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

Title	時間情報と周波数情報を用いた雑音環境における基本周波数推定に関する研究
Author(s)	石本，祐一
Citation	
Issue Date	2004-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/951
Rights	
Description	Supervisor:赤木　正人，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

時間情報と周波数情報を用いた
雑音環境における音声の基本周波数推定に関する研究

指導教官　赤木 正人 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

石本 祐一

2004 年 1 月 22 日

Copyright © 2004 by Yuichi ISHIMOTO

要 旨

本研究では、雑音環境においても頑健で高精度な基本周波数推定を目指し、ヒトの聴覚機構と同様に時間情報と周波数情報の両方を利用した雑音に対しても頑健な手法と、帯域幅可変櫛形フィルタによる雑音抑圧、及び瞬時周波数の不動点を利用した高精度な手法の 3 つを組み合わせた基本周波数推定法を構築する。

音声の基本周波数は多くの音声情報処理において用いられる重要な特徴であるため、これまでにも様々な基本周波数推定法が提案されている。しかし、雑音の存在に対応した決定的な手法はまだ確立されていない。これらの多くは、基本周波数を表す情報として、時間領域に現れる音声の周期性の特徴 (時間情報) と周波数領域に現れる調波性の特徴 (周波数情報) のどちらかを利用している。低周波数帯域に強い雑音のエネルギーが存在すると音声波形は雑音の影響により歪み、音声波形の自己相関法のような時間情報のみを利用する推定法では推定誤差が大きくなる。一方、中・高周波数帯域の雑音エネルギーが強い場合には音声スペクトルの調波構造が歪み、ケプストラム法のような周波数情報のみを利用する推定法は雑音の影響を受けやすい。実環境雑音は特定の時間-周波数帯域に偏ったエネルギー分布を持つため、雑音の特性によって基本周波数抽出に有用な音声の特徴や周波数帯域は異なり、時間情報と周波数情報のどちらかのみを利用する従来手法では実環境雑音に対応できない。ヒトは雑音が存在していても基本周波数に対応する音の高さを知覚できるが、ヒトの音高知覚には時間情報と周波数情報の両方が用いられると考えられている。本研究では、上記の実環境雑音の特性から音声のエネルギーに対して雑音エネルギーが相対的に小さい帯域が存在すると考え、その帯域の時間情報と周波数情報の両方を利用し統合することで、雑音に対して頑健な推定を行う。ただし、この推定はクリーン音声に対する高精度な手法と比べると推定精度の点で劣るため、次に雑音を抑圧した後に高精度な基本周波数推定を行うことを考える。実環境雑音の特性は時々刻々と変化するために、雑音の特性を予め仮定する雑音抑圧手法は適していない。そのため、時間情報と周波数情報を用いて得られた推定基本周波数を利用して構築した櫛形フィルタを用いて雑音を抑圧する。その後、雑音抑圧音声に対して高精度な基本周波数推定を行う。以上の方策より、提案手法は次の 3 つの段階から構成される。

(1) 初期推定部 - 雑音を含む音声に対して瞬時振幅の時間-周波数表示に現れる

時間情報と周波数情報を利用し雑音に対して頑健な基本周波数推定を行う。雑音エネルギーの小さい帯域に現れる音声の基本周波数の時間情報と周波数情報を利用することで、雑音が存在する環境でも頑健な推定が可能となる。この初期推定部における推定結果は信号対雑音比が 3 dB 程度であっても推定可能区間の減少を抑えることができる。

(2) 雑音抑圧部 - 初期推定部で得られた基本周波数を用い、音声の調波成分を残す帯域幅可変櫛形フィルタによって雑音付加音声の雑音抑圧を行う。このとき、初期推定部で推定された基本周波数にはある程度の誤差が含まれているため、音声の調波成分が誤って除去されないように櫛形フィルタを設計する必要がある。よって、初期推定基本周波数の誤差の度合を推測し、櫛形フィルタの帯域幅を調節する。

(3) 最終推定部 - 雑音抑圧された音声に対して、瞬時周波数の不動点を利用した高精度な基本周波数推定を行う。瞬時周波数は雑音の影響を受けやすい特徴であるが、雑音抑圧部によって調波成分付近以外の雑音を抑圧できるため、雑音の存在にかかわらず高い精度を保ったままの基本周波数推定が可能となる。

計算機シミュレーションにより提案手法の耐雑音性能と推定精度を調査した結果、白色雑音やピンク雑音のような従来から評価に用いられていた人工的な雑音だけではなく、走行自動車内雑音のようにエネルギー分布が大きく偏った実環境雑音に対しても、クリーン音声に対する基本周波数推定とほぼ同程度の高い推定精度を得ることができ、本研究による手法が「雑音に対して頑健」と「高精度」の両方を満たす推定が可能であることがわかった。

