修士論文

時間周波数領域の瞬時振幅・瞬時周波数を利用した基本周波数推定法の検討

1510755　山口 敏浩

主指導教員　鵜木 祐史
審査委員主査　鵜木 祐史
審査委員　赤木 正人
党 建武

北陸先端科学技術大学院大学
情報科学研究科
(情報科学)

2019 年 8 月

Abstract

The background of this study is the formation of natural speech communication between human and machine. In order to refine the machine and the person's communication, it is necessary to provide the machine with the function of combining naturalness with clarifying what is also naturalness. However, its naturalness is not yet fully understood. Prosodic information such as intonation and accent included in speech is deeply related to the features that form the naturalness of human communication. This prosody combines the three attributes of pitch, loudness, and timber and from the synergistic effect, humans perceive communication as natural. It is pointed out that pitch is causal between the pitch and the fundamental frequency of the sound. In other words, it suggests that the possibility of quantitative understanding of human senses is hidden in fundamental frequency (F0). For this reason, F0 estimation is an indispensable estimation technique for applications such as speech recognition, speech synthesis, voice conversion, voice quality evaluation, and sound source separation. In addition, this is also used to provide objective evaluation measures. F0 estimation is well-known to be one of the studies that is considered extremely difficult, for the following two reasons. The first reason is that F0 cannot be observed directly. The second reason is that the vocal cord vibration that is the source of F0 is quasi-periodic instead of a constant periodic vibration. For these reasons, the F0 estimation requires robustness and accuracy. This study aims to discuss how to propose a method for robustly and precisely estimating F0 of speech under noisy or reverberant conditions. The F0 of speech can be utilized as a significant feature to represent the source information (glottal waveform) of speech sound in various speech-signal processes. These are in speech analysis/synthesis systems, automatic speech recognition (ASR) systems, and speech emphasis methods. Therefore, a particularly important issue in these applications is to robustly and accurately estimate the F0 of target speech in real environments. The F0 estimation method with high performance has been proposed recently, which realized either robustness or accuracy. For example, those representative examples are FreeDAM (Fundamental fRequency Estimation mEthod using Demodulation of Amplitude Modulation) with high robustness, and a method with high accuracy proposed by Dhiman et al. FreeDAM is an F0 estimation method inspired by the concept of human's pitch perception called "Missing fundamental" and Modulation Transfer Function (MTF). The method proposed by Dhiman et al. is a method to extract temporal variation of F0 smoothly from a sound spectrogram. The Riesz transform is used to transform a sound spectrogram into instantaneous information. Since this instantaneous information is a complex signal composed of an amplitude term and a phase term, instantaneous

amplitude (IA) and instantaneous frequency (IF) is decomposed from each term. However, both FreeDAM with its focus on robustness, and the method proposed by Dhiman et al. with high accuracy have their own problems. From this reason, the aim of this research is to make a basic study to solve each other's problems in that way, which each other's advantages strengthen each other's problems. That can realize the F0 estimation method that has both robustness and accuracy. The results obtained in this study contribute to providing key technologies to realize natural speech communication between human and machine.

目 次

図 目 次

表 目 次

第1章　序論

1.1　はじめに

かつて，音声コミュニケーションと言えば，専らヒトとヒトとの間でとり交わされてきたが，近年のそのありかたは様変わりし，相手はヒトに限定されなくなりつつある．コールセンターにおける自動音声案内，人工知能アシスタントやおもてなしロボットによる音声サービスの提供など，その完成度はさておき，機械がコミュニケーション機能を備え，その代わりにヒトを相手にコミュニケーションをとることが，実現されつつある時代が到来しようとしている．しかし，それらのコミュニケーションの利用先は，現時点では極めて単純な要件のやり取りに限定されており，ヒトとヒトのコミュニケーションのような自然さが欠けている．

自然な発話には，個人性（性別，年齢）を表現する，抑揚（音調，イントネーション）やアクセントなど，音声の時間的な変化として見られる韻律情報が欠かせない [1]．韻律や個人性は，音の高さ（ピッチ），音の大きさ（ラウドネス），音色（ティンバー）と呼ばれる聴知覚の三大属性が三位一体となり，相手に伝わる情報であるとされている [2]．これは，音声コミュニケーションの本質が，単に言語的な情報を伝えることではなく，話者の意図を伝えることであるからである．感情的な情報を韻律にのせることで，相手の聴覚が刺激を受け，話者の意図がはじめて相手に伝わる．伝えたい意図に適する韻律情報を，音声にのせて発話することで，音声の持つ質的な情報量が格段に増し，聴者は知覚した言語情報以上に，話者の意図を推し量ることを可能とする．自然な音声コミュニケーションにおいて，この韻律情報は不可欠な要素であり，音声の生成と知覚は，ヒトの高度な技術の結晶である [3]．

一方，現段階でのヒトと機械の音声コミュニケーションは，必ずしも自然ではない．確かにこれらの音声システムが提供する合成や分析機能は，言語的な側面に限定すれば，完成度が高いと言える．例えば，前述したおもてなしロボットについては，情報通信端末の契約店や，ホテルの受付などで，身近な接客に活用されつつあり，ヒトとヒトではなく，ヒトと機械のコミュニケーションを図る第一歩が，機能として具現化されていると言える [4, 5]．だが，ヒトとヒトとの音声コミュニケーションにおける自然さや，潜在意識下で行われる巧みな非言語情報のやりとりと比較すると，より高度化，洗練化の余地があることは明白である．

ヒトと機械の音声コミュニケーションを，ヒトとヒトの音声コミュニケーション並みに，非言語情報が盛り込まれた高度なものにする第一歩は，ヒトが知覚す

る “音声の自然さを形成するメカニズム” を明らかにすることである．聴覚の刺激を通してこの正体を探り，追求すべき自然さを明らかにすることで，音声コミュニケーションが進むべき目標が明確になる．議論する上で忘れてはならないことは，この “自然さ” いう課題が，認知科学や生理学などの自然科学と，物理学や音声学の人間科学の学問分野の狭間にあることである [2]．音声の自然さとは，ヒトの感覚に基づく心理量で，評価者の主観により千差万別であり，最終的にはヒトが判断する感覚である [6]．この自然さが心理的な要素を持っていることを認識しつつも，自然さの正体を明らかにする上で，何らかの公平な尺度が求められるということが，この課題の難解な点である．

これを踏まえて本研究は，基本周波数（Fundamental Frequency: F0）に着目した．F0 は，おおよそピッチに関係することが知られている物理量であり，その時間的変化には，韻律に係わる特徴が含まれていることが報告されている [7]．このため，物理量である F0 に着目し，音声の観測波形から F0 を推定する方法について，深く検討することを本研究の対象とした．F0 は，自然な音声コミュニケーションの形成だけでなく，図 1.1 に示す声質評価や音声合成といった重要な技術にも利用される．本研究で得られる成果は，ヒトと機械の自然なコミュニケーションの実現に向けて，要素技術の提供に貢献するものである．

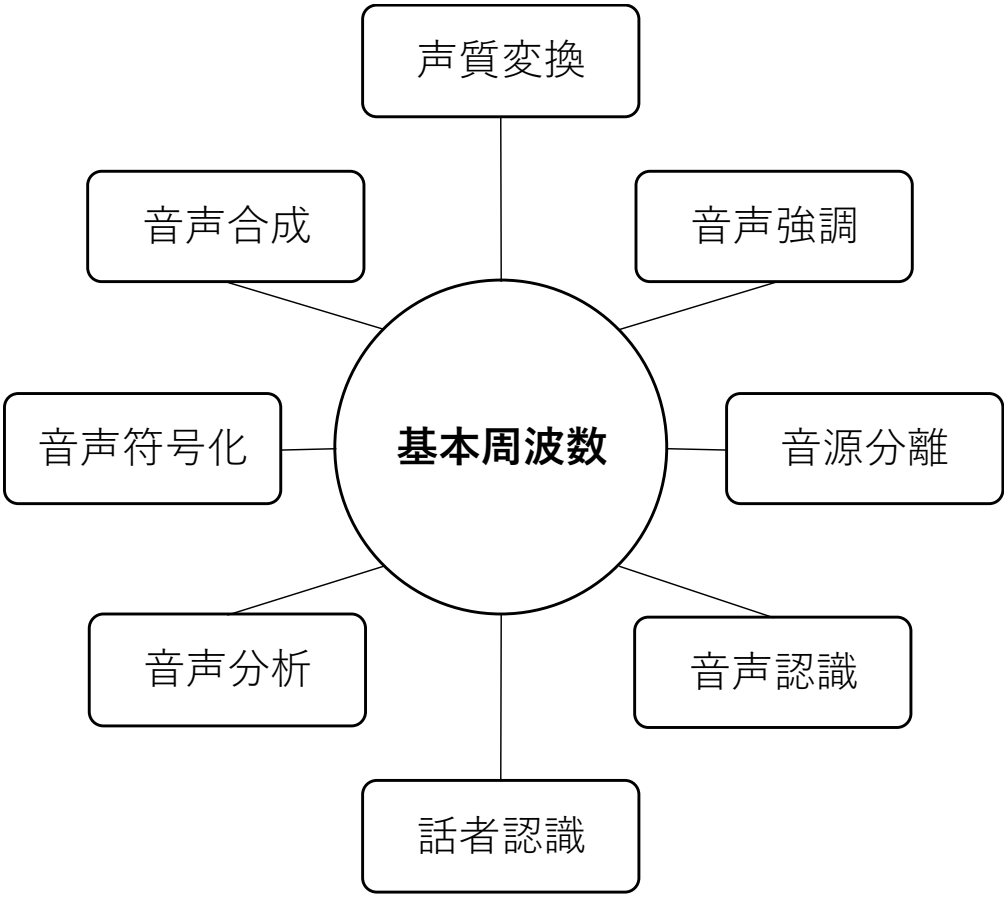

図 1.1: 基本周波数推定法の応用例

1.2 研究の背景

F0 推定とは，観測した音声波形（出力）から，何らかの方法で声帯波形（入力）を推理することである．F0 は，声帯の振動周期の逆数であり，母音に見られる周波数である．ヒトの母音の生成過程を説明するには，Fant が体系化した理論に基づく音源フィルタの利用が，端的であり望ましい [8]．

図 1.2 は，音源フィルタモデルの機能概要である．現在も F0 推定に係る議論に，しばしば取り上げられるモデルであり，音情報信号処理を支える理論と，係わりが深いとされている [9, 10]．声帯音源 $G(\omega)$ と声道フィルタ $H(\omega)$ が独立である仮定のもとに，ヒトが生成する周波数領域 ω における音声波形 $S(\omega)$ は，式 (1.1) に示す線形モデルとして表現される [9]．

$$S(\omega) = H(\omega)\ G(\omega) \tag{1.1}$$

肺から流入する空気による準周期的な声帯振動が，声帯音源である．最も扱いやすい声帯音源は，正弦波による F0 とその高調波で構成される調波複合音を仮定したものである．声道フィルタの周波数特性は，共振周波数（フォルマント周波数）と周波数の傾斜特性（スペクトル包絡）であり，性別や年齢などに応じて異なる声道の形状によるものである．しかし，声道フィルタや声帯音源は，ヒトの体内器官であり，直接計測することは困難であるから，より精度の高い F0 推定法が求められるわけである．現在，F0 の複雑な時間的変化を推定できる高精度な F0 推定法が提案されているが，その原点となる F0 推定法の一例であるゼロ交差法も，音源フィルタに基づく方法である [11]．

一方，音源フィルタモデルの考え方は，工学的にも応用できる．例えば，声帯振動による基本周波数を搬送波として，声帯振動の準周期的な時間変化やフォルマント周波数を情報信号に見立てると，周波数変調（Frequency Modulation: FM）や振幅変調（Amplitude Modulation: AM）と見なすことが可能であり，無線通信分野の変調理論や復調理論が適用できる．つまり音声生成と知覚の過程を，それぞれ送信機と受信機として工学的に解釈できる [12, 13]．

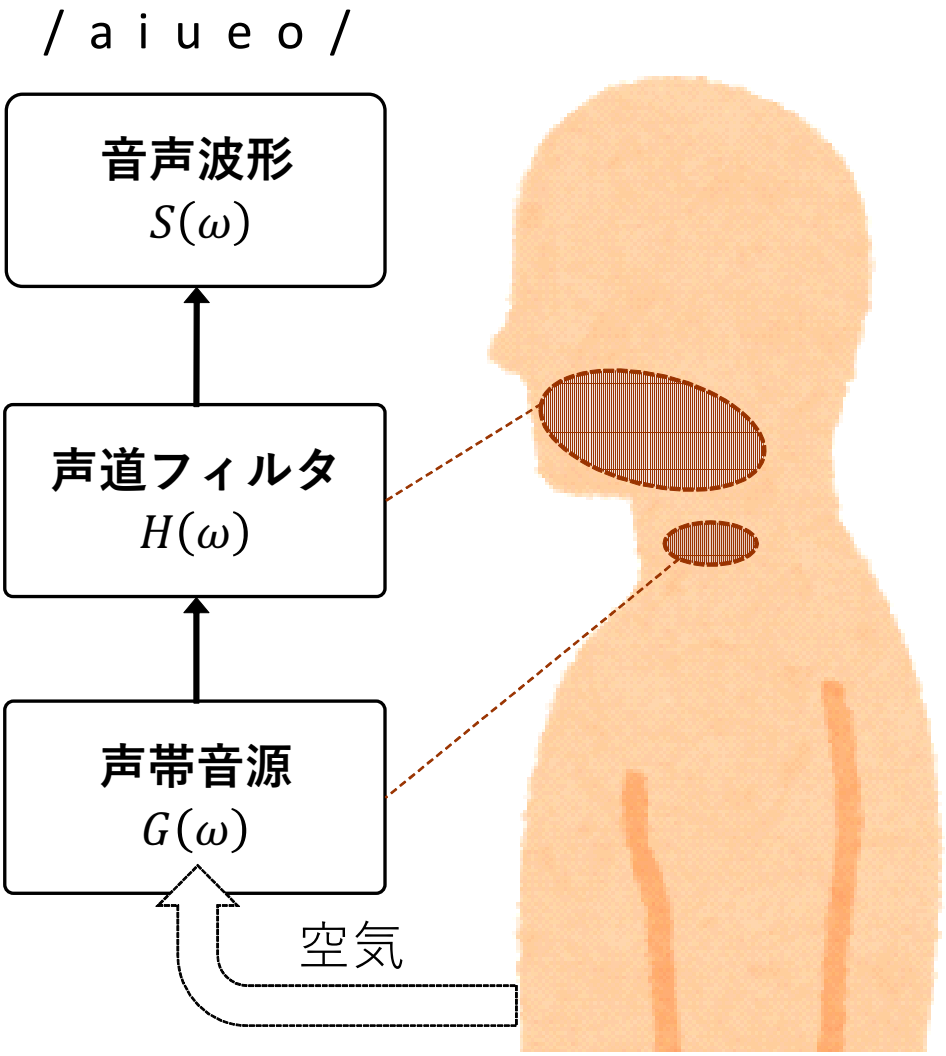

図 1.2: 音源フィルタモデルの概要

1.3 研究の目的

本研究の目的は，正確性と頑健性を備えるF0推定法を検討することである．F0を推定する環境は，必ずしも静音な環境であるとは限らないためである．つまり，雑音や残響の影響を受けた観測波形からF0を正確に推定することは，従来のF0推定法だけでは困難であるからだ．例えば，屋外でF0を推定する場合，求められる耐性は雑音と残響に頑健であることである．加えて，F0の時間的変動は，特に韻律に係わるとされており，自然な音声を表現する情報として欠かせない[1]．雑音や残響が付加された観測波形から，F0の滑らかな時間的変動を正確に推定する方法を検討し，さらにその評価結果から妥当性が示せれば，F0推定の実用化に寄与できる．そのため，まずは近年提案されたF0推定法の中から，正確性や頑健性が高いと報告される推定法の原理を調査する．これらの推定法の実装により，正確性と頑健性に有利に働く特徴を掘り下げ，互いの弱点が合理的に克服できる方法の骨格を形成する．先行研究の調査から得られる知見に基づき，最善の方策を多面的に検討し，計算機による実装と評価により，実用的なF0推定法の有効性を示すことが重要である．

1.4 論文の構成

図1.3は，本論文の構成を示すものである．

本章は序論である．F0に関わる応用技術と，本研究の背景と目的を述べる．

第2章では，代表的なF0推定法と近年の提案法を俯瞰し，代表的なF0推定法に立脚する先行研究が抱える問題点を述べる．

第3章では，先行研究から得られる知見に基づき，正確性を損なわずに頑健性を高めるF0推定の改良法を，原理に基づく観点から提案する．

第4章では，第3章の改良法を計算機に実装し，観測波形を入力に与えた場合における，正確性と頑健性に係わる評価結果と考察を述べる．

第5章は結論である．本研究が明らかにしたことと，残された課題を整理して述べる．

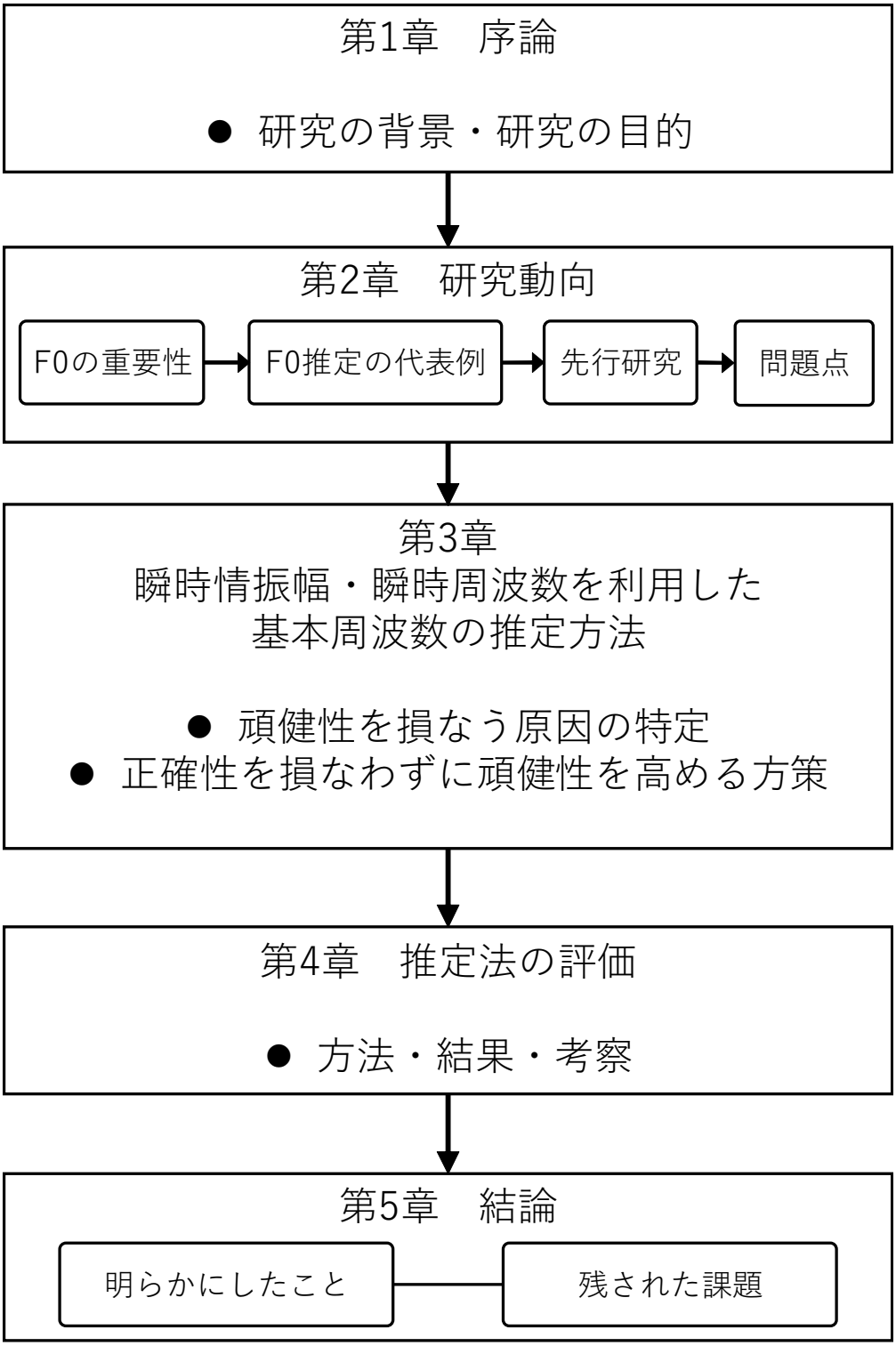
第1章　序論
● 研究の背景・研究の目的
第2章　研究動向
F0の重要性
F0推定の代表例
先行研究
問題点
第3章
瞬時情振幅・瞬時周波数を利用した
基本周波数の推定方法
● 頑健性を損なう原因の特定
● 正確性を損なわずに頑健性を高める方策
第4章　推定法の評価
● 方法・結果・考察
第5章　結論
明らかにしたこと
残された課題

図 1.3: 論文の構成

第2章　研究動向

2.1　基本周波数推定法の重要性

実用的なF0推定法が確立されると，音声コミュニケーションに係わる問題を解決できる可能性が高まる．このためには，F0が準周期性であることを踏まえたうえで，正確性，頑健性を満たすF0推定法が必要である．

2.1.1　準周期性

準周期性とは，声帯振動の周期が時々刻々と変化し，定周期ではない性質である．F0の準周期性は，平均的な周波数を中心として，正負に不規則に変動することが知られている[14]．観測波形から分析する際に，変動する周期を考慮して，時間分解能や周波数分解能を適切に設定することが重要である．

2.1.2　正確性

正確性とは，F0の時間的な変動を観測波形から抽出し，その軌跡を詳細に推定できる性能である．F0は，声帯振動の周期の逆数として定義される周波数である．F0の準周期性の性質から，その周期が時間と共に変化することを前提として，声帯振動に起因する変動成分を，少ない誤差で観測波形から推定しなければならない．音声認識や個人性の識別などの観点では，F0の時間的変動を忠実に推定できることが望ましい．このためには，フォルマント周波数とスペクトル包絡にみられる周波数特性を，観測波形から取り除かなければ，声帯音源が正確に推定できない．

2.1.3　頑健性

頑健性とは，雑音や残響などの外乱に，F0の推定結果が左右されないことを示す性能である．たとえ残響や雑音の影響が観測波形に混入したとしても，F0の周波数成分だけを抽出する選択性と，残響や雑音に対する耐性の両立が不可欠である．

2.2 代表的な基本周波数推定法

F0 は，周期性または調波性という，単位の異なる物理量として観測できる．そのため F0 推定処理の領域は，時間，周波数，または時間周波数に大別される．表 2.1 に列挙した推定法は，これまでに提案されてきた，高い正確性を有する代表的な F0 推定法の一例である [14].

