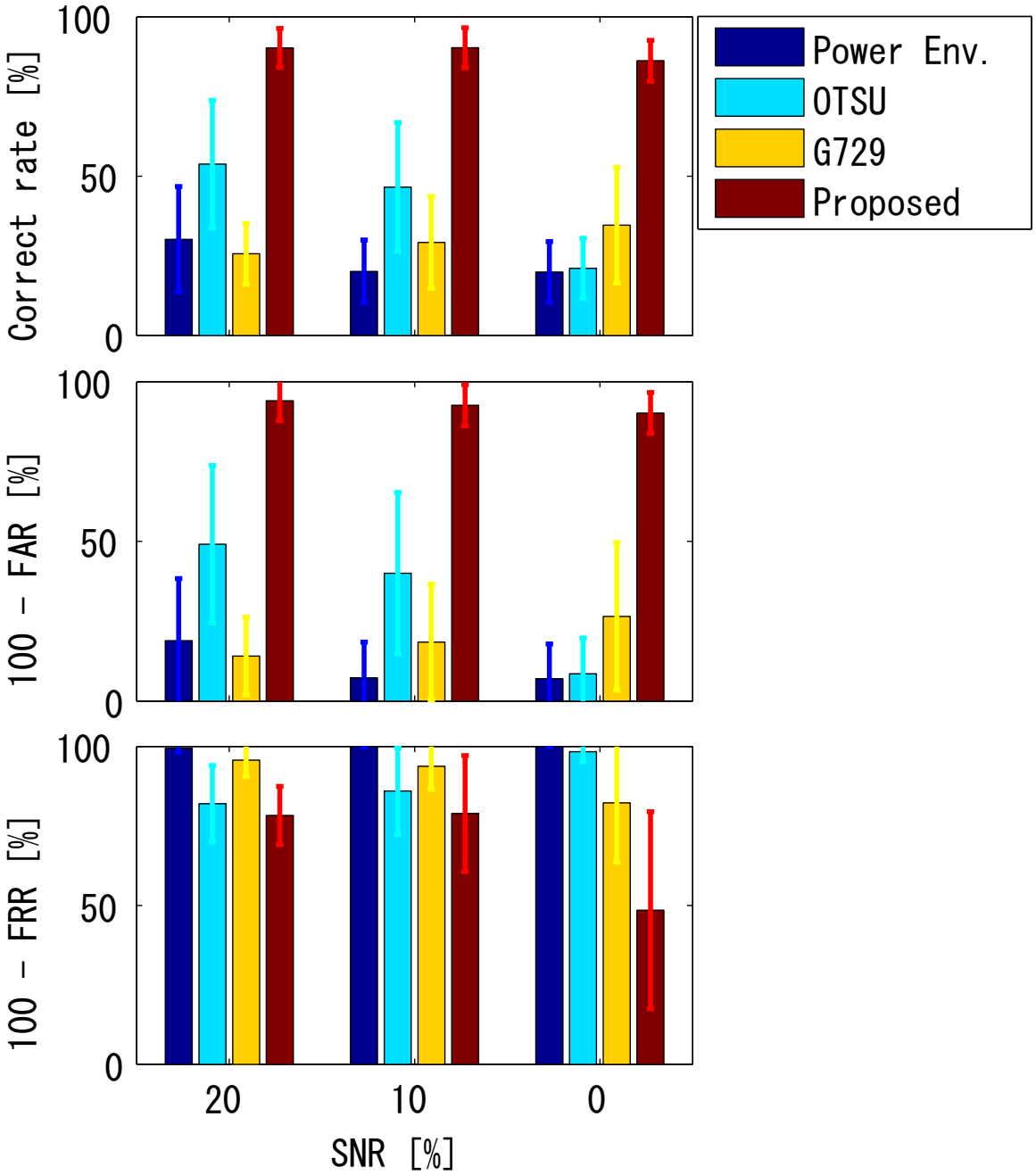

図 6.10: 実験 5 の結果（実環境条件）

第7章　まとめ

7.1　本研究で明らかになったこと

本研究では，経験的モード分解（EMD）を背景雑音除去に，変調スペクトル分析（MSA）を音声／非音声の判別に用いる，二段階の音声区間検出（VAD）法を提案し，その耐雑音性に関する評価を行った．その結果，以下のことを明らかにした．

- EMD を用いることで，正解の音声区間のデータを用いなくとも，定常雑音除去を行うことができる．その結果，背景雑音が存在する環境下でも高い精度で音声区間検出ができる．
- MSA を用い，音声の区間と非音声の区間を判別することで，音声以外の信号が存在する環境でも，高い精度で音声区間の検出ができる．
- EMD と MSA を二段階で用いることで，非音声信号が混入し，背景雑音が存在する環境でも高い精度で音声区間の検出ができる．また，背景雑音として非定常信号が存在するような，実環境を想定した環境であっても，SNR が低い環境でなければ音声区間検出を高い精度で行うことができる．