雑音環境においても音声の基本周波数を推定できるという本研究の成果は、音声情報処理の幅広い分野に応用可能である。例えば、一般に雑音環境での自動音声認識では認識精度が大幅に低下することが問題とされているが、韻律辞書を作成し基本周波数の概形を特徴量として認識器に用いることにより認識精度の向上を図ることができる。また、基本周波数推定の精度が合成音声の自然性を左右する音声分析合成符号化においては、雑音の存在する実環境でのシステムの運用に貢献できる。さらに、様々な音が混じりあった状態から目的音声を抽出するカクテルパーティ効果の概念を基にした音源分離では、基本周波数を音源の違いを示す特徴として用いているため、本研究は聴覚情景解析の研究に対しても貢献できる。

目 次

1 序論 1

1.1 基本周波数抽出の重要性 2

1.2 基本周波数抽出の問題 5

1.3 従来の基本周波数推定法 7

1.3.1 周期性の特徴を利用した基本周波数推定法 8

1.3.2 調波性の特徴を利用した基本周波数推定法 10

1.3.3 耐雑音性能と推定精度 11

1.4 ヒトのピッチ知覚 21

1.5 実環境雑音の特性 24

1.6 実環境雑音に対する基本周波数推定法の構築 26

1.6.1 時間-周波数領域の基本周波数情報 26

1.6.2 基本周波数を利用した雑音抑圧 28

1.7 本研究の目的 29

1.8 本論文の構成 31

2 雑音環境における基本周波数推定法の構成 34

2.1 はじめに 35

2.2 雑音環境における基本周波数推定法 36

2.2.1 雑音に頑健な基本周波数推定 (初期推定部) 36

2.2.2 基本周期を利用した雑音抑圧 (雑音抑圧部) 39

2.2.3 高精度な基本周波数推定 (最終推定部) 40

2.3 まとめ 40

3 瞬時振幅の周期性・調波性を利用した基本周波数推定 42
3.1 はじめに . . . 43
3.2 瞬時振幅の周期性・調波性を利用した基本周波数推定の概要 . . . 43
3.3 瞬時振幅の時間-周波数表現 . . . 45
3.4 周期性・調波性に対する自己相関処理 . . . 48
3.5 自己相関係数の統合方法 . . . 52
3.6 基本周波数推定実験 . . . 55
3.6.1 雑音に対する頑健性に関する検証 . . . 56
3.6.2 推定精度と相関係数の値の関係の検証 . . . 60
3.7 まとめ . . . 62

4 帯域幅可変櫛形フィルタによる雑音抑圧 64
4.1 はじめに . . . 65
4.2 櫛形フィルタの定式化 . . . 65
4.3 帯域幅パラメータの決定 . . . 67
4.3.1 初期推定基本周波数の誤差と櫛形フィルタの帯域幅の関係 . 67
4.3.2 櫛形フィルタの帯域幅の決定方法 . . . 71
4.4 基本周期を一定とする波形の時間伸縮 . . . 72
4.5 雑音抑圧性能の検証 . . . 73
4.5.1 実験条件 . . . 73
4.5.2 実験結果 . . . 73
4.6 まとめ . . . 75

5 雑音環境における有効性検証 76
5.1 はじめに . . . 77
5.2 計算機シミュレーション . . . 77
5.2.1 実験条件 . . . 77
5.2.2 白色雑音 . . . 79
5.2.3 ピンク雑音 . . . 79
5.2.4 走行自動車内雑音 . . . 82
5.2.5 デパート内雑音 . . . 82

5.2.6 考察 85
5.3 まとめ 88

6 結論 89
6.1 本論文の要約 90
6.2 今後の課題 92

付録 94
A 時間領域に現れる周期性の特徴を利用する基本周波数推定法 94
A.1 複数窓幅から得られた自己相関関数を用いる推定法 94
A.2 YIN 96
B 周波数領域に現れる調波性の特徴を利用する基本周波数推定法 . . . 98
B.1 移動平均と帯域制限を用いたケプストラムによる推定法 . . 98
B.2 STRAIGHT-TEMPO 100

謝辞 104

参考文献 105

本研究に関する発表論文 114

第 1 章

序論

1.1 基本周波数抽出の重要性

音声は肺から押し出された呼気が声帯や声道を通って作られる音波であり、一般に、声帯の振動を伴う有声音と振動を伴わない無声音に大別される。図 1.1 に発声器官の模式図を示す。有声音を発声するときには、まず肺からの呼気が気管支を経て気管に送られ、声帯の開閉運動によって断続流 (励振波) を生じる。この流れは、声帯の部分から喉頭、咽頭、口腔を通って唇にいたる声道内を伝播する間に、声道の伝達特性によって特定の周波数成分が強調されたり減衰したりして、唇あるいは鼻から放射される。すなわち、図 1.2 に示すように、声帯振動による音源の生成、声道の形による調音、唇または鼻孔からの放射の作用によって音声波形が生成される [1, 2]。有声音には図 1.3 のようなほぼ相似的な波の繰返しがみられ、この繰返しの周期を一般に基本周期と呼び、その逆数を基本周波数と呼んでいる。基本周期は声帯振動によって作られる励振波の周期に等しい。また、有声音の周波数スペクトルは、図 1.4 のような、基本周波数とその整数倍の成分周波数からなる、いわゆる調波構造になっている [3, 4]。