- YIN[15]
  YIN は，時間波形の周期性を利用する F0 推定法の代表例である．波形を観測する窓長の最適値が，F0 推定精度に依存する．

- SWIPE (A Sawtooth Waveform Inspired Pitch Estimator)[16]
  SWIPE は，スペクトル構造の調波性を利用する F0 推定法の代表例である．誤差を低減するための計算コストが大きいとされている [17].

- TEMPO (Time domain Excitation extractor using a Minimum Perterbation Operator)[18]
  TEMPO は，代表的な音声合成分析法 STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) を構成する分析部である．瞬時周波数（Instantaneous Frequency: IF）を利用して，フィルタの中心周波数と IF の写像空間に基づく固定点が利用される．

- PHIA (Periodicity and Harmonicity using Instantaneous Amplitude)[19]
  PHIA は，周期性と調波性を利用する F0 推定法である．瞬時振幅（Instantaneous Amplitude: IA）と IF を利用し，得られる周期性と調波性の確からしさを，Dempster の結合規則 [20, 21] で統合し，F0 推定値を決定する．

これら代表的な F0 推定法の正確性は，静音な音環境で有効である．しかし，背景雑音や残響といった音環境に対する頑健性について，実用化の視点から，深く検討する必要があるとされている [14].

表 2.1: 代表的な基本周波数推定法の特徴

推定法	処理領域	参考文献
YIN	時間（周期性）	[15]
SWIPE	周波数（調波性）	[16, 17]
TEMPO	周波数	[18]
PHIA	時間（周期性）・周波数（調波性）	[19, 20, 21]

2.3 頑健性の高い基本周波数推定法

近年，F0 推定法の頑健性を向上させる試みとして，FreeDAM (Fundamental fRequency Estimation mEthod using Demodulation of Amplitude Modulation) が提案された [22, 23]．FreeDAM が着目した原理は，ミッシングファンダメンタルと呼ばれる，ヒトのピッチ知覚である．ミッシングファンダメンタルは，複合音の基本波の有無によらず，ヒトが知覚するピッチが変わらないとされる現象であり，このことは，Schouten の実験などで示されている [7, 25]．さらに FreeDAM は，その方法に音声変調伝達関数（Modulation Transfer Function: MTF）の概念を取り入れた．MTF の概念に基づくと，雑音や残響の影響を受けることで，AM 音の変調度は低下するが，情報信号の周期性が保持されることが示されている [26]．したがって，音の調波構造を AM 変調のスペクトラムに見立てると，AM の復調技術が適用できる．つまり F0 の推定処理は，隣り合う 3 本 1 組の任意の調波スペクトルを抜き出し，両側波帯に相当する振幅包絡線情報を抽出することである．

F0 の推定処理は，図 2.1 に示す機能で構成され，時間波形から切り出すフレーム毎に実行される [24]．

1). 帯域制限
フレーム内に存在する調波を，3 本 1 組として抽出する処理である．調波の抽出に，帯域制限フィルタ（Band Pass Filter: BPF）を利用する．

2). 閾値処理
帯域内の雑音を抑圧する処理である．抽出した調波から，中央値を下回る成分を減じて，フレーム内に重畳した雑音を抑圧する．

3). 同期検波
AM 音から包絡線情報を抽出する処理である．3 本 1 組の中央の調波に同期する搬送波を再生する．再生した搬送波と AM 音を乗算し，包絡線情報を抽出する．包絡線情報の抽出に，低域通過フィルタ（Low Pass Filter: LPF）を利用する．

4). 逆フィルタリング
MTF の概念に基づく，変調度を改善する処理である．雑音や残響の影響を受けることで低下した包絡線に逆フィルタを適用し，波高値を回復する [27]．

5). 特徴抽出・最終判定
抽出した F0 を，フレームの推定値として決定する処理である．推定値は，回復した包絡線における，直流成分の含有率，周波数一致率，時間波形の形状一致率から決定される．

FreeDAM の頑健性は，調波構造が外乱に対して頑健である性質に起因するものである [24]．

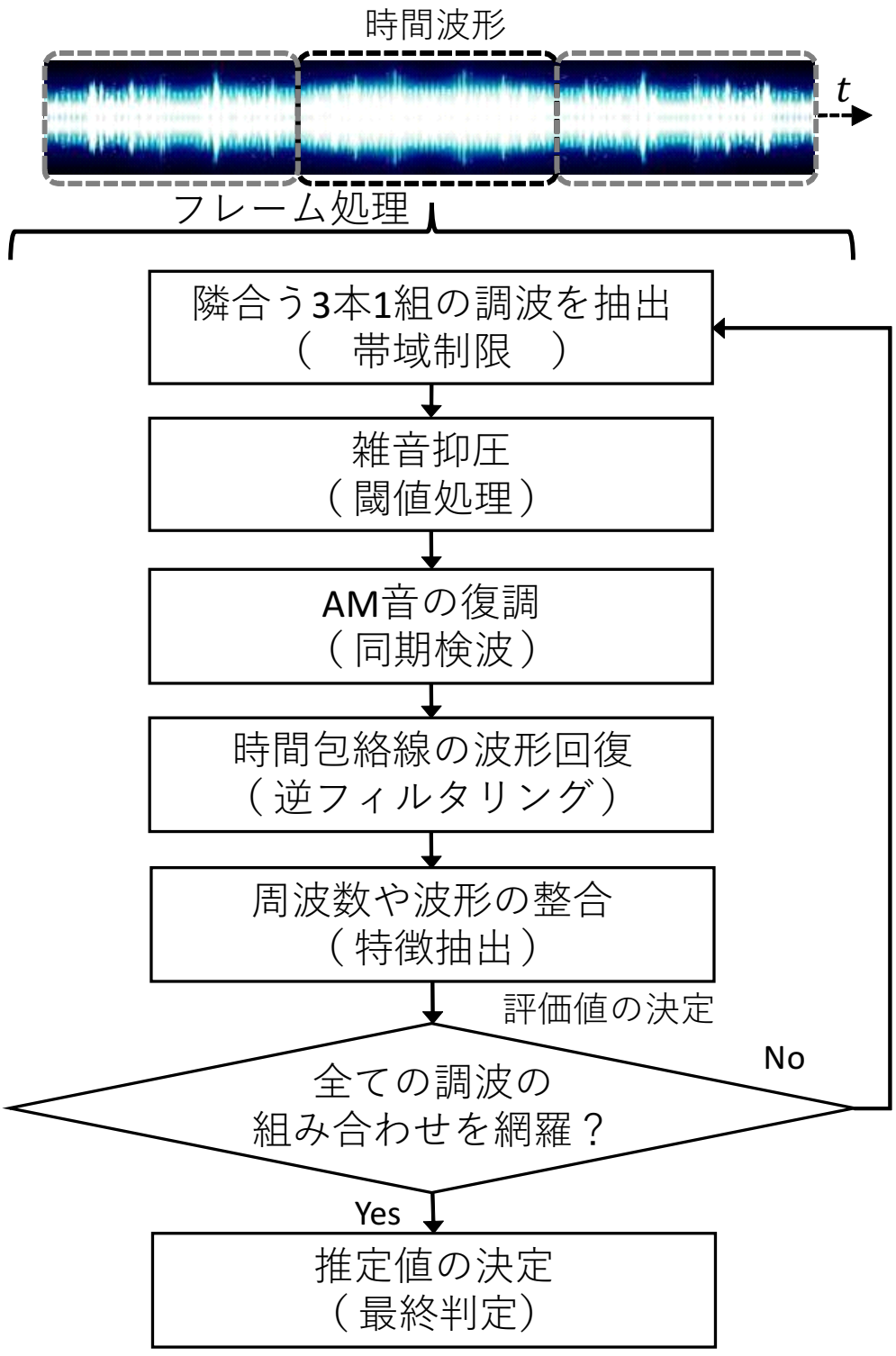

図 2.1: 推定法の概略手順（FreeDAM）

2.4 正確性の高い基本周波数推定法

一方，正確性を向上させる F0 推定法が，Dhiman らによって提案されている [28, 29]．この方法は，複素 Riesz 変換（Compelx Riesz Transform: CRT）を利用して，サウンドスペクトログラムから F0 の時間変動を抽出する方法である．CRT は，Hilbert 変換を多次元信号に拡張した概念であり，信号の位相を 90 度シフトする性質が受け継がれる [29]．Dhiman らの方法は，この性質に着目してサウンドスペクトログラムを瞬時複素信号（Instantaneous Complex Signal: ICS）に変換し，IA と IF に分解するものである．IA と IF の支配的な成分は，それぞれ声道フィルタと声帯音源である．つまり，サウンドスペクトログラムから声道フィルタの影響を取り除き，F0 の滑らかな時間的変動が IF の軌跡として抽出される．Dhiman らの方法は，図 2.2 に示す機能で構成される．最終判定を除く処理は，サウンドスペクトログラムの局所的な領域（パッチ）毎に実行される [32]．

1). 帯域制限
空間周波数 [30] 領域に BPF を適用し，帯域通過成分として IA と IF を抽出する処理である．適用する BPF は，10 次のバタワースフィルタである．

2). 複素 Riesz 変換
抽出した帯域通過成分を，ICS に変換する処理である．ICS の実部は，帯域通過成分である．ICS の虚部は，Riesz カーネルと呼ばれるフィルタ核 [31] と，帯域通過成分の畳み込みである．

3). 固有値算出
構造テンソルの固有値から，コヒーレンスマップを生成する処理である．コヒーレンスマップは，時間周波数領域の干渉の度合いを表す指標であり，調波性が選別できるとされている [33, 34]．構造テンソルの要素は，ICS とガウス平滑化フィルタ [35] の畳み込みである．

4). 螺旋操作
構造テンソルの最大固有値に対応する固有ベクトルを，ICS の虚部に乗じる操作である．この操作で ICS の局所方位を補正することで，振幅項と位相項からそれぞれ IA と IF が分離できる [28]．

5). 最終判定
重み付けされた IF（Weighted Instantaneous Frequency: WIF）から，F0 の推定値を判定する処理である．最終判定に利用する WIF は，IF にコヒーレンスマップを乗じ，0.05 の閾値を上回る値である [28]．F0 の推定値は，WIF に表れる調波性の平均値から算出し，時間フレーム毎に決定される．

Dhiman らの方法は，音源フィルタモデルの概念を，2 次元の時間周波数領域に拡張 [36] し，CRT で得られる瞬時情報に基づき，正確性を高める F0 推定法である．

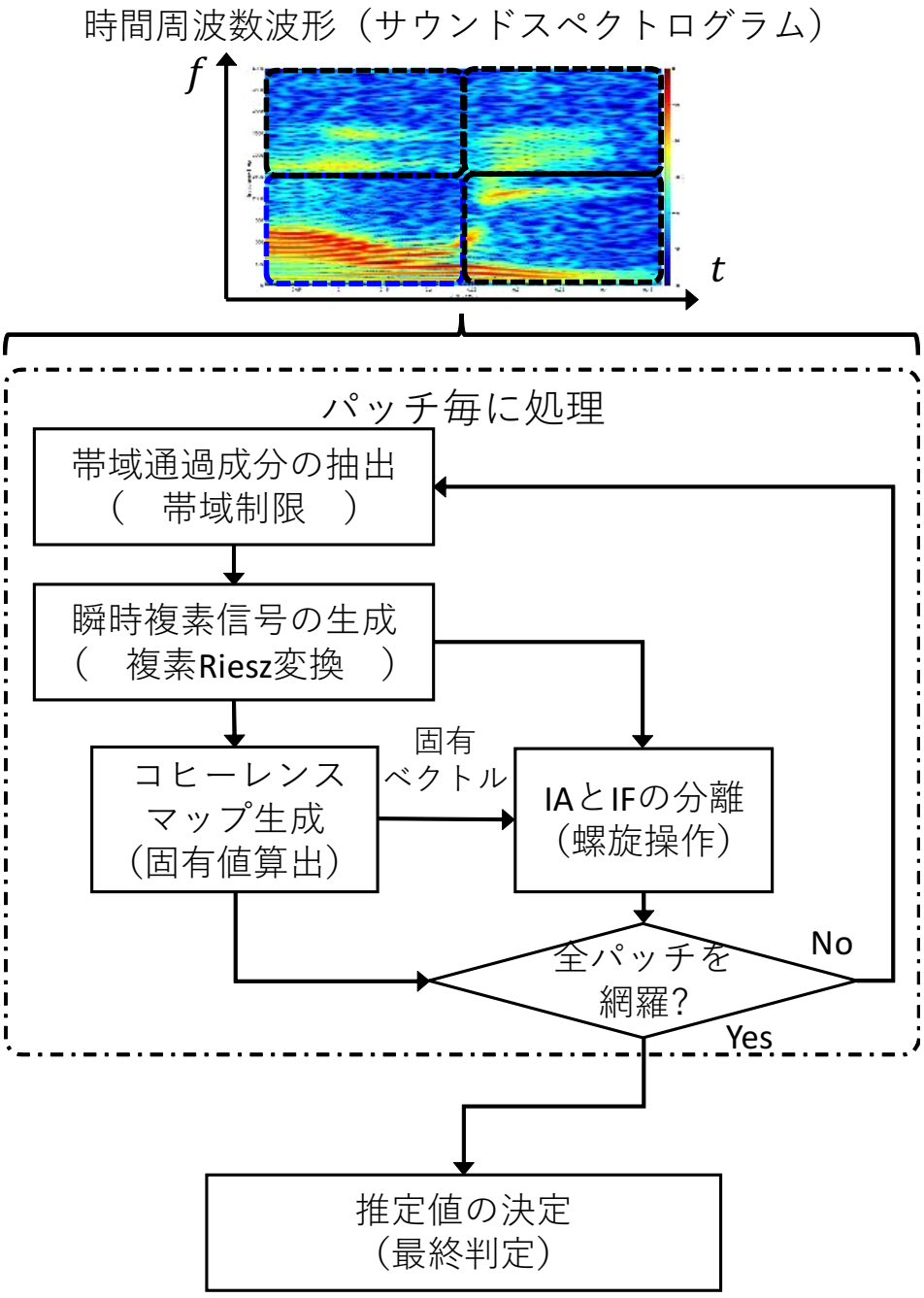

図 2.2: 推定法の概略手順（Dhiman らの方法）

2.5 問題点

頑健性の高い FreeDAM と，正確性の高い Dhiman らの方法は，それぞれ問題点を抱えている．図 2.3 は，FreeDAM による F0 推定値の例示である．FreeDAM は，AM 音のピッチ知覚に基づく雑音残響に頑健な F0 推定法であるが，正確性の観点から問題点が残されている．実線で示す FreeDAM の F0 推定値は，フレームシフト長である 50 msec の間隔で一定であり，F0 の基準値である破線に比べて直線的である．根底にある FreeDAM の問題点は，復調区間が頑健性を高める一方で，その区間の長さが正確性の低下を招くというジレンマである．AM 音の復調に必要な信号区間は，少なくとも振幅包絡線の周期以上であることが前提となる．ところが，振幅包絡線の周期は F0 の基本周期に比べて長く，F0 に応じて自ずと復調区間を広げなければならない．また，フォルマント周波数や有声区間と無声区間の識別が考慮されていないことも，残された課題として実用性の観点から取り上げられている [22]．

一方，Dhiman らの方法の頑健性は，雑音環境や残響環境でどの程度頑健であるかは，明らかにされていない．しかし，Dhiman らの方法の正確性は，代表的な方法と遜色ないことが報告されており，FreeDAM の正確性の向上に寄与する効果が，基本原理に見込まれる．このため次の章では，Dhiman らの方法の頑健性を確認する．Dhiman らの方法の頑健性の確認は，F0 推定法の正確性と頑健性の両立に向けた取り組みの第一歩である．

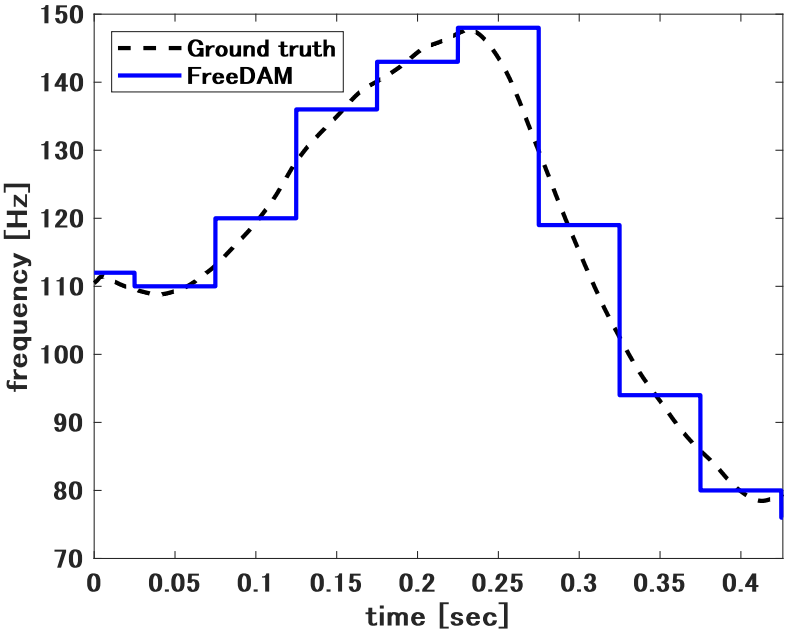

図 2.3: FreeDAM による F0 推定値の例示（静音環境）

第3章　瞬時振幅・瞬時周波数を利用した基本周波数の推定法

3.1　正確性が高い基本周波数推定法の頑健性評価

計算機上に Dhiman らの F0 推定法を実装し，頑健性を評価する．

3.1.1　評価条件

- 実装方針
  原著論文に提示された Dhiman らの F0 推定法の範囲は，原理に係わる必要最小限の情報にとどめられており，実装に係わる処理や諸元の一部が，開示されていない．このため本研究では，Dhiman らの方法の安定的な動作を優先するために，“自己相関法 [38] による安定的なピーク検出” と “コヒーレンスマップ生成時における閾値の設定” を代替的に実装する．

- 諸元
  表 3.1 は，実装に適用する諸元である．

- 入力信号
  入力信号は，図 3.1 に示すサウンドスペクトログラムである．このサウンドスペクトログラムは，観測波形から有声区間を切り出し，短時間フーリエ変換（Short Term Fourier Transform: STFT）で生成したものである．

- 評価指標
  頑健性の指標である，Gross Pitch Error (GPE)[40] を採用する．GPE は，許容誤差率 p を上回るフレームが存在する割り合いであり，式（3.2）で算出する．F0 の基準値は TEMPO の推定値，本節で採用する許容誤差率は 20%とする．

$$GPE_p = \frac{1}{N}\sum_{n=1}^{N} e_p(n) \tag{3.1}$$

$$e_p(n) = \begin{cases} 1, & \left|\dfrac{\hat{f}_0(n) - f_0(n)}{f_0(n)}\right| > p \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

ここで，N はSTFTの全フレーム数，$\hat{f}_0(n)$ と $f_0(n)$ は，それぞれ n 番目の時間フレームにおけるF0の推定値と，F0の基準値である．ただし，図3.2は，静音環境におけるDhimanらの方法によるF0推定値である．GPEの値より，許容誤差率20%を上回るフレームの割り合いは，全フレームの5.5%である．

- 雑音環境と残響環境
  雑音環境のモデルは，加法性白色ガウス雑音であり，信号対雑音電力比（Signal to Noise Ratio: SNR）は，0 dBとする．
  残響環境は，式(3.3)で生成する統計的室内インパルス応答 $h(t)$ と，実音声の畳み込みで構築する [24, 39]．

$$h(t) \;=\; a \exp\left(\frac{-6.9\ t}{T_R}\right) n(t), \tag{3.3}$$

$$a \;=\; \sqrt{\frac{1}{\int_0^T \exp\left(\frac{-13.8\ t}{T_R}\right)}} \tag{3.4}$$

  ここで，残響時間 T_R は，振幅の減衰量が60 dBに到達するまでの時間であり，2.0 secとする．$n(t)$ は，正規分布に従う乱数である．

表 3.1: 設定した諸元

項目	諸元
窓	Hanning
分析窓長	30 msec
シフト長	1 msec
標本化周波数	20 kHz
FFT長	8192
推定に利用する調波の範囲	60 Hz - 1000 Hz
コヒーレンスマップの閾値	10^{-6}
ガウスカーネルの標準偏差	0.5
パッチのサイズ	600 Hz × 100 msec