以上の結果から，背景雑音と非音声信号の混入にロバストな VAD の開発という目的を達成することができた．

7.2　残された課題

提案法は，EMD による雑音除去の後，パワーエンベロープの閾値処理により，音声区間の候補を検出する．そのため，最終的な VAD の結果，とくに誤棄却率（FRR）は，このときの音声区間候補を検出精度の影響を非常に大きく受ける（この後の MSA で音声区間候補の中から非音声区間を棄却することで誤受理率（FAR）を下げることはできるが，候補として検出しそこなった音声区間を改めて検出する機会が存在しないため，音声区間候補検出時より FRR を下げることができないため）．そのため，閾値の調整やパワーエンベロープの閾値処理以外の方法を用いることで，提案法の音声区間検出精度をより高められる可能性がある．また，音声区間候補の検出を行わず，MSA を用い各フレームごと

に音声／非音声の判別を行うことで，雑音除去した信号から直接音声区間検出を行うことも可能であり，その手法についても検討する必要がある．

また，各信号の種類における変調スペクトルの特徴の再検討も行う必要がある．今回，提案法は背景雑音が非定常信号で SNR が低いとき，FRR に増加がみられた．パラメータの値を変更したり，新たな判断基準を検討することで音声区間検出精度の向上が期待できる．

さらに，複数の信号が混在する環境において，EMD により各信号の成分がどの様に IMF に分解されるかも調査する必要がある．この結果と，前述した変調スペクトルの特徴の再検討結果を用いることで，EMD による雑音除去をより正確に行うことができると期待できる．

そして，音声区間検出精度に関しても，より多くの手法と比較する必要がある．背景雑音にロバストな VAD として有名なＥｘｐ AR 法，また，非音声信号の混入にロバストな VAD などとも性能を比較する必要がある．

7.3 今後の展望

提案法では，観測信号を IMF に切り分け，非定常成分のみを再合成することで雑音除去を行っている．そのため，再合成信号には観測信号の非定常成分は依然として混ざったままである．もし，音声と非音声の成分が異なる IMF に分解されていた場合，IMF 上で音声区間検出を行うことができれば，非定常成分の影響を受けずに音声の特徴をより正確にとらえられることが期待できる．それが可能なのか，もし可能ならば，各 IMF で得た音声区間を元の一つの信号の音声区間としてまとめるにはどのような手法をとるべきかを検証する必要がある．

また，本研究では背景雑音，非音声雑音という，主に雑音に対する頑健性に対して研究を行った．しかし，実環境での信号処理を考えると，雑音以外にも残響が大きく影響してくる．耐雑音性だけではなく，耐残響性も考慮した VAD を開発することだできれば，より実環境で正確に音声区間の検出が可能になる．

EMD は計算コストが高いため，リアルタイムに情報を処理するようなアプリケーションには不向きである．しかしメディア処理などリアルタイム処理を必要としないアプリケーションなら，背景雑音，非音声信号に頑健な VAD は大いに利点がある．また，前述の耐残響性を実現できれば，音楽などのデータから音声区間を取り出し，後の検索や要約に用いるメタデータの作成や，会議の発言データから「誰がいつ発言したか」を自動推定する話者決定技術などに応用することが可能であると考えられる．

謝辞

本研究を進めるにあたり，多大な助言と懇切丁寧かつ，熱心なご指導をしていただきました鵜木祐史准教授に心から感謝します．研究を進める中で，様々な助言と熱心な指導をしていただきました赤木正人教授に心から感謝します．本研究に関して多くのご助言をいただきました，宮内良太助教，博士後期課程の木谷俊介氏，森田翔太氏，濱田康弘氏，そして，有意義な討論，御助言を賜った赤木・鵜木研究室の皆様に心から感謝いたします．

参考文献

[1] 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,” **音響誌**, 537–543, 2009.

[2] Remrez, J., Grriz, J. M., and Segura, J. C., “Voice Activity Detection. Fundamentals and Speech Recognition System Robustness,” in Grimm M. and Kroschel, K., Robust Speech Recognition and Understanding, 1–22, 2007.

[3] Otsu, N. “A threshold selection method from gray-level histogram,” *IEEE Trans. Syst. Man.*, SMC(9), 62–66, 1979.

[4] Benyassine, A., Shlomot, E., Su, H.Y., Massaloux, D., Lamblin, C. and Petit, J. P., “ITU-T recommendation G.729 annex B: A silencee compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applocation,” *IEEE Commum. Mag.*, 35, 64–73, 1997.

[5] ETSI EN 301 708 v7.1.1, Digital cellular telecommunications system; Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channeles, 1999.

[6] Kentaro Ishizuka and Hiroko Kato, ”A feature for voice activity detection derived from speech analysis with the exponential autoregressive model,” *ICASSP2006*, 1, 789-792, 2006.

[7] Sawaguchi, T. and Unoki, M., “Investigation of a method of speech signal analysis using empirical mode decomposition and its application,” *J. Signal Processing*, 14(4), 273–276, 2010.