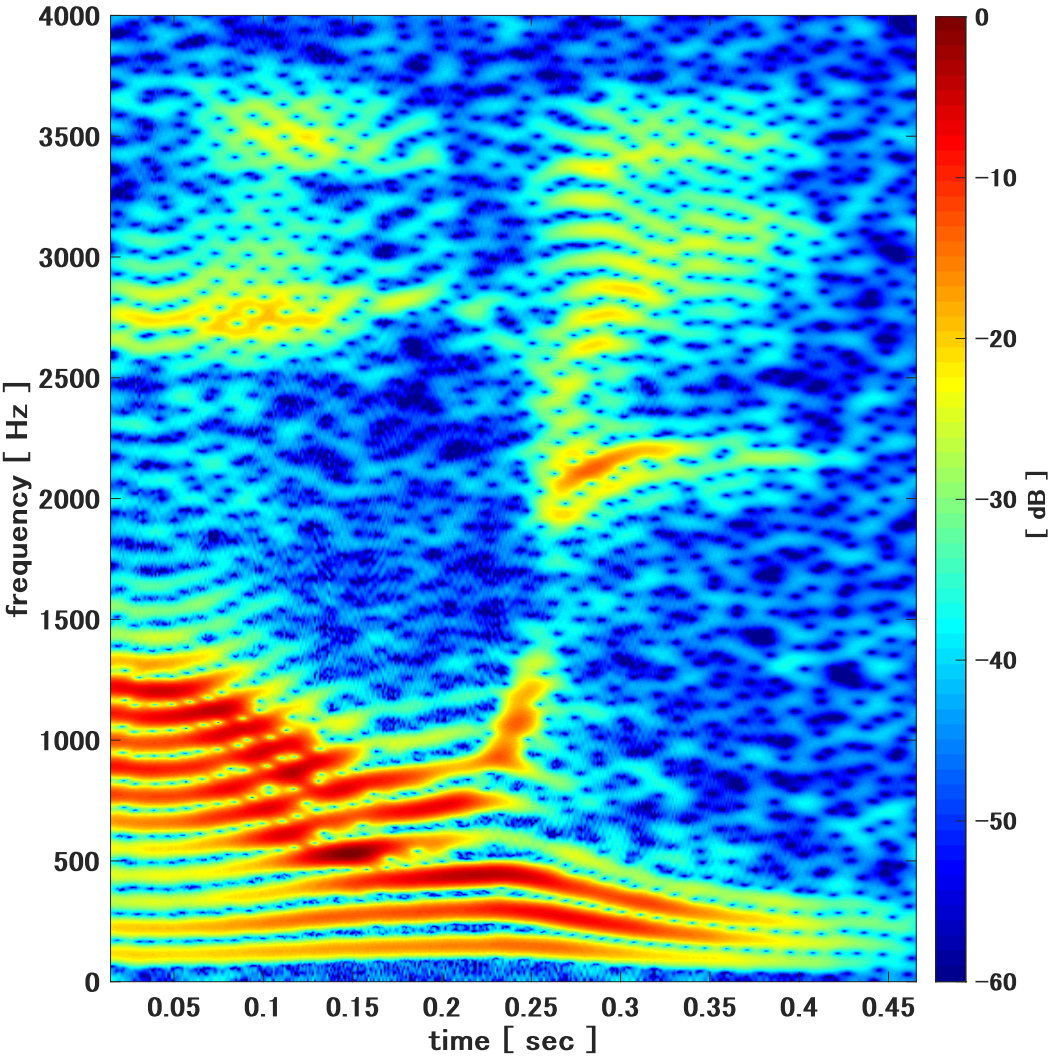

図 3.1: 静音環境におけるサウンドスペクトログラム（男性話者の/aoi/）

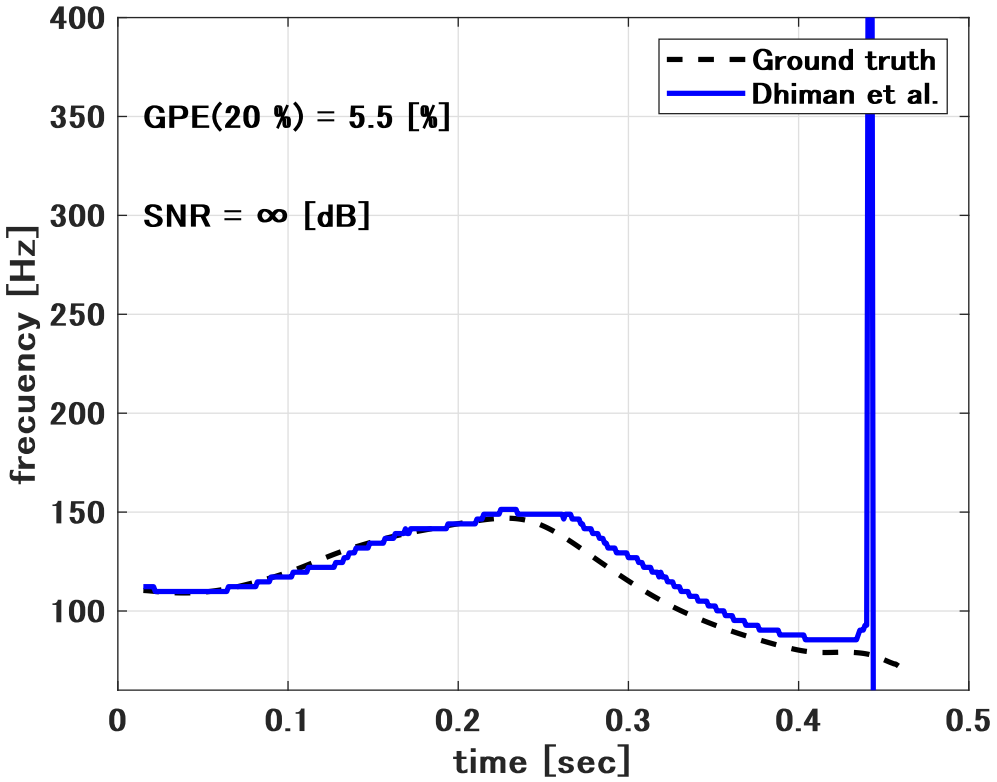

図 3.2: Dhiman らの方法の F0 推定値（SNR ∞ [dB] の場合）

3.1.2　頑健性の確認

図 3.3 は，雑音環境における Dhiman らの方法の F0 推定値である．GPE は 34.4% であり，静音環境と比べて，許容誤差率 20%を上回るフレームの増加分は，28.9%である．この増加は，特に 0.3 sec 付近から増加する誤差による影響が支配的である．

一方図 3.4 は，残響環境における Dhiman らの方法の F0 推定値である．GPE は 39.0%であり，SNR を 0 dB とした雑音環境と比べて，許容誤差率 20%を上回るフレームは，さらに 4.6%増加した．雑音環境で観測された傾向と同様に，時間の経過と共に誤差の増加が認められる．0.3 sec 以降の F0 推定値に認められる直線的な F0 推定値が，GPE の増加要因である．

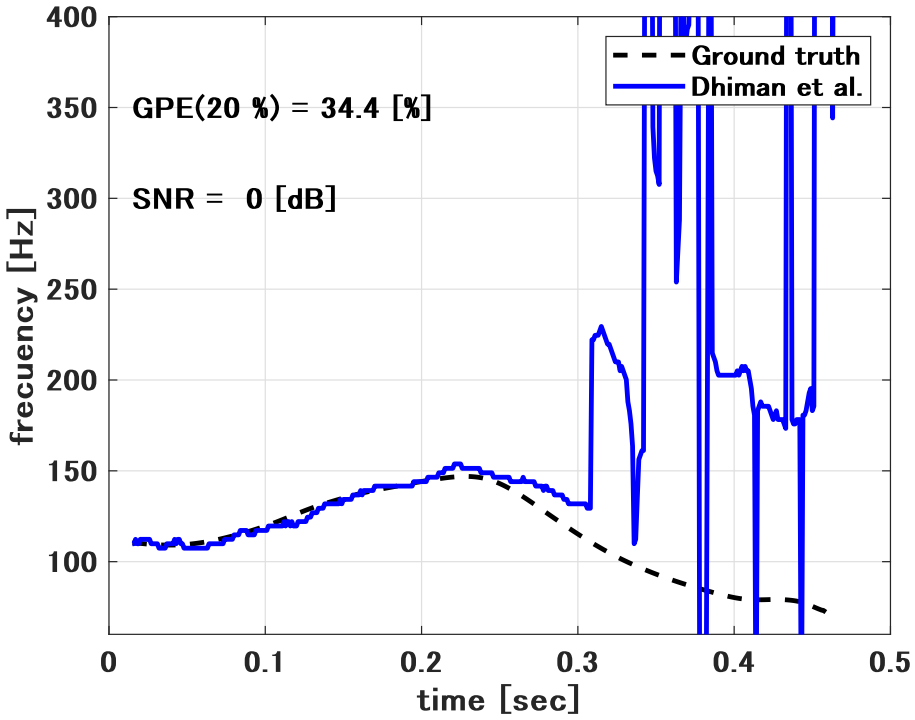

図 3.3: Dhiman らの方法の F0 推定値（SNR 0 [dB] の場合）

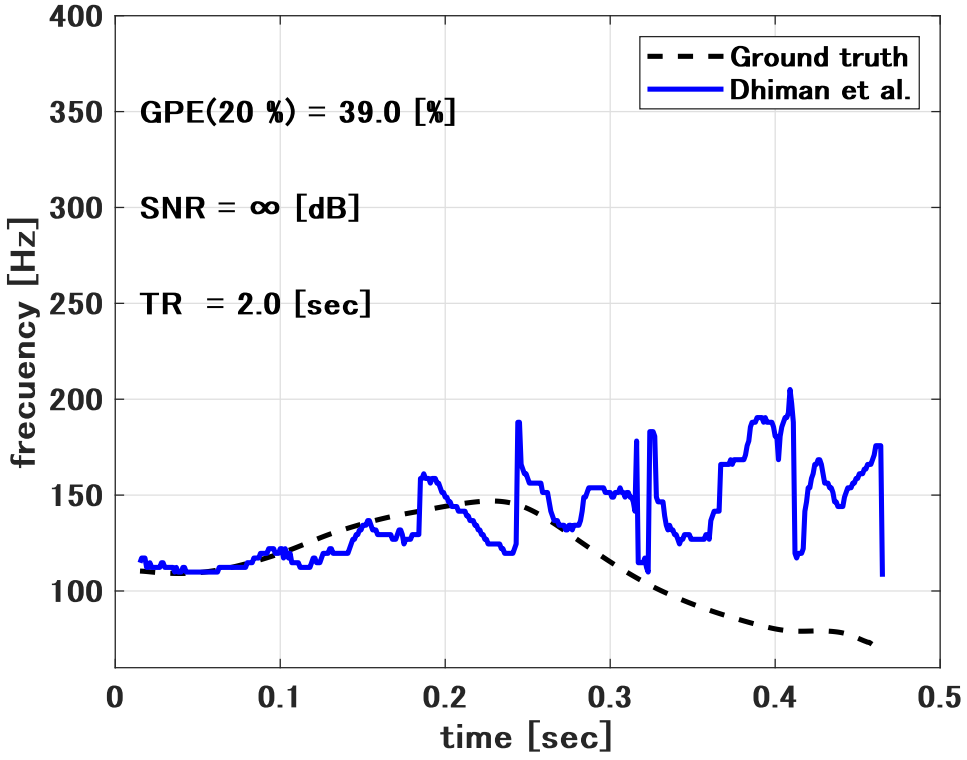

図 3.4: Dhiman らの方法の F0 推定値（SNR ∞ [dB], TR 2.0 [sec] の場合）

3.1.3 頑健性の分析

WIF から F0 推定値を決定する Dhiman らの方法において，調波性を鮮明に抽出することが，GPE の低減に不可欠である．特に，WIF を生成するコヒーレンスマップを乗じる処理は，いわばコヒーレンスマップによる調波性の照合過程であり，WIF とコヒーレンスマップの分析が重要である．図 3.5 は，静音環境における Dhiman らの方法の F0 推定処理の過程であり，上段と下段はそれぞれ WIF とコヒーレンスマップである．同様に，図 3.6 および図 3.7 は，それぞれ雑音環境および残響環境における，Dhiman らの方法の F0 推定処理の過程である．最も鮮明な調波性が確認できる図 3.5 は，WIF とコヒーレンスマップから調波性の所在が容易に識別できる．このため，GPE が最低値を示している．図 3.6 の WIF は，0.3 sec 以降における大部分の調波性が確認できず，加えてコヒーレンスマップが一様である．図 3.7 の WIF の調波性は，直線的である．コヒーレンスマップは，図 3.5 より不鮮明ではあるが，図 3.6 のコヒーレンスマップに比べて，濃淡が表れる．

一方，全ての WIF の一部の領域に，周期性に相当する縦縞模様が観測されている．調波性を利用する Dhiman らの方法において，この縦縞模様に係わる成分は，頑健性を阻害する要因である．BPF による適切な空間フィルタが，Dhiman らの方法の頑健性を支える根幹である．