[8] Pek Kimhuoch, 荒井隆行, 金寺登, “変調スペクトルによる音楽が付加された音声の自動検出の検討,” **音響論**, 1-Q-4, 119–122, 2010.

[9] Takeda, K., Sagisaka, Y., Katagiri, S. Abe, M. andKuwabara, H., Speech Detabase User‘s Manual, ATR Technical Report TR-I-0028, 1988.

[10] http://www.speech.cs.cmuedu/comp.speech/Ssction1/Data/nosex.html

[11] 後藤真考, 橋口博樹, 西村拓一, 岡隆一, “RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース,” **情処学論**, 45(3), 728–738, 2004.

[12] http://avocet.zoology.msu.edu/

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

[14] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, Henry H. Liu. “ The Empirical Mode Decomposition and the Hilbert Spectrum for nonlinear and nonstationary time series analysis,” *Proceedings of the Royal Society: Mathematical, Physical and Engineering Sciences*, A454, 903–995, 1998.

[15] Md. Khademul Islam Molla, Keikichi Hirose, “Robust Voiced/Unvoiced Classication of Speech Signal Using Hilbert-Huang Transformation,” *Journal of signal processing*, 12, 6, 473–482, Nov 2008.

[16] Taufig Hasan, Student Member, IEEE, Md. Kamrul Hasan, Senior Member, IEEE, “Suppression of Residual Noise From Speech Signals Using Empirical Mode Decomposition,” *IEEE signal processing letters*, 16, issue 1, 2–5, 2008.

[17] 松田 徹也, 広瀬 啓吉, 峯松 信明, “経験的モード分解による主構造抽出を介 した雑音環境下における音声信号の基本周波数推定,” **電子情報通信学会技術研 究報告**, 109, 57, 49–54, May 2009.

[18] Zhuo-Fu Liu, Zhen-Peng Liao,En-Fang Sang, “Speech Enhancement Based on Hilbert-Huang Transform,” *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International* , 8, 4908–4912, Aug. 2005.

[19] 石田 皓之, 高橋 友和, 井手 一郎, 村瀬 洋, “経験的モード分解を用いた Hilbert ワープ法による指動作映像の認識,” **画像の認識・理解シンポジウム （*MIRU2008*）**, 1162–1168, Jul. 2008.

[20] 斉藤 佑樹, 田中 聡久, 曹 建庭, “経験的モード分解を用いた準脳死患者の脳波解析,” *2007***年電子情報通信学会総合大会論文集**, A–4–7, Mar. 2007.

[21] R. Drullman, J. M. Festem, and R. Plomep, “Effect of Reducing Slow Temporal Modulations on Speech Reception,” *J. Acoust. Soc. Am.*, 95, 2675–2680,1994.

[22] T. Houtgast and H. J. M. Steeneken, "A Review of the MTF Concept in Room Acoustics and its Use for Estimationg Speech Intelligibility in Auditoria," *J. Acoust. Soc. Am.*, 77, 1069–1077, 1985.

[23] B. C. J. Moore, *An Introduction to the Psychology of Hearing* (Academic, New York, 1989).

[24] H. Fletcher, *Speech and Hearing in Communication* (Krieger, Huntington, NY, 1953).

[25] T. Arai, S. Greenberg, "The temporal properties of spoken Japanese are similar to those of English," *Proc. Eurospeech*, 2, 601–604, 2004.

[26] L. Xugang, M. Unoki, R. Isotani, H. Kawai, and S. Nakamura, "Voice activity detection in a regulatized reproducing kernel Hilbert space," *Interspeech 2011*, 3086–3089, Sep. 2011.

[27] L. Xugang, M. Unoki, R. Isotani, H. Kawai, and S. Nakamura, "Adaptive regularization framework for robust voice activity detection," *Interspeech 2011*, 2653–2656, Aug. 2011.

[28] M. Unoki, L. Xugang, R. Petrick, S. Morita, M. Akagi, R. Hoffmann, "Voice activity detection in MTF–based power envelope restoration," *Interspeech 2011*, 2609–2612, Aug. 2011.

発表リスト

- 金井康昭, 澤口知希, 鵜木祐史, “経験的モード分解と変調スペクトル分析を用いた頑健な音声区間検出の検討,” **電気関係学会北陸支部連合大会講演論文集**, G–6, 2011.
- 金井康昭, 澤口知希, 鵜木祐史, “経験的モード分解と変調スペクトル分析を用いた頑健な音声区間検出の検討,” **電子情報通信学会 第２６回信号処理シンポジウム講演論文集**, B1–2, 2011. (信号処理学生奨励賞)
- Y. Kanai, M. Unoki, “Study on Robust Voice Activity Detection Using Empirical Mode Decomposition and Modulation Spectrum Analysis,” *2012 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing*, 2012.（to be appeared）
- 金井康昭, 鵜木祐史, “経験的モード分解と変調スペクトルを用いたロバスト音声区間検出,” *2012* **年春季研究発表会講演論文集**, 2012.（to be appeared）