これらの分析結果から，Dhiman らの方法の頑健性を阻害する主な要因は，次の二点である．

- 空間フィルタによる不要な成分の抽出
- 外乱によるコヒーレンスマップの一様性

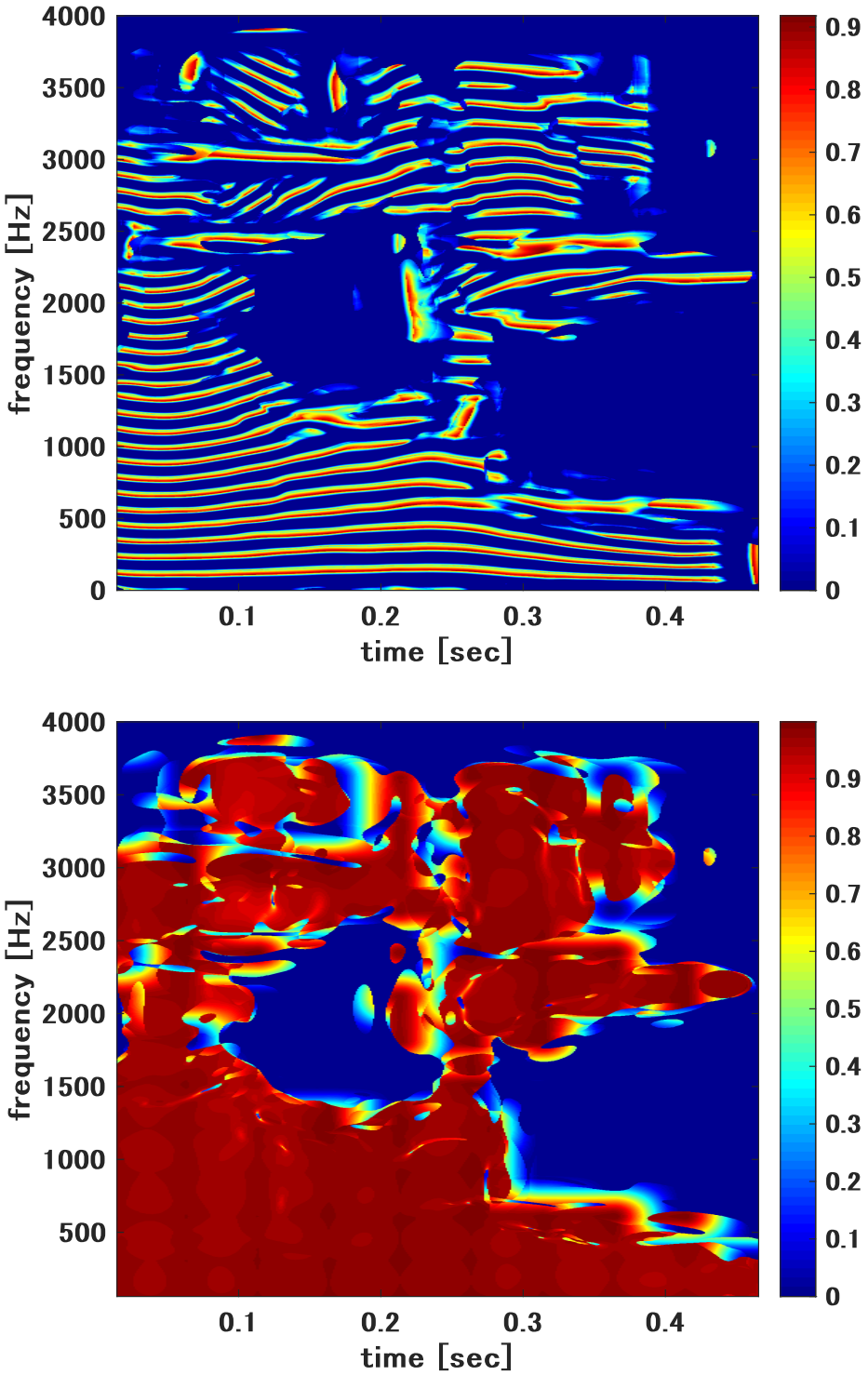

図 3.5: Dhiman らの方法による F0 の推定過程（SNR ∞ [dB] の場合）

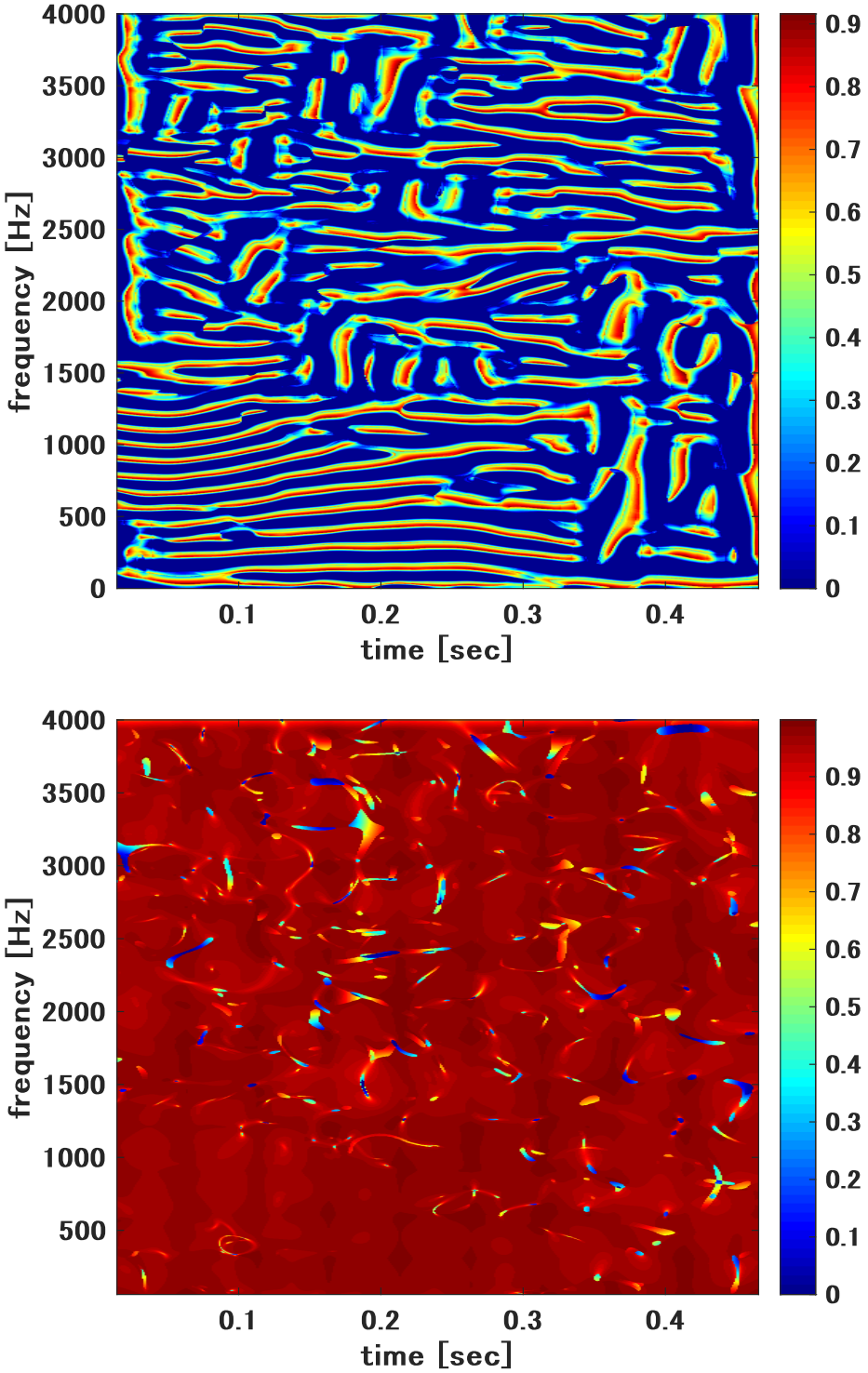

図 3.6: Dhiman らの方法による F0 の推定過程（SNR 0 [dB] の場合）

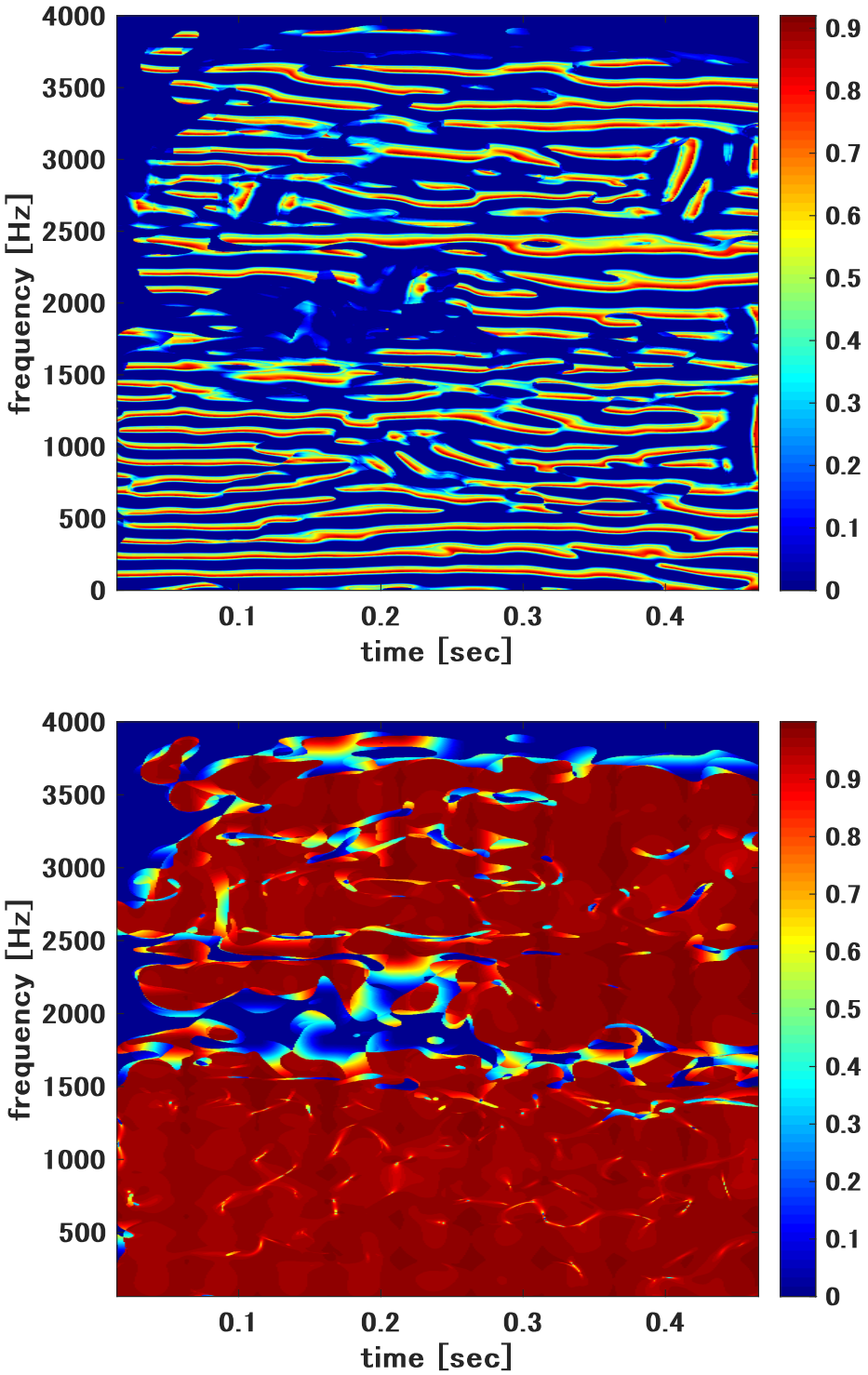

図 3.7: Dhiman らの方法による F0 の推定過程（SNR ∞ [dB], TR 2.0 [sec] の場合）

3.2 着眼点

Dhiman らの方法の頑健性を高めるための着眼点は，次の二点である．

- 基本周波数（F0）と空間周波数の相互作用
- 瞬時振幅（IA）が指し示す領域

3.2.1 基本周波数と空間周波数の相互作用

図 3.8 は，基本周波数を内包する空間周波数領域の幾何学配置である．空間周波数領域の横軸と縦軸は，それぞれ垂直成分（Vertical: V）と水平成分（Horizontal: H）である．Dhiman らの方法の信号モデルは，この空間周波数領域の信号 $S(\boldsymbol{\omega})$ を，式 (3.5）で近似するものである [32]．

$$\begin{aligned} S(\boldsymbol{\omega}) &\approx S_l(\boldsymbol{\omega}) + S_b(\boldsymbol{\omega}) \\ &= \alpha_0 V(\boldsymbol{\omega}) + V(\boldsymbol{\omega}) \cos \Phi(\boldsymbol{\omega}) \end{aligned} \tag{3.5}$$

ただし，$S_l(\boldsymbol{\omega})$ と $S_b(\boldsymbol{\omega})$ は，それぞれ空間周波数領域における低域成分と帯域通過成分である．$\alpha_0, V(\boldsymbol{\omega}), \cos \Phi(\boldsymbol{\omega})$ は，それぞれバイアス，IA，IF である．$\boldsymbol{\omega} = (t, \omega) \in \mathbb{R}^2$ であり，t と ω はそれぞれ時間と周波数である．$S_l(\boldsymbol{\omega})$ は，図 3.8 の上段と下段の原点に表れる成分である．一方 IF を含む $S_b(\boldsymbol{\omega})$ は，図 3.8 の上段と下段において，中心から半径 r_s，角度 θ に表れる成分である．IF を含む $S_b(\boldsymbol{\omega})$ は，$S(\boldsymbol{\omega})$ に BPF を適用して抽出される．

これらの原理から，次の二点が最適な空間フィルタを設計するための核心である．

- $S_l(\boldsymbol{\omega})$ を適切にマスクすること [35]
- その上で，BPF の中心周波数に $S_b(\boldsymbol{\omega})$ を指定すること

$S_l(\boldsymbol{\omega})$ の適切なマスクは，画像の直流成分を除去する処理 [35] に着想する手法であり，例えばヒトの F0 の範囲に基づくマスクの指定法が考えられる．加えて，調波構造の間隔が一定である仮定のもとに，空間周波数領域の幾何学配置と基本周波数 f_0 の間に，式（3.6）の関連性が知られている [37, 41]．

$$r_s \approx \frac{2\pi f_s}{N_{STFT}\ f_0\ \cos\theta} \tag{3.6}$$

ここで，f_s と N_{STFT} は，それぞれ標本化周波数と STFT のフレーム数である．式（3.6）より，BPF の中心周波数に頑健性の高い FreeDAM の F0 推定値が利用できる．この着眼点から，課題を抱える Dhiman らの方法の空間フィルタリングに，頑健性を引き上げる効果が期待できる．

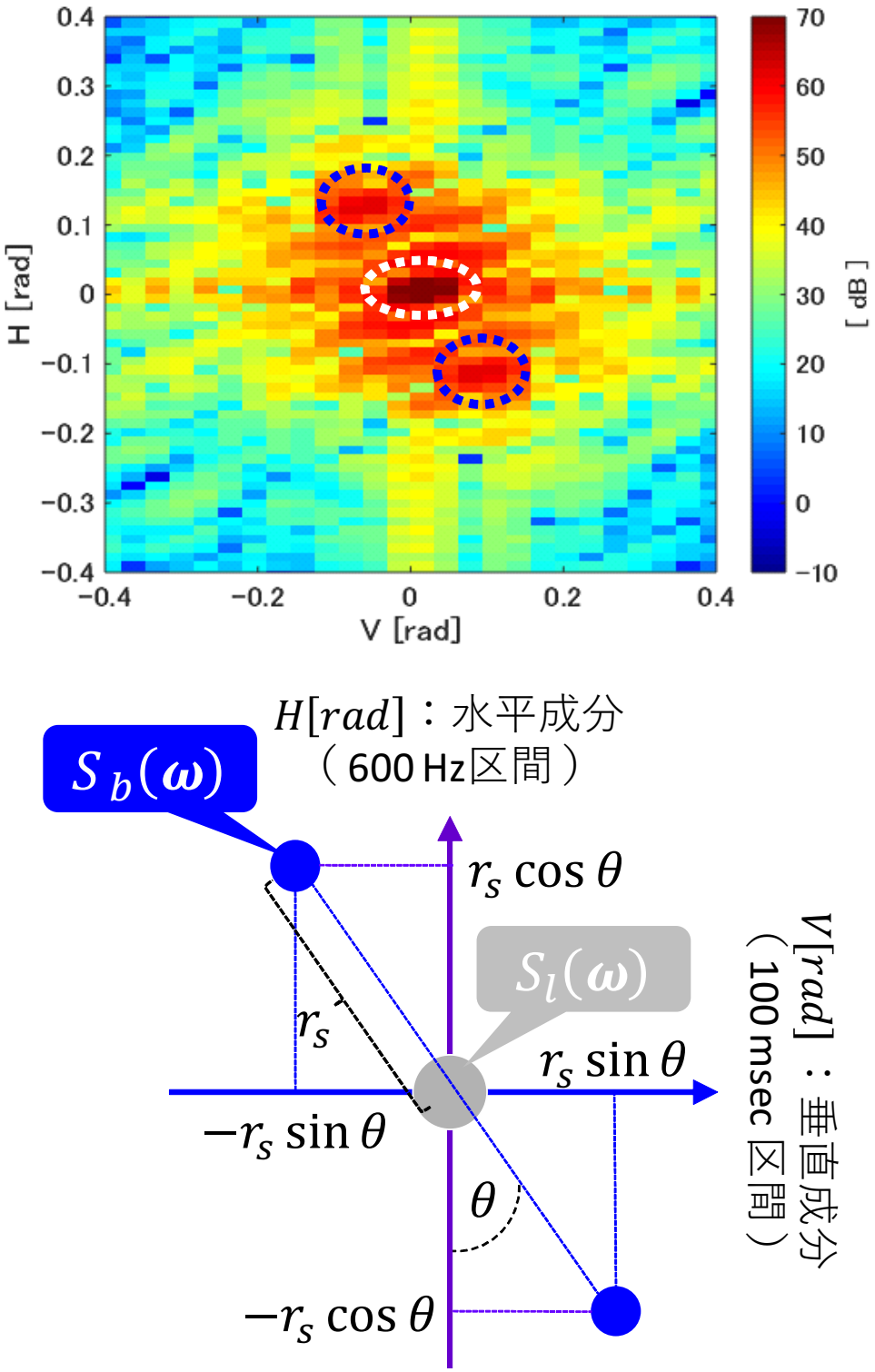

図 3.8: 帯域通過成分の幾何学配置

3.2.2　瞬時振幅が指し示す領域

雑音環境や残響環境の分析で確認された一様なコヒーレンスマップから，雑音や残響の程度による最適な閾値が，コヒーレンスマップの生成に存在することが考えられる．しかし，想定する全ての雑音環境や残響環境を整備し，最適な閾値を発見することは困難である．このため本研究においては，コヒーレンスマップを補強する方針とする．

コヒーレンスマップの補強法として，IA が示すエネルギー集中領域に着目する．図 3.9 は，図 3.1 に示したサウンドスペクトログラムから分離した IA である．IA の支配的な成分は声道フィルタであり，高いエネルギーが密集する領域として，低周波領域が識別できる．コヒーレンスマップの目的は，時間周波数領域から調波性を選別することである．調波性を示す IF に比べて IA の SNR は高く，雑音や残響に頑健である．調波性を識別する手段に IA を用いて，一様なコヒーレンスマップの特徴化を図ることで，推定可能な F0 のダイナミックレンジを広げる効果が期待できる．特に，ICS から IF を分離する際に得られる IA は，Dhiman らの方法に利用されない成分である．

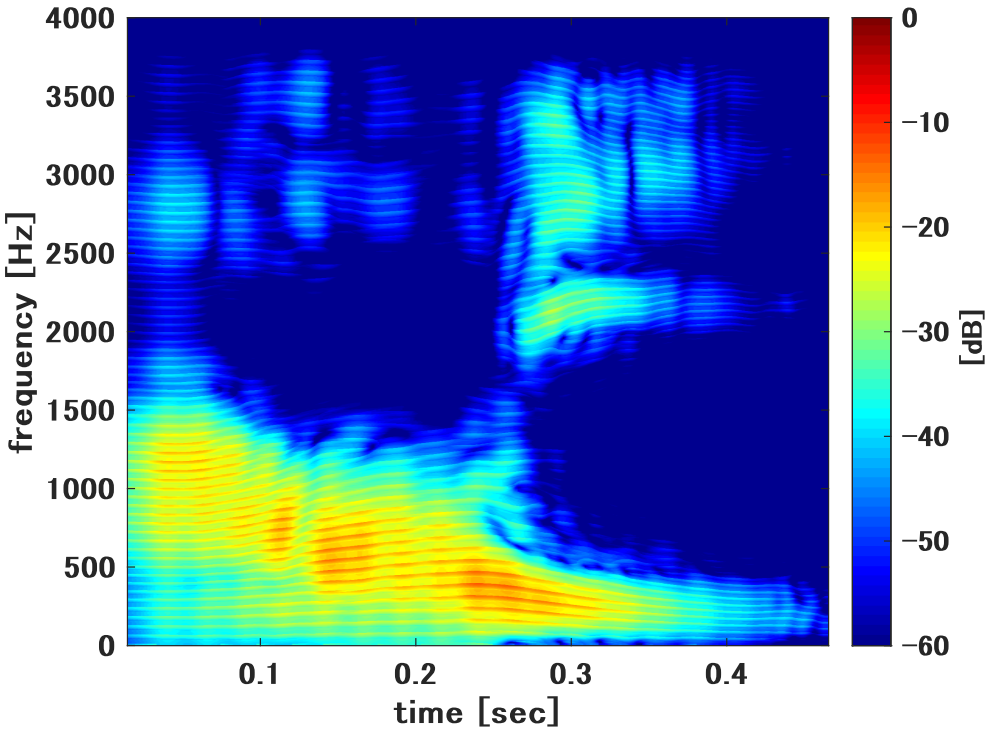

図 3.9: 静音環境における瞬時振幅

3.3 改良法

3.3.1 外乱に頑健性な中心周波数の指定

第一に，Dhiman らの方法の正確性を損なうことなく，頑健性を高めるために，FreeDAM の F0 推定値を用いる改良法を提案する．ただし提案する上で，BPF の帯域幅を決定するために，時間フレーム内で変動する F0 の最大値と最小値が，既知であることを仮定する．

図 3.10 は，改良法の概念図である．上段と下段の図はそれぞれ，外乱の影響を受けた空間周波数領域と，上段の幾何学配置である．改良法の主な特徴は，次の三点である．

- パッチサイズの見直し
  改良法のパッチサイズは，800 Hz × 50 msec である．600 Hz × 100 msec である，Dhiman らの方法のパッチサイズに対して，改良法のパッチサイズは，水平成分を伸ばし垂直成分を縮めている．一般的にヒトの F0 は，概ね 60 Hz から 400 Hz であることが知られており [42]，改良法のパッチサイズにより，2 本以上の調波が密に観測できる．

- マスク領域の明確化
  BPF の中心周波数の取り違えを防止する観点から，空間周波数にマスク領域を適用する．適用するマスク領域は，IA 成分を含む正の水平成分と，負の水平成分である．F0 の推定範囲は男性と女性で異なり，それぞれ概ね 60 Hz - 200 Hz，150 Hz - 400 Hz とされている [42]．この F0 の推定範囲の上限周波数値を基準として，男性と女性でそれぞれ 200 Hz と 400 Hz を上回る水平成分が，マスクの境界条件である．マスク領域を指定することで，BPF の適用範囲を事前に限定する効果が得られる．

- 頑健な中心周波数の指定法
  BPF の中心周波数に，FreeDAM の頑健な F0 推定値を利用する指定法である．外乱の影響を受けた IF 成分は，空間周波数領域に散在し，BPF を適用するべき中心周波数が定まらない．この状況に，頑健性の高い FreeDAM が推定する F0 を適用する．FreeDAM の F0 推定値は，長時間のフレーム長から得られるため，時間平均的な効果により，頑健性に有利に働く．空間周波数領域において，原点からの距離 $\hat{r}_s$ 離れた H 軸上の点を，BPF の中心周波数とする．BPF の帯域幅である d は，時間フレーム内における F0 の最大値と最小値の差から決定すれば良い．式 (3.6) より，FreeDAM の F0 推定値を $\hat{f}_0$，角度 θ を 0 rad として，BPF の中心周波数 $\hat{r}_s$ が算出できる．調波性を表す水平成分を基準に BPF の直径 d を決定するため，角度 θ は 0 rad とする．

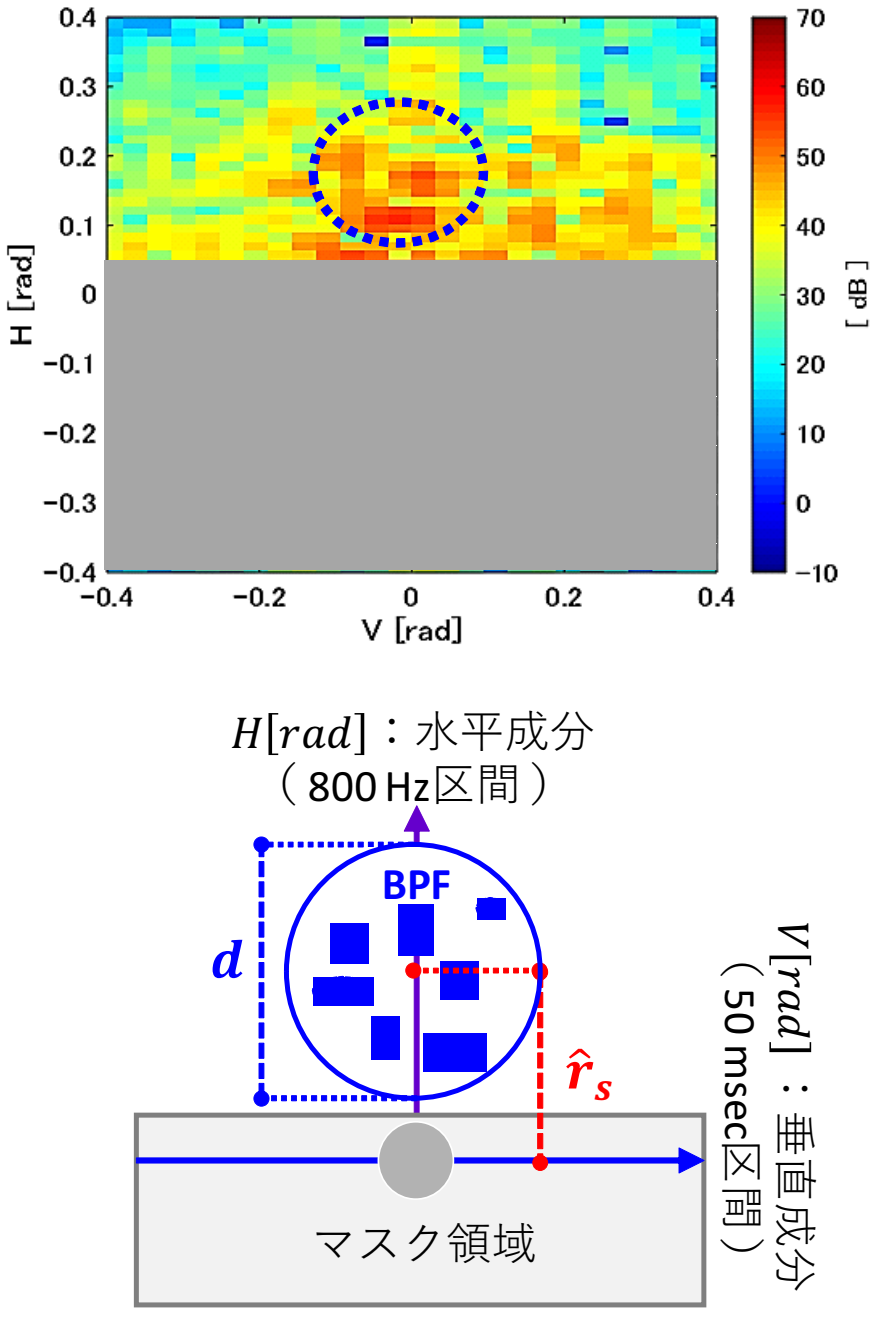

図 3.10: 空間フィルタにおける中心周波数の頑健な指定法

3.3.2　瞬時振幅を利用するマスク

つぎに，コヒーレンスマップを補強する，IA を利用するマスク法（IA マスク）を提案する．この提案は，コヒーレンスマップに IA マスクを乗じることで，新たにコヒーレンスマップの補強効果を狙うものである．図 3.11 は，IA マスクを生成する閾値の決定法を図示するものである．横軸は，0 dB を基準とする IA の減衰量である．左側の縦軸は IA の相対度数であり，右側の縦軸は IA の累積度数である．本研究に採用する IA マスクの閾値は，IA の相対度数に見られる分布形状と，累積度数 5%に相当する IA の振幅から決定する．双峰性の分布形状において，累積度数 5%に相当する IA の閾値から，フォルマント周波数の大局的な位置を識別する．図 3.11 において，累積度数 5%に相当する IA の減衰量は，21.8 dB である．つまり振幅閾値として，-21.8 dB を下回る IA が，F0 推定範囲から除外されるマスク領域である．

IA の支配的な成分はフォルマント周波数であり，時間周波数領域の一部に見られる SNR が高い．このため，フォルマント周波数が重畳する領域から，SNR の高い鮮明な調波性が得られる．

図 3.12 は，-21.8 dB の振幅閾値で生成した IA マスクの例示である．図 3.13 は，この IA マスクで補強した WIF であり，IA マスクが指し示す SNR の高い領域により，IF の鮮明な調波性が確認できる．

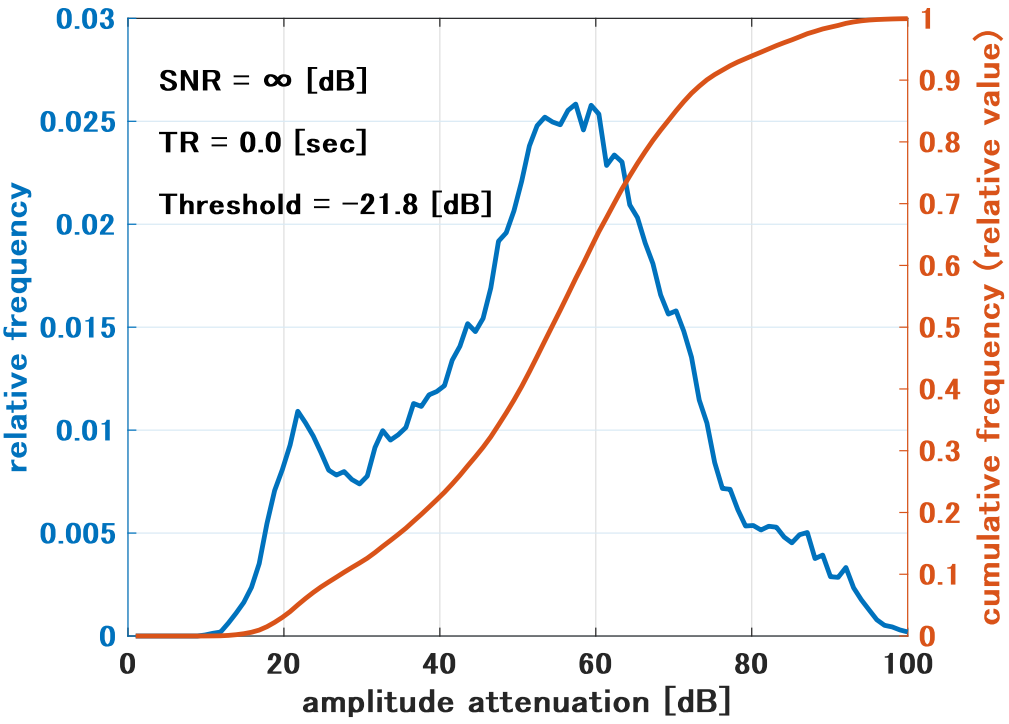

図 3.11: 瞬時振幅を利用する閾値（静音環境における累積度数 5%の場合）

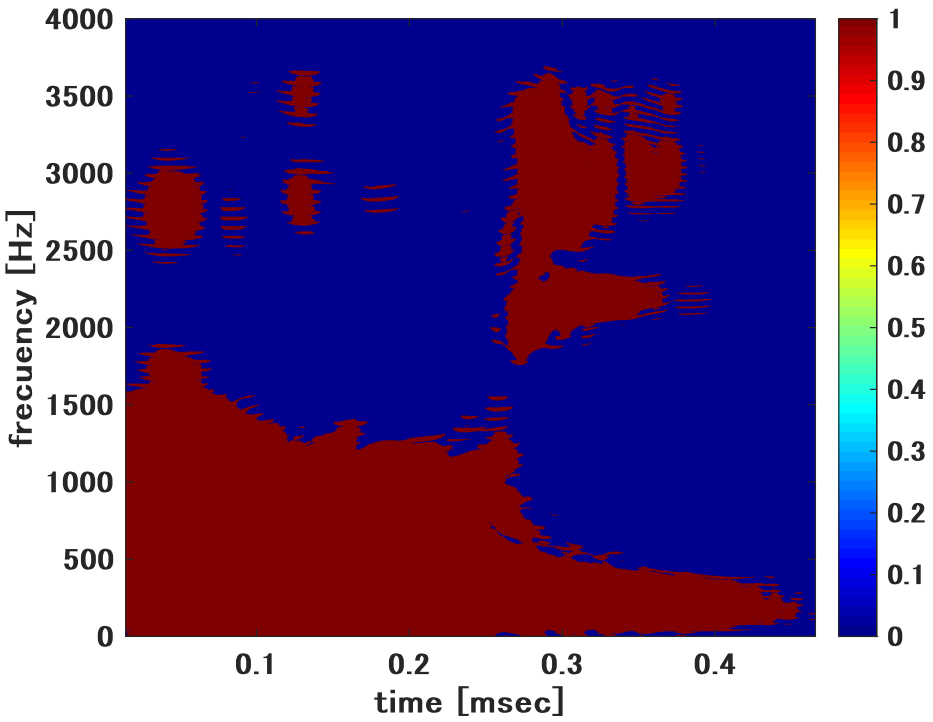

図 3.12: 瞬時振幅を利用するマスク（SNR ∞ [dB] の場合）

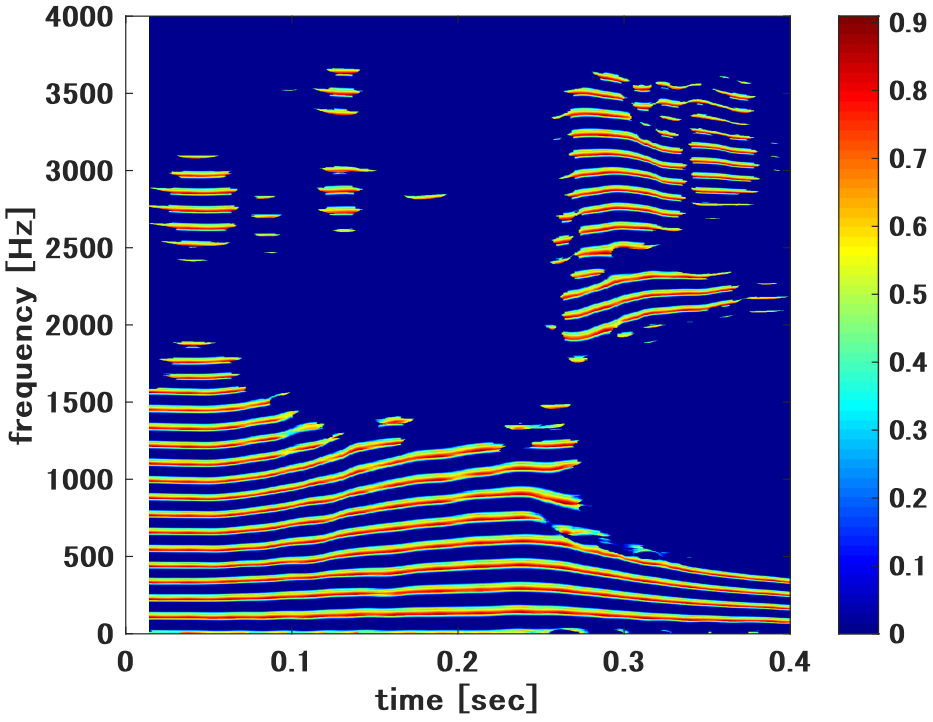

図 3.13: IA マスクで補強した WIF（SNR ∞ [dB] の場合）

第4章　提案法の評価

4.1　評価方法

計算機に提案法を実装し，Dhiman らの方法の頑健性に係わる改善効果を評価する．評価方法は，第 3 章で採用した “評価条件” の諸元に準じるものとする．特筆する提案法の評価条件は，次の六点である．

- F0 の推定範囲
  60 Hz から 400 Hz とする [42].
- FreeDAM の推定結果と F0 の変動幅
  TEMPO の F0 推定値を利用する．FreeDAM の推定結果は，時間フレーム毎の平均値とする．F0 の変動幅は，時間フレーム毎の最大値と最小値とする．
- 雑音環境
  SNR は，20 dB，10 dB，0 dB の三種類とする．
- 残響環境
  残響時間 T_R は，0.5 sec，1.0 sec，2.0 sec の三種類とする．ただし SNR は，観測波形に雑音を加えていないことを示す，∞ とする．
- 評価指標
  GPE の他に，正答率（Correct Ratio: CR)，FPE (Fine Pitch Error) を追加採用する．

  CR は，F0 の推定値 f_0 が許容誤差率 p に収まるフレーム数の割り合いであり，式 (4.1) で算出する，

$$CR_p = \frac{1}{N}\sum_{n=1}^{N} e_p(n) \tag{4.1}$$

$$e_p(n) = \begin{cases} 1, & \left|\dfrac{\hat{f}_0(n) - f_0(n)}{f_0(n)}\right| \leq p \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

FPE は推定精度の指標であり，許容誤差率 p を満たすフレームの誤差率平均値として，式 (4.3) で算出する，

$$FPE_p = \frac{1}{N_{e,p}} \sum_{n=1}^{N} e_p(n) \tag{4.3}$$

$$e_p(n) = \begin{cases} \left| \dfrac{\hat{f}_0(n) - f_0(n)}{f_0(n)} \right|, & \left| \dfrac{\hat{f}_0(n) - f_0(n)}{f_0(n)} \right| \leq p \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

ただし $N_{e,p}$ は，許容誤差率 p を満たすフレーム数である．

- 許容誤差率
  CR，GPE，FPE に採用する許容誤差率 p は，20%と 5%とする．

4.2　評価結果

図 4.1 は，静音環境における F0 の推定結果である．横軸と縦軸は，それぞれ時間と周波数である．さらに図中では，赤色実線，緑色実線，青色実線，黒色破線により，それぞれ提案法，Dhiman らの方法，TEMPO による F0 推定値，F0 の真値を識別する．

提案法と Dhiman らの方法において，0.25 sec 以降の真値との間に一定の差が認められる．また，0.45 sec 付近に見られる局所的な F0 の変動を除き，提案法と Dhiman らの方法の評価結果に，いずれも大きな違いは認められない．

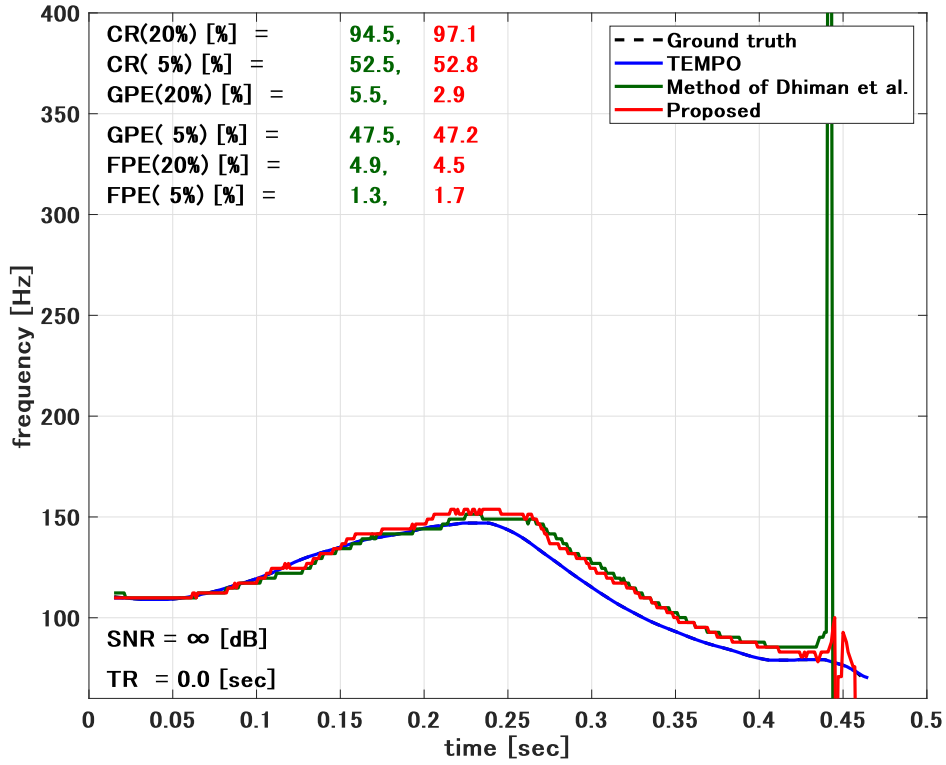

図 4.1: 提案法の F0 推定値（SNR ∞ [dB] の場合）

4.2.1　雑音環境の評価結果

図 4.2，図 4.3，図 4.4 は，それぞれ SNR 20 dB，SNR 10 dB，SNR 0 dB における F0 推定値である．

図 4.2 の評価結果は，図 4.1 の静音環境と同等の値を示しており，SNR 20 dB による影響は認められない．

図 4.3 の TEMPO の評価結果に，10%程度の F0 推定誤差の増加が確認できる．一方，提案法と Dhiman らの方法に認められる F0 推定誤差の増加は，数%程度である．

図 4.4 の許容誤差率 20%の評価結果において，TEMPO と Dhiman らの方法の F0 推定誤差が，数 10%以上増加している．一方，提案法の F0 推定誤差の増加は，数%程度である．

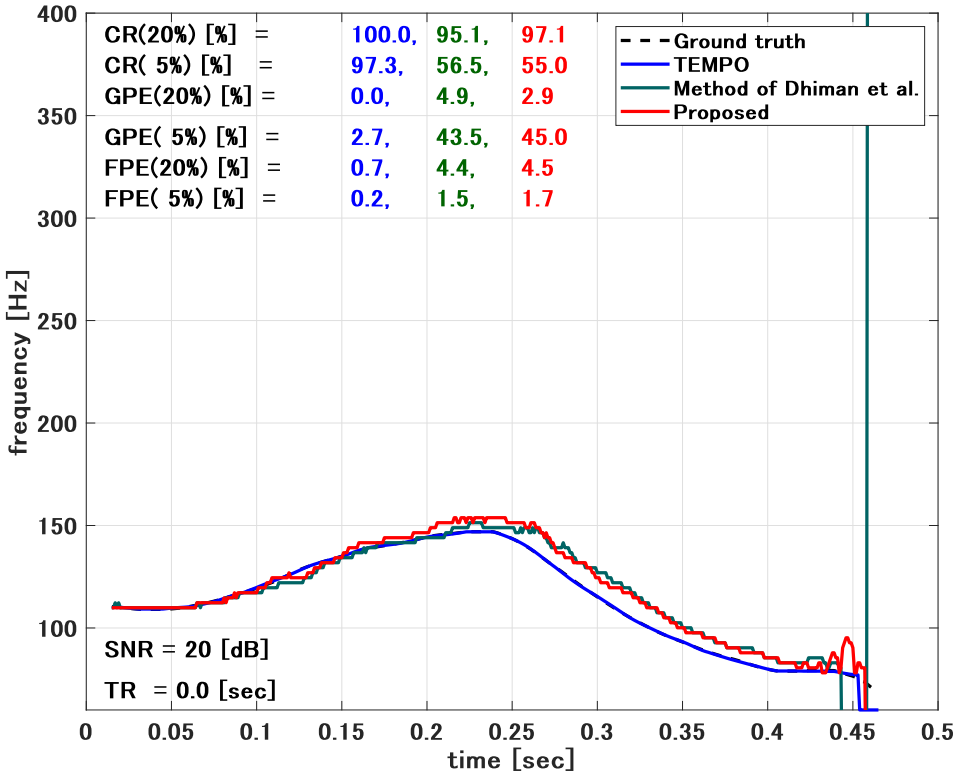

図 4.2: 提案法の F0 推定値（SNR 20 [dB] の場合）

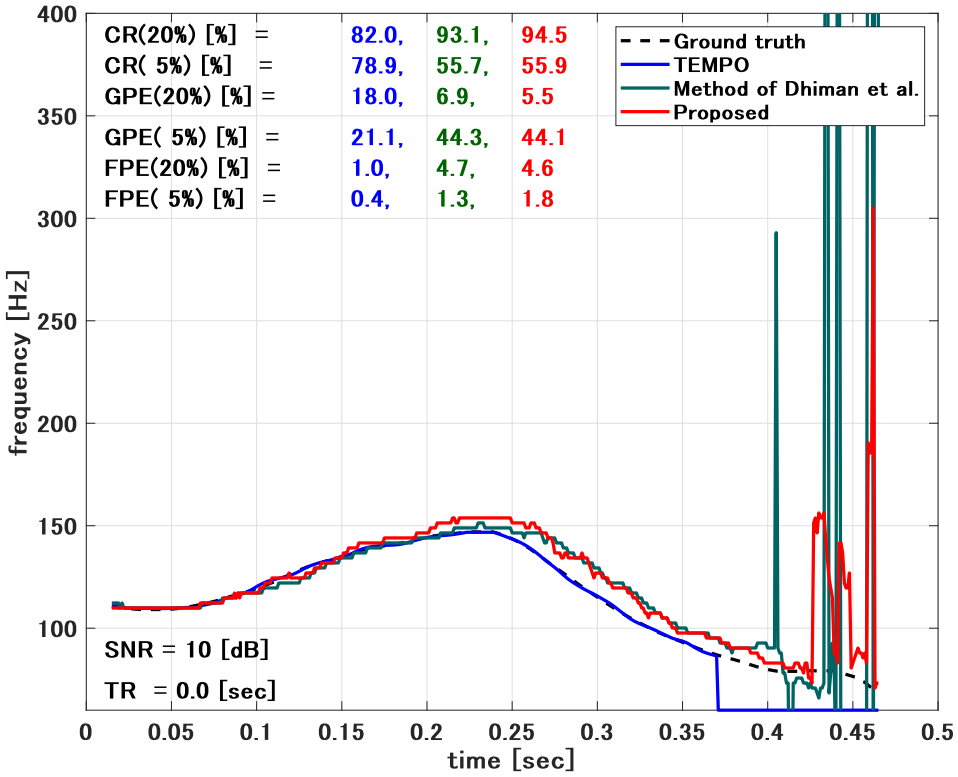

図 4.3: 提案法の F0 推定値（SNR 10 [dB] の場合）

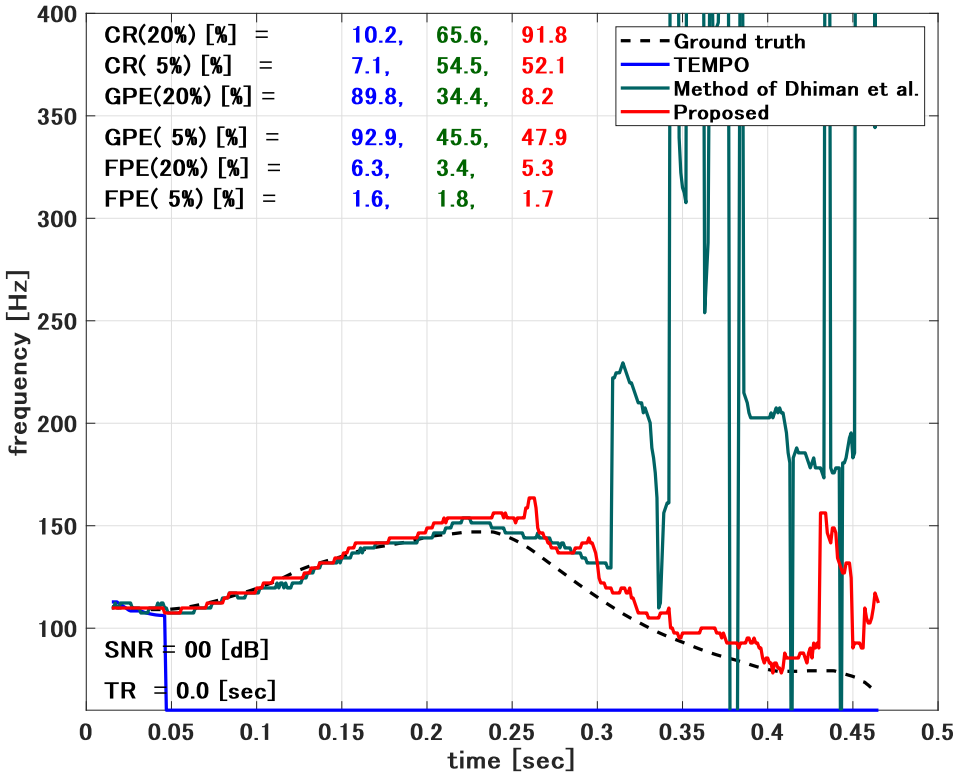

図 4.4: 提案法の F0 推定値（SNR 0 [dB] の場合）

4.2.2　残響環境の評価結果

図 4.5，図 4.6，図 4.7 は，残響環境における F0 推定値である．それぞれの残響時間 T_R は，0.5 sec，1.0 sec，2.0 sec である．

図 4.5 と図 4.1 の静音環境における評価結果を比較すると，全ての F0 推定法において 10%以上の F0 推定誤差の増加が認められる．

図 4.6 の 0.3 sec 以降は，TEMPO の値域を超える領域である．この 0.3 sec 以降に観測される F0 推定誤差は，Dhiman らの方法においても同様の傾向を示しており，大幅な F0 推定誤差の変動が認められる．一方提案法の F0 推定誤差は，図 4.1 の静音環境に比べて 10%以上の増加が認められるが，真値に追従する傾向が認められる．

図 4.6 と同様に，図 4.7 の 0.25 sec 以降の F0 推定誤差において，真値に追従する大局的な F0 の軌跡が，提案法に認められる．

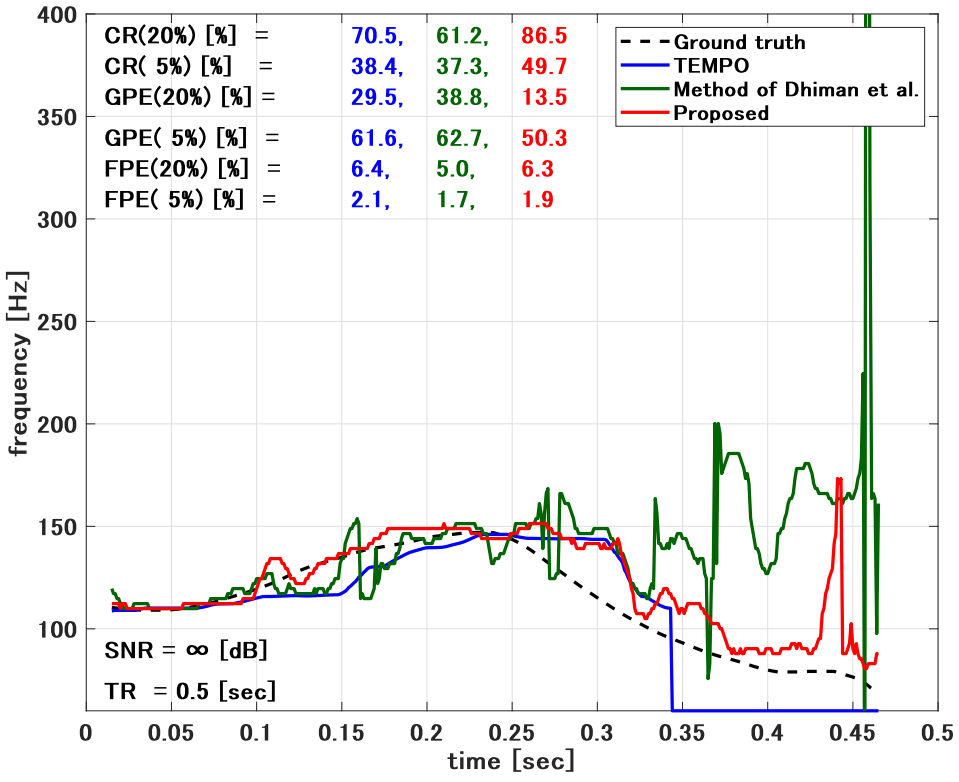

図 4.5: 提案法の F0 推定値（SNR ∞ [dB], TR 0.5 [sec] の場合）

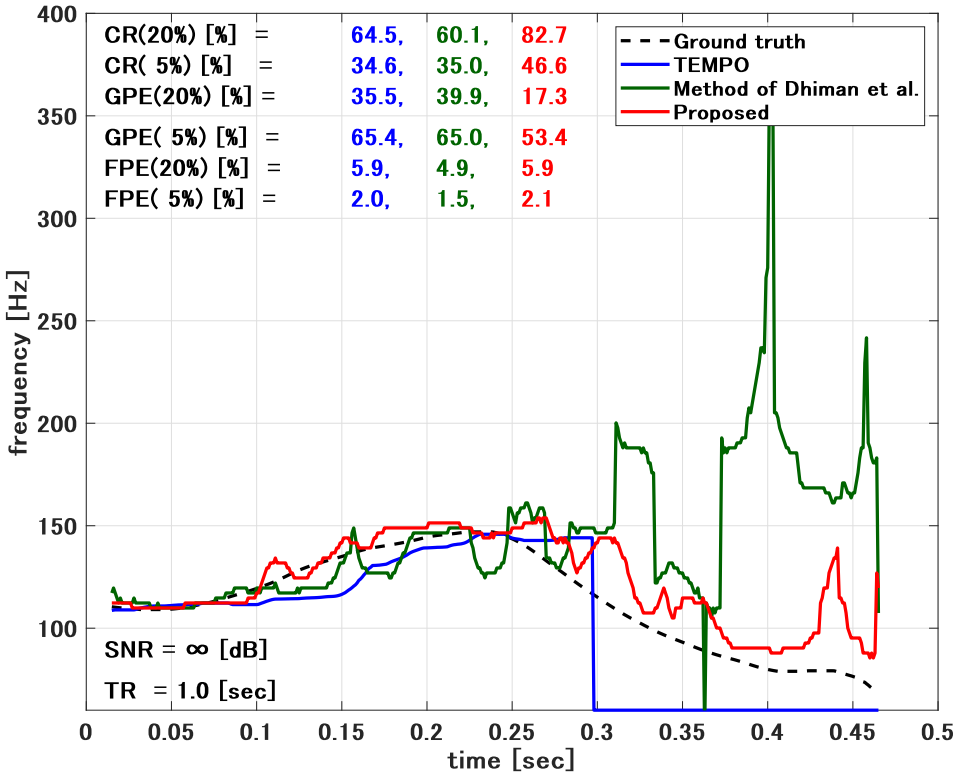

図 4.6: 提案法の F0 推定値（SNR ∞ [dB], TR 1.0 [sec] の場合）

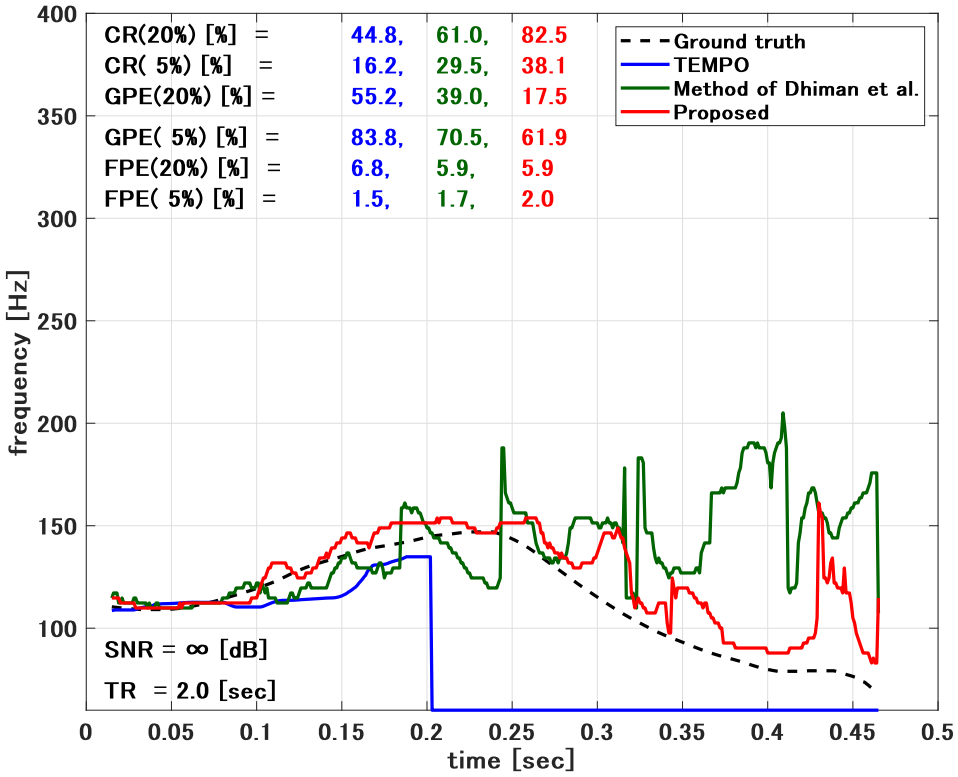

図 4.7: 提案法の F0 推定値（SNR ∞ [dB], TR 2.0 [sec] の場合）

4.3 考察

- 静音環境
  許容誤差率 20%において，Dhiman らの方法と提案法の CR は 90%以上，GPE と FPE は 10%以下であり良好な結果が認められる．さらに全ての評価指標において，大きな差が認められないことから，正確性は TEMPO に比べて遜色ない水準である．ただし，Dhiman らの方法と提案法には，0.25 sec 以降の真値との差が見られる．この差の理由の一つは，0.25 sec 以降の低次の調波の数が低下することに起因するものである．F0 推定に利用する調波の数を増やすことで，許容誤差率 5%を採用する評価指標が改善される．なお，0.45 sec 付近に見られる F0 の局所的な変動は，音声の終端に見られる SNR の低下によるものである．これらの事象から，提案法の正確性は，Dhiman らの方法と同等であることが確認された．

- 雑音環境
  SNR が 0 dB の場合において，提案法の優位性が認められる結果となった．ただし，許容誤差率 5%の F0 推定誤差は，Dhiman らの方法と提案法に違いは認められない．これらの事象から，SNR が低い場合において，F0 推定誤差の改善効果が確認された．

- 残響環境
  Dhiman らの方法と提案法の F0 推定誤差において，FPE に大差は認められない．一方，FPE を除く全ての評価指標において，Dhiman らの方法に比べて，提案法の F0 推定誤差に，概ね 10%以上の改善効果が認められた．これらの事象から，残響環境において，改善効果が確認された．

静音環境，雑音環境，残響環境において，Dhiman らの方法の正確性を損なうことなく頑健性を向上する効果が，提案した改良法の評価結果から認められた．

第5章　結論

5.1　本研究で明らかにしたこと

本研究の目的は，近年提案されたF0推定法の中から，正確定と頑健性を満たす性質を掘り下げ，互いの弱点を合理的に克服する実用的なF0推定法の骨格を形成することであった．この目的の中で，本研究が明らかにした成果は，次の四点である．

- Dhimanらの方法の雑音耐性と残響耐性の脆弱性
- Dhimanらの方法の頑健性を損なう主な要因の特定
- 瞬時振幅（IA）に着目した調波性の識別能力の改善効果
- FreeDAMの性質に着目したDhimanらの方法の頑健性の増強法

本研究の主な成果は，Dhimanらの方法の頑健性を減退させる要因を特定し，改良法により提案内容の妥当性を，実装評価から明らかにしたことである．正確性の高いDhimanらの方法の頑健性を損なう主な要因の一つは，雑音や残響によりBPFの中心周波数が定まらないことである．さらにコヒーレンスマップに見られる調波性の減退が，外乱の程度による付随現象として増長する事象も挙げられる．本研究では，これらの性質に着目し，FreeDAMのF0推定値とDhimanらの方法では重要視されていなかったIAを，BPFの中心周波数の指定法と，コヒーレンスマップによる調波性の識別法に統合するという，合理的な改良法を提案した．FreeDAMが推定するF0は，いわば分析フレーム毎の代表値であり，声道フィルタの成分が支配的となるIAは，フォルマント周波数の影響を受けて，大部分のエネルギーが集中する領域を識別する．これらの性質を積極的に応用することで，F0推定法の頑健性が合理的に高まる．

5.2 残された課題

本研究の残された課題は，次の三点である．

- 大規模データベースを利用した改良法の評価と洗練化
- 洗練化で得られた知見に基づく FreeDAM の拡張
- 拡張した FreeDAM と Dhiman らの方法の有機的な結合による提案法の確立

F0 推定法の実用化に向けた取り組みとして，大規模データベースを利用し，改良法の洗練化が必要である．本研究で明らかにした成果は，有声区間の単独話者の母音による評価結果から認められた範囲に限定されるためである．特に，話者の性差，年齢，地域や言語等による個人性から，F0 の時間的軌跡が異なる．また，F0 は有声音の母音にのみ観測される特徴量であり，音声から無声音や子音を正確に識別する手法が，実用的な観点から不可欠である．大規模なデータベースを評価に利用することは，F0 推定に係わるより多くの知見の獲得につながり，その結果から提案法の洗練化が可能になると考えられる．さらに，頑健性が高いことが知られている FeeDAM は，一方で正確性に課題を抱えており，提案法の洗練化による波及効果として，FreeDAM の課題解決に繋がる新しい知見が獲得できる可能性も見込める．このように FreeDAM と Dhiman らの方法の相補的な観点もあり，継続的に検討を進めることが重要である．

より自由な発想のもとに，FreeDAM と Dhiman らの方法の有機的な結合に向けて，絶え間なく検討していくことが，正確性と頑健性を両立させるプロセスであると考える．

参考文献

[1] 鈴木久喜，“ピッチ抽出の今昔，” 日本音響学会誌，vol. 56, no. 2, pp. 121–128, 2000.

[2] 大串健吾，音のピッチ知覚，音響サイエンスシリーズ 15，コロナ社，2016.

[3] 筧一彦， 辰巳格， 皆川泰代， 持田岳美， 渡辺眞澄，聞くと話すの脳科学，音響サイエンスシリーズ 17，廣谷定男（編），コロナ社，2017.

[4] SoftBank, “Pepper,” https://www.softbank.jp/robot/consumer/products/, (2019-07-26 閲覧)

[5] 変なホテル，“ヒト型ロボット，” https://www.hennnahotel.com/, (2019-07-26 閲覧)

[6] 香田徹，日比野浩，任書晃，倉智嘉久，入野俊夫，鵜木祐史，鈴木陽一，牧勝弘，津崎実，聴覚モデル，音響サイエンスシリーズ 3，森周司，香田徹（編），コロナ社，2011.

[7] 大串健吾，“音のピッチ知覚について，” 日本音響学会誌，vol. 73, no. 12, pp. 758–764, 2017.

[8] Gunnar Fant, “The source filter concept in voice production,” STL-QPSR. vol. 22, no. 1, pp. 21–37, 1981.

[9] 正木信夫，元木邦俊，松崎博季，北村達也，音声生成の計算モデルと可視化，音響テクノロジーシリーズ 14，鏑木時彦（編），コロナ社，2010.

[10] Bishnu S. Atal and Suzanne L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” J. Acoust. Soc. Am., vol. 50, no. 2, pp. 637-655, 1971.

[11] Bernard Gold and Larence R. Rabiner, “Parallel processing techniques for estiming pitch periods of speech in the time domain,” J. Acoust. Soc. Am., vol. 46, no. 2, pp. 442-448, Aug. 1969.

[12] 斉藤洋一，ディジタル無線通信の変復調，電子情報通信学会（編），コロナ社，東京，1996.

[13] 安部素嗣，安藤繁，“共有 FM-AM の時間周波数統合に基づく聴覚情景解析 (I) :Lagrange 微分特徴量とその周波数軸統合，” 電子情報通信学会論文誌，Vol. 83, no. 2, pp. 458-467, 2000.

[14] 鵜木祐史，石本祐一，赤木正人，“残響音声からの基本周波数推定に関する検討,” JAIST Research Report，IS-RR-2005-007, March 2005.

[15] Alain de Cheveigne and Hideki Kawahara, “Yin, a fundamental frequency estimator for speech and music,” J. Acoust. Soc. Am., vol. 111, no. 4, pp.1917–1930, 2002.

[16] Arturo Camacho and John G. Harris, “A sawthooth waveform inspired pitch estimator for speech and music,” J. Acoust. Soc. Am., vol. 124, no. 3, pp. 1638–1652, 2008.

[17] 森勢将雅, “2-2 基本周波数推定（歌声研究に関する視点から）,” 電子情報通信学会知識ベース，2 群-9 編-2 章, 2010.

[18] Hideki Kawahara and Haruhiro Katayose, Alain de Cheveigne, and Roy D. Pattrson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity,” Proc. Eurospeech’99, vol. 6, pp. 2781–2784, 1999.

[19] Yuichi Ishimoto, Masashi Unoki, and Masato Akagi, “A Fundamental Frequency Estimation Method for Noisy Speech Based on Instantaneous Amplitude and Frequency,” Proc. EuroSpeech2001, pp. 2439–2442, 2001.

[20] Arthur P. Dempster, “Upper and Lower Probabilities Induced by a Multivalued Mapping,” Annals of Mathematical Statistics, vol. 38, no. 2, pp. 325–339, 1967.

[21] 浅沼克紀，大西正輝，小島篤博，福永邦雄，“色情報と領域追跡情報を用いた人物の顔と手の領域の抽出,” 電気学会論文誌 C, vol. 119-C, no. 11, pp. 1351–1358, 1999.

[22] 三輪賢一郎，鵜木祐史，“振幅変調のピッチ知覚に基づいた調波複合音の基本周波数推定法,” 電子情報通信学会論文誌（A），vol. J98-A, no. 12, pp. 668–679, 2015.

[23] Kenichiro Miwa and Masashi Unoki, “Robust method for estimating F0 of complex tone based on pitch perception of amplitude modulated signal, ” Proc. Interspeech2017. pp.2311–2315, 2017.

[24] 三輪賢一郎，“振幅変調特性に着目した雑音残響に頑健な基本周波数推定法，” 北陸先端科学技術大学院大学　情報科学研究科　博士論文，Nov. 2018.

[25] Schouten J. Frederik, “The Residue, a new Component in Subjective Sound Analysis,” Proc. Koninkl. Ned. Akad. Wetenschap. vol. 43, pp. 356-365, 1940.

[26] 鵜木祐史，山崎悠，赤木正人，“雑音残響環境下における MTF ベース・パワーエンベロープ回復処理の検討,” 日本音響学会春季講演論文集, pp. 853–856, 2010.

[27] 鵜木祐史，“変調伝達関数に基づく音声信号処理（1）パワーエンベロープ逆フィルタ処理の原理とその応用について,” 信号処理学会誌，vol. 12, no. 5, pp. 339–348, 2008.

[28] Jitendra Kumar Dhiman, Nagaraj Adiga, and Chandra Seelamantula, “A Spectro-Temporal Demodulation Technique for Pitch Estimation,” Proc. Interspeech2017. pp. 2306–2310, 2017.

[29] Candra Sekhar Seelamantula, Nicolas Pavillon, Christian Depeursinge, and Michael Unser, “Local demodulation of holograms using the Riesz transform with application to microscopy,” J. Opt. Soc. Am. A., vol. 29, no. 10, pp. 2118–2129, Oct. 2012.

[30] 尾知博，シミュレーションで学ぶディジタル信号処理，CQ 出版社，2004.

[31] Anne Sedlazeck, “Local feature detectionby higher order Riesz transforms on images,” Thesis, University of Kiel, 2008.

[32] Haricharan Aragonda and Candra Sekhar Seelamantula, “Demodulation of narrowband speech spectrograms using the Riesz transform,” IEEE Trans. Audio, Speech, Lang. Process., vol. 23, no. 11, pp. 1824–1834, Nov. 2015.

[33] Bernd Jähne, “Digital Image Processing,” Springer-Verlag, Berlin, 2005.

[34] Karthika Vijayan, Jitendra Kumar Dhiman, and Chandra Sekhar Seelamantula, “Time-Frequency Coherence for Periodic-Aperiodic Decomposition of Speech Signals,” Proc. Interspeech2017. pp. 329–333, 2017.

[35] 村松正吾，MATLAB による画像&映像信号処理，CQ 出版社，2007.

[36] 多次元信号とシステム，ディジタル信号処理ハンドブック，電子情報通信学会（編），オーム社，東京，1993.

[37] Tianyu. T. Wang and Thomas. F. Quatieri, "Two-dimensional speech-signal modeling," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 6, pp. 1843 - 1856, Aug. 2012.

[38] Larence R. Rabiner, "On the use of autocorrelation analysis for pitch detection," IEEE Trans. on Acoustics, Speech, Signal Process., vol. ASSP-25, no. 1, pp. 24-33, Feb. 1977.

[39] Masashi Unoki, Hosorogiya Toshihiro, "Estimation of fundamental frequency of reverberant speech by utilizing complex cepstrum analysis," Journal of Signal Processing, vol. 12, no. 1, pp. 31-44, Jan. 2008.

[40] Lawrence R. Rabiner, Michael J. Cheng, Aaron E. Rosenberg, and Carol A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. Acoustic, Speech, Signal, Process., vol. ASSP-24, no. 5, pp. 399–418, 1976.

[41] Tianyu T. Wang and Thomas F. Quatieri, "Towards co-channel speaker separation by 2-D demodulation of spectrograms," in Proc. IEEE Workshop on Applications of Signal Process to Audio and Acoustics, Oct 2009, pp. 65 - 68, 2009.

[42] 赤羽誠，石川畑，大河内正明，粕谷英樹，桑原尚夫，田中和世，新田恒雄，矢頭隆，渡辺隆夫，音声工学，板橋秀一（編），森北出版社，2005.

謝辞

本研究の遂行にあたり，熱心なご指導を賜りました，北陸先端科学技術大学院大学の主指導教員である鵜木祐史教授に，深甚な感謝の意を申し上げます．

本研究を進めるにあたり，貴重なご助言を賜りました，北陸先端科学技術大学院大学の赤木正人教授，党建武教授に，厚く謝意を申し上げます．

本研究を通じてお世話になった，北陸先端科学技術大学院大学の鵜木・赤木研究室の皆様，党研究室の皆様，職員の皆様に心よりお礼申し上げます．

修学においてご支援いただいた，株式会社光電製作所の皆様に，心より感謝いたします．

最後に，温かく見守ってくれた両親と，修業と就業の両立を支えてくれた妻に，低頭してここに謝意を表します．

付 録A

A.1　入力波形

図 A.1 は，ATR ディジタル音声データベースに収録された，男性話者の実音声（/aoi/）であり，図 3.1 に示すサウンドスペクトログラムの生成元である．

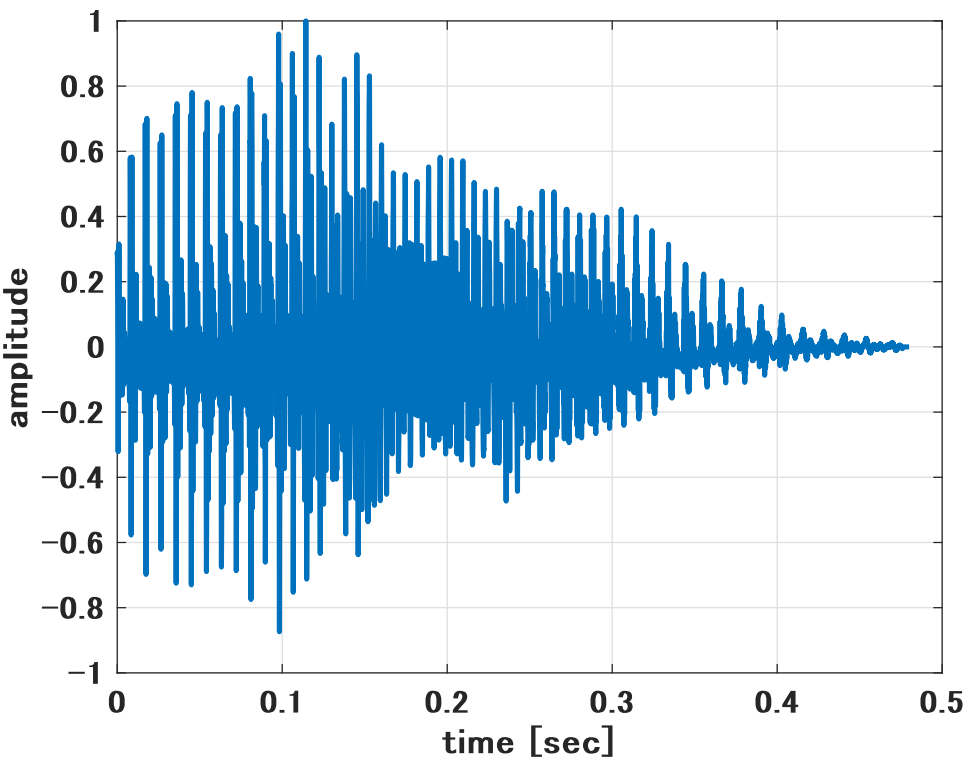

図 A.1: 実音声（男性）の時間軸波形

A.2　評価に利用したF0の基準

図 A.2 は，TEMPO で求めた F0 の基準値である．

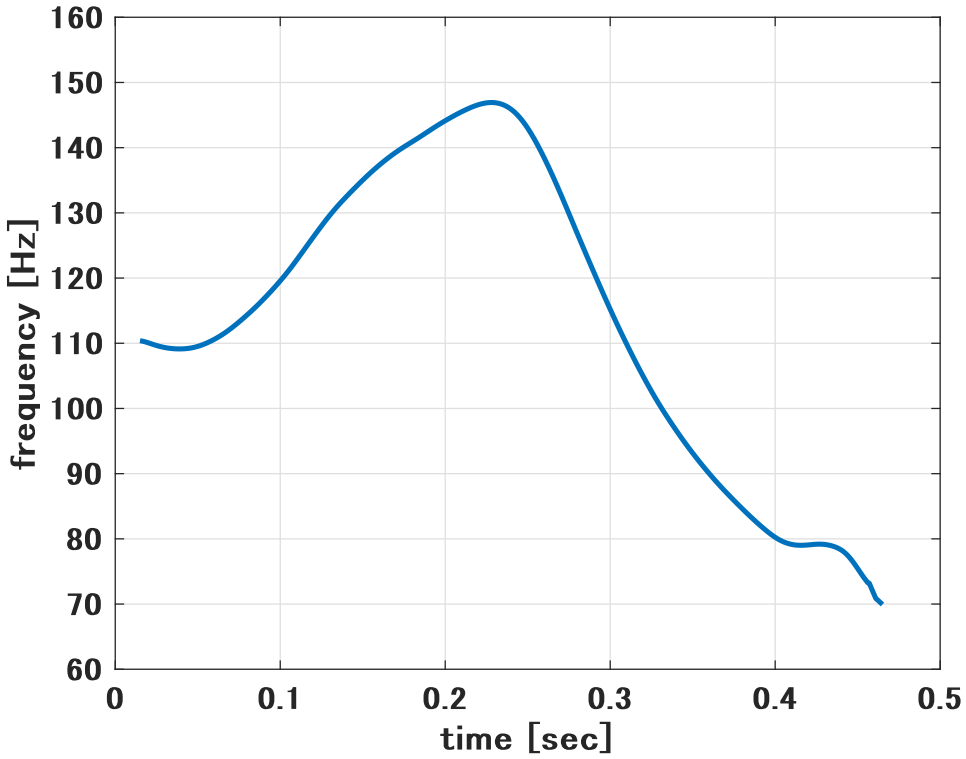

図 A.2: 実音声（男性）の基本周波数（F0）の軌跡

A.3　Riesz カーネル

図 A.3 は，CRT に用いる Riesz カーネルの位相特性である．軸上の V と H は，それぞれ垂直成分と水平成分である．

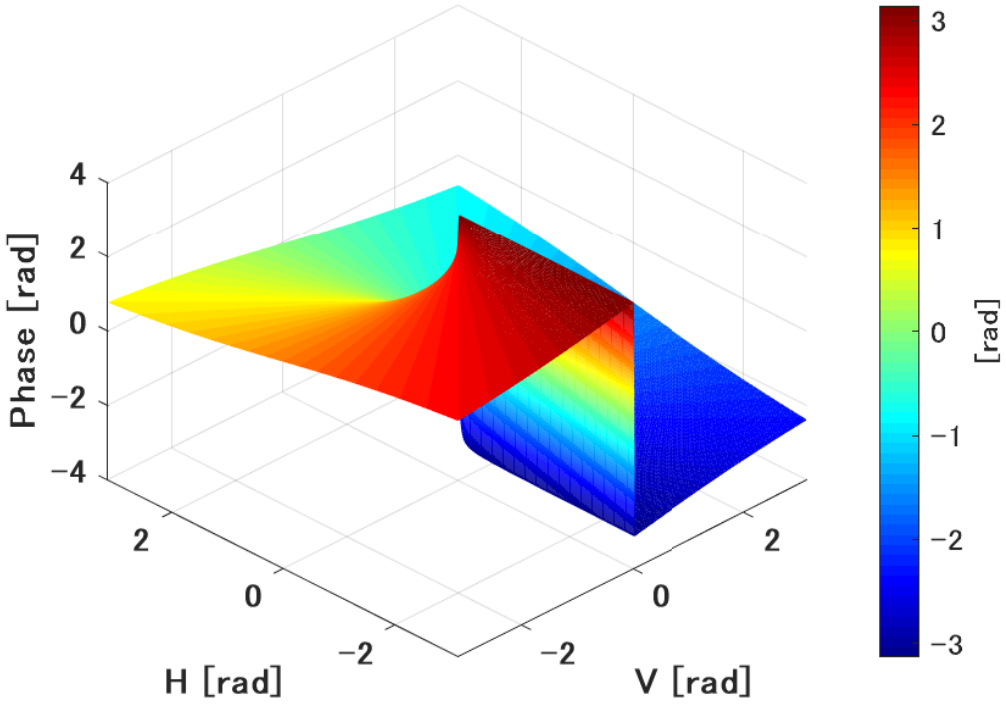

図 A.3: Riesz カーネルの位相特性 [29]

A.4　バタワースフィルタ

図 A.4 は，帯域通過成分を抽出する BPF の振幅特性である．BPF は，10 次のバタワースフィルタである．

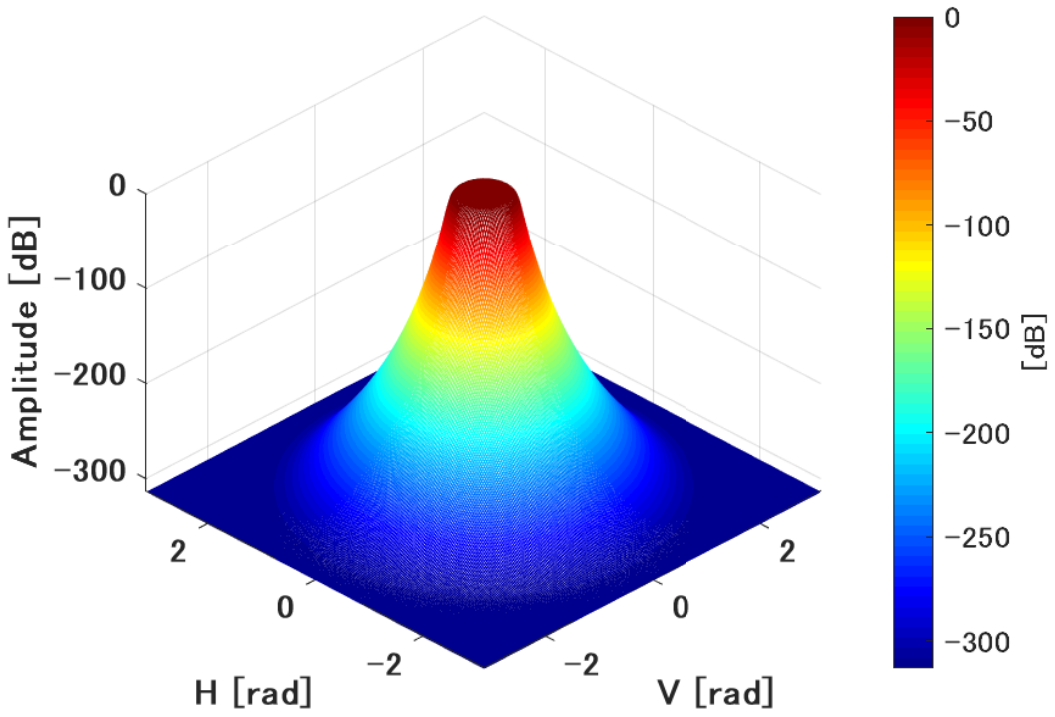

図 A.4: バタワースフィルタの振幅特性（10 次）

A.5　ピーク検出

図 A.5 と図 A.6 は，それぞれ Dhiman らの方法によるピーク検出と，自己相関法によるピーク検出の例示である．図の横軸と縦軸は，それぞれ周波数と振幅である．図 A.5 の凹凸は複雑であるのに対して，図 A.6 のピークは，100 Hz 付近の一箇所だけである．図の横軸と縦軸は，それぞれ周波数と相関値である．

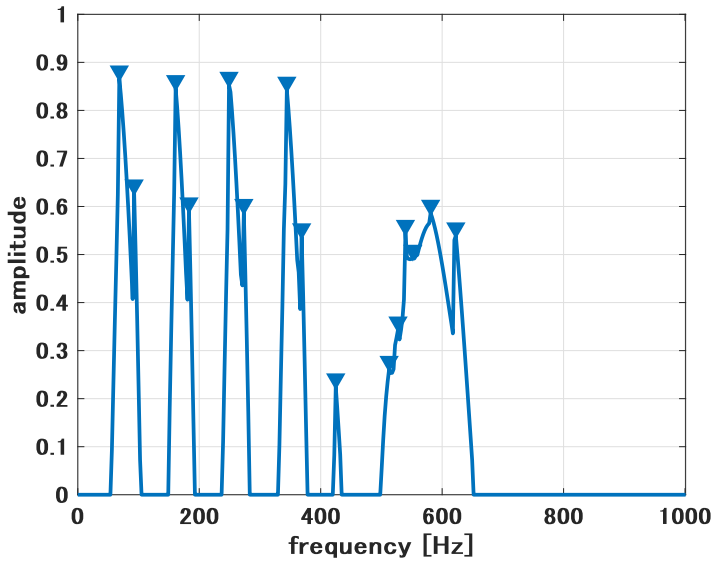

図 A.5: Dhiman らの方法のピーク検出

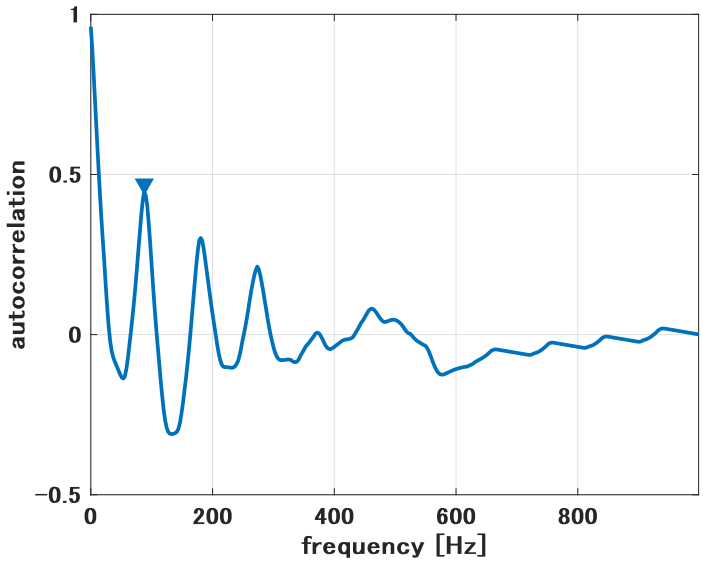

図 A.6: 自己相関法によるピーク検出

A.6　空間フィルタの効果

A.6.1　適切な空間フィルタ（調波性を抽出）

図 A.7 は，静音環境に見られた BPF と IF の対応例である．上段の横軸と縦軸は，それぞれ垂直成分（V）と水平成分（H）である．下段は，空間フィルタで抽出した成分から分離した IF である．

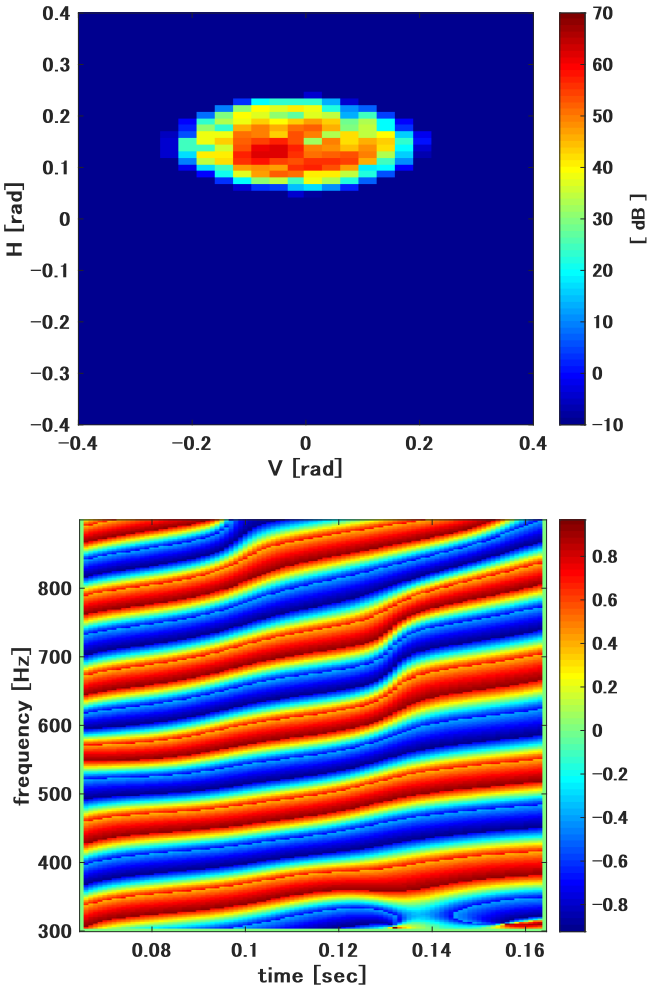

図 A.7: 帯域を制限した後（SNR ∞ [dB] の場合）

A.6.2　不適切な空間フィルタ（周期性を抽出）

図 A.8 は，雑音環境に見られた BPF と IF の対応例である．空間フィルタの領域を，原点に近しい左側の領域に適用すると，V 成分が IF に強く表れてしまう．

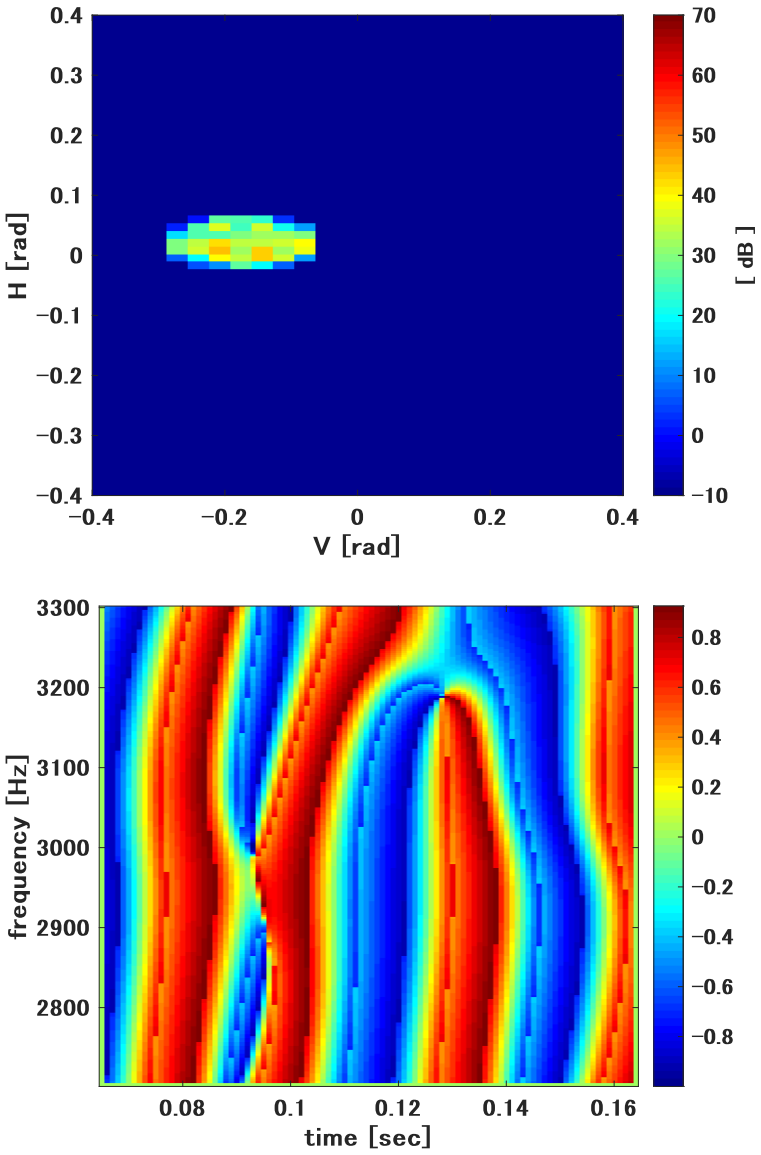

図 A.8: 誤って抽出された周期性の事例（SNR 0 [dB] の場合）

A.7　瞬時振幅によるマスク

A.7.1　雑音環境で生成したIA マスク

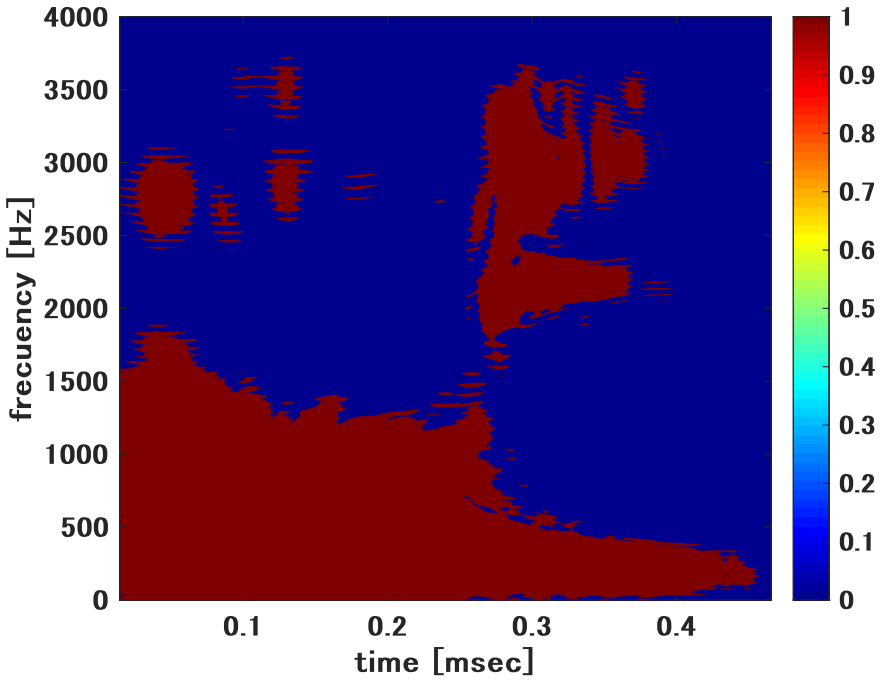

図 A.9: 瞬時振幅を利用するマスク（SNR 20 [dB] の場合）

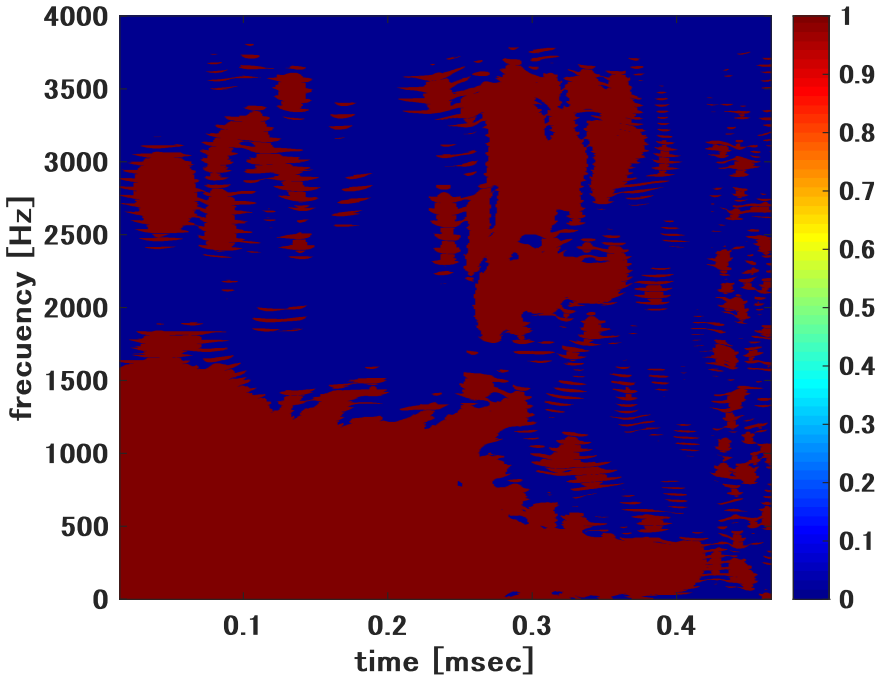

図 A.10: 瞬時振幅を利用するマスク（SNR 10 [dB] の場合）

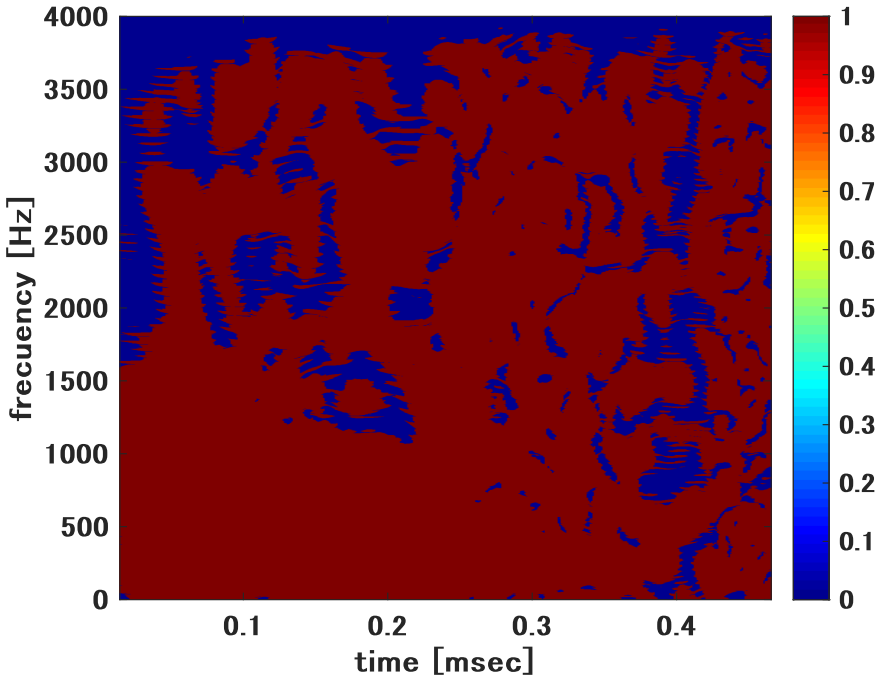

図 A.11: 瞬時振幅を利用するマスク（SNR 0 [dB] の場合）

A.7.2　残響環境で生成したIAマスク

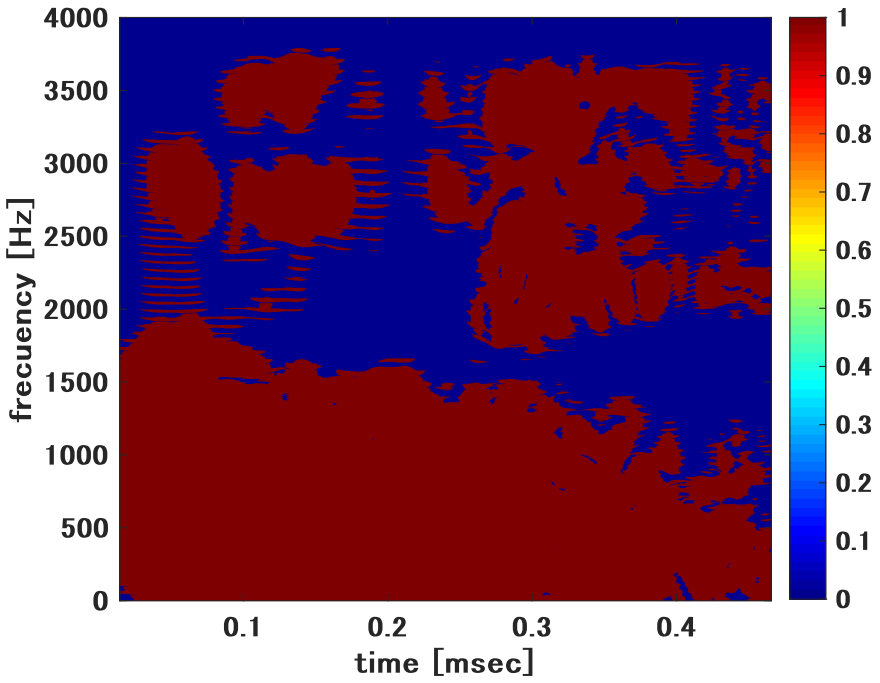

図 A.12: 瞬時振幅を利用するマスク（SNR ∞ [dB], TR 0.5 [sec] の場合）

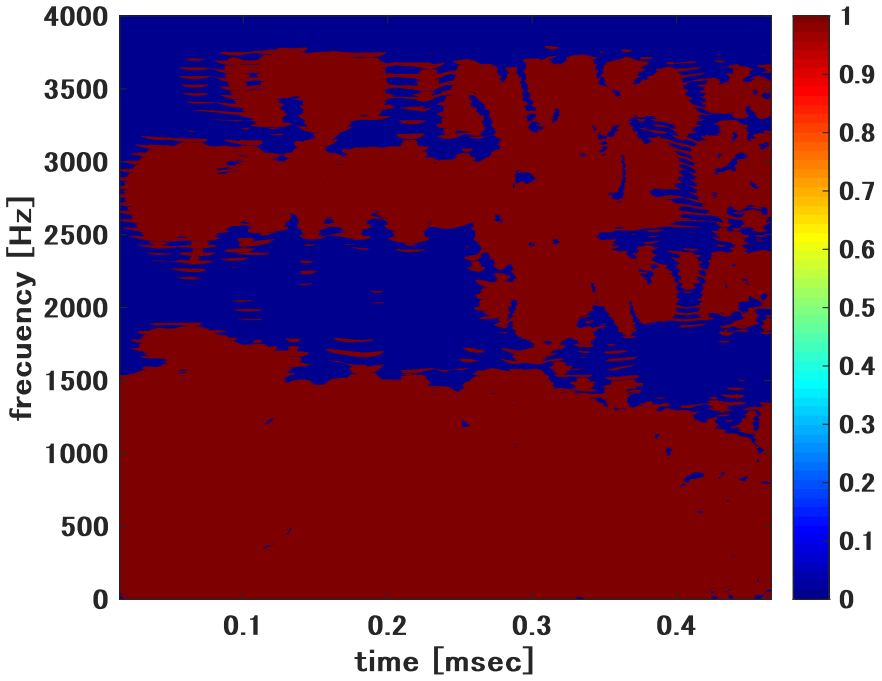

図 A.13: 瞬時振幅を利用するマスク（SNR ∞ [dB], TR 1.0 [sec] の場合）

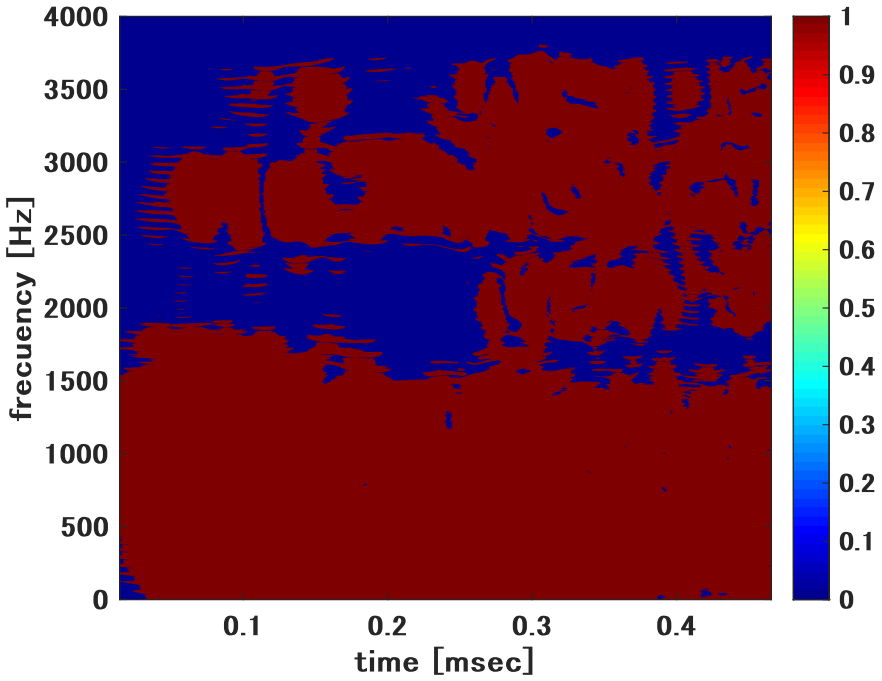

図 A.14: 瞬時振幅を利用するマスク（SNR ∞ [dB], TR 2.0 [sec] の場合）

A.8　コヒーレンスマップ

A.8.1　実装したコヒーレンスマップの生成条件

次式は，実装時に適用したコヒーレンスマップの生成条件を，Dhiman らの方法のコヒーレンスマップ $C(\boldsymbol{\omega})$ の生成式 [28] に反映したものである．

$$C(\boldsymbol{\omega}) \triangleq \begin{cases} \left(\dfrac{\lambda_1(\boldsymbol{\omega}) - \lambda_2(\boldsymbol{\omega})}{\lambda_1(\boldsymbol{\omega}) + \lambda_2(\boldsymbol{\omega})}\right)^2, & \lambda_1(\boldsymbol{\omega}) + \lambda_2(\boldsymbol{\omega}) \ > \ 10^{-6} \\ 0, & \text{otherwise} \end{cases} \tag{A.1}$$

ただし，$\lambda_1(\boldsymbol{\omega})$ と $\lambda_2(\boldsymbol{\omega})$ は，それぞれ構造テンソルの固有値である．

A.8.2　改良法により重み付けされた瞬時周波数（雑音環境）

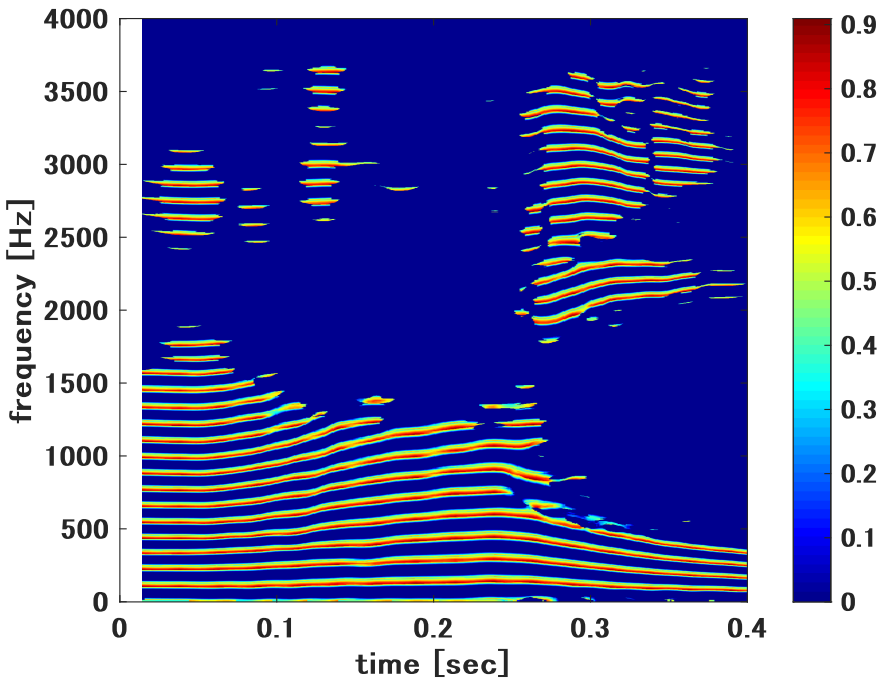

図 A.15: IA マスクで補強した WIF（SNR 20 [dB] の場合）

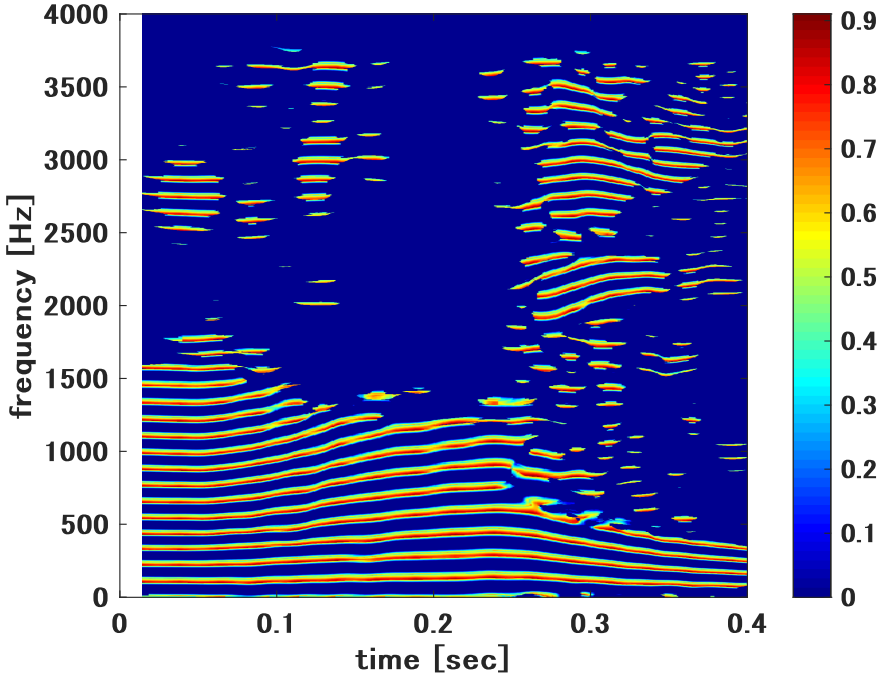

図 A.16: IA マスクで補強した WIF（SNR 10 [dB] の場合）

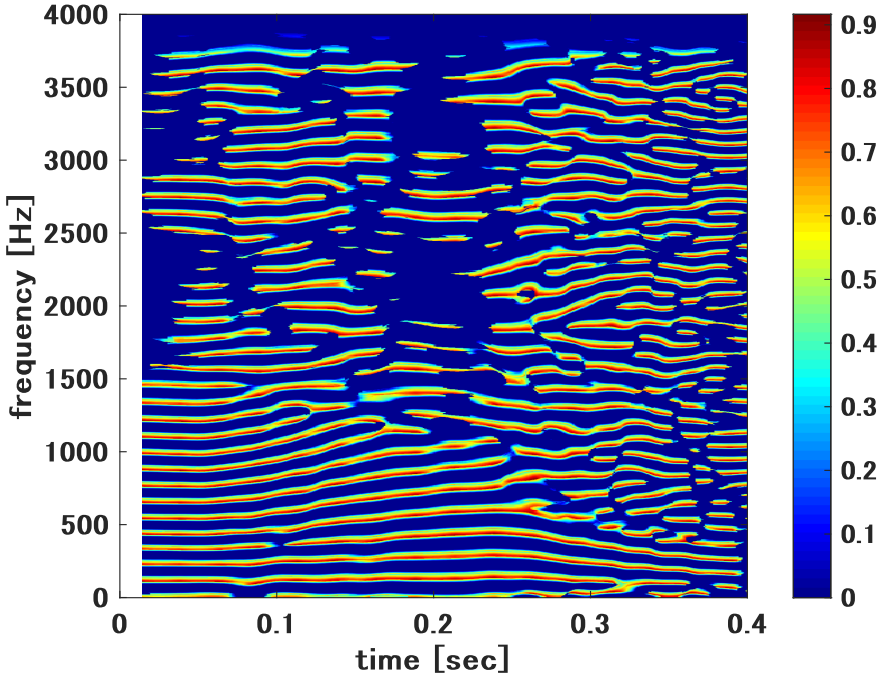

図 A.17: IA マスクで補強した WIF（SNR 0 [dB] の場合）

A.8.3　改良法により重み付けされた瞬時周波数（残響環境）

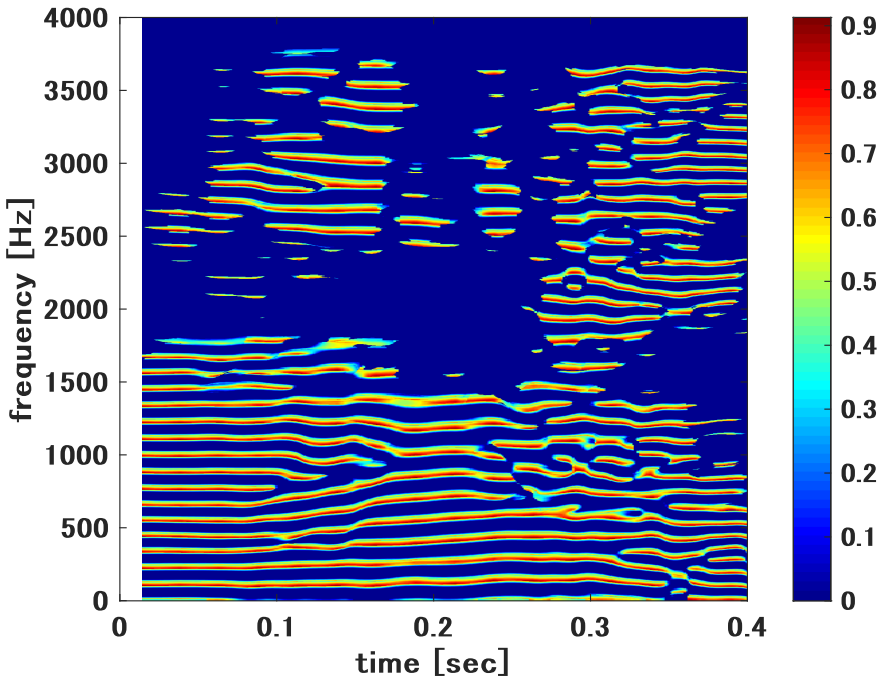

図 A.18: IA マスクで補強した WIF（SNR ∞ [dB], TR 0.5 [sec] の場合）

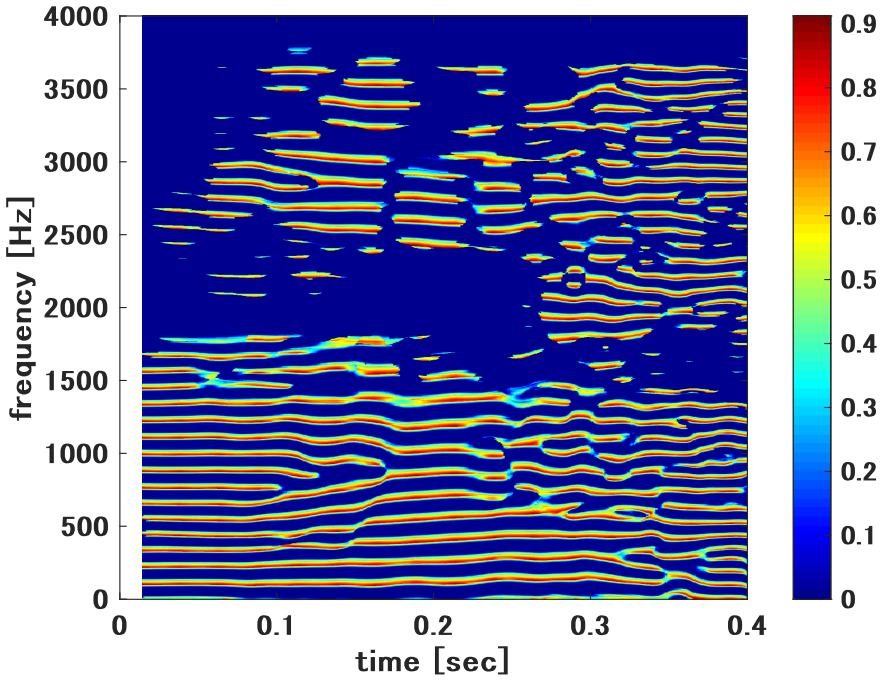

図 A.19: IA マスクで補強した WIF（SNR ∞ [dB], TR 1.0 [sec] の場合）

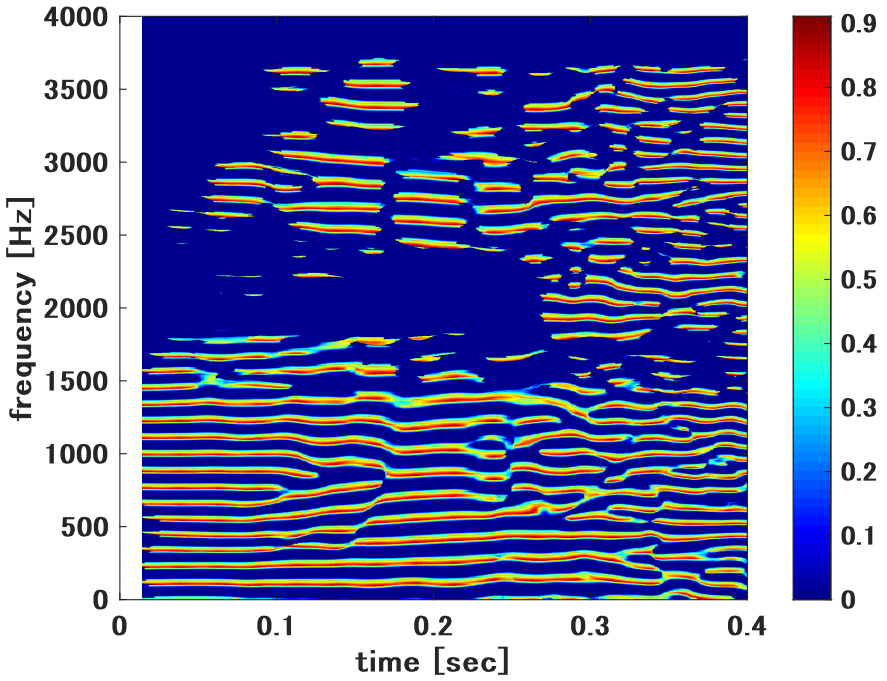

図 A.20: IA マスクで補強した WIF（SNR ∞ [dB], TR 2.0 [sec] の場合）