基本周波数は聴感上では音の高さ (ピッチ) に対応し、基本周波数の緩やかな時間的変化は抑揚となる。従って、基本周波数の違いは男声女声や個人間の音色の違いとして現れ、基本周波数の時間的変化は話者のくせや方言などの特徴として現れる。そのため、基本周波数は個人個人を識別する手がかりの一つとなる。その他にも、アクセントによる高さの高低差で意味が変わる言葉 (「橋」と「箸」、「飴」と「雨」など) もあり、音の高さは言葉の意味の形成にも用いられている。このように、人間が音声を知覚する上で韻律情報が大きな役割を担っていることから、基本周波数はアクセントやイントネーションによる言語学的な研究対象となっているが、その他にも工学的応用として話者認識や音声認識への利用が考えられている。

例えば、自動音声認識においては、音声の韻律的特徴を句や単語の区分化に利用したり、韻律辞書を作成し基本周波数の概形を利用することで、認識精度を向上させることができる [5, 6, 7]。他にも、音声は声帯振動からなる音源情報と声道特性からなるスペクトル包絡情報によって生成されていると考えられることから、ボコーダ型の音声分析合成においては、音声を分析部で音源情報とスペクトル包

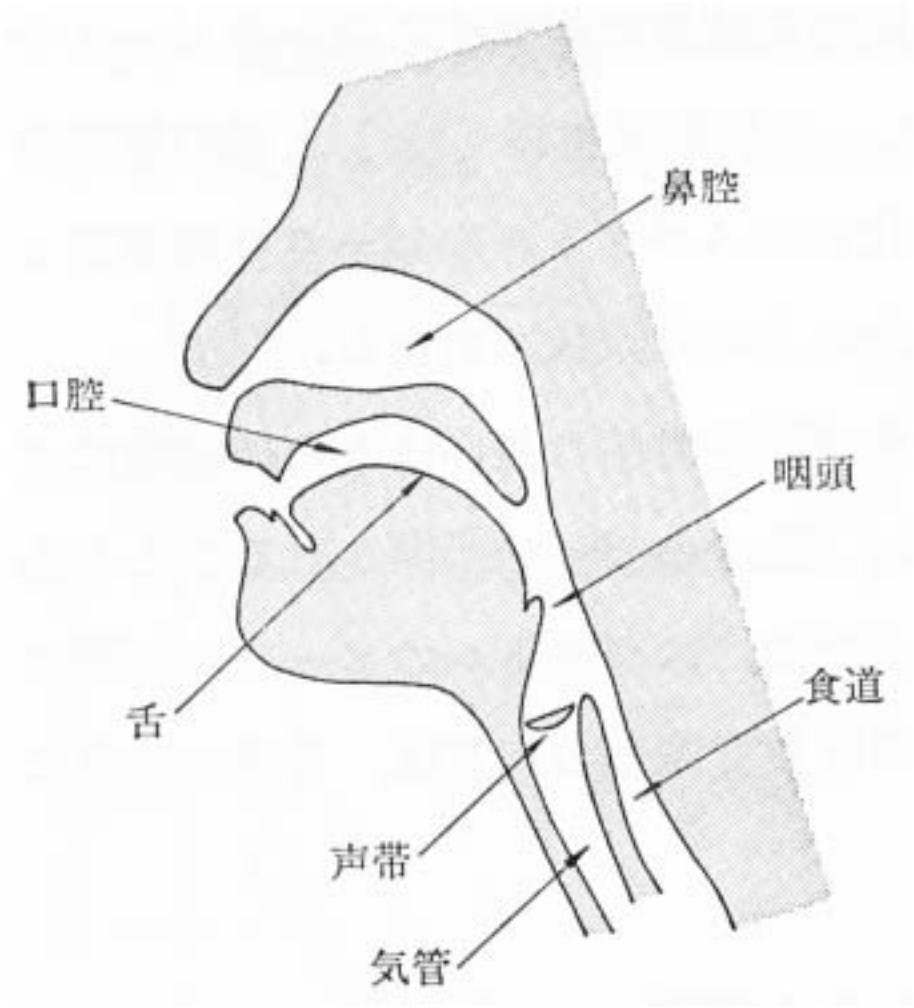

図 1.1: 発声器官の模式図 (城戸ら [1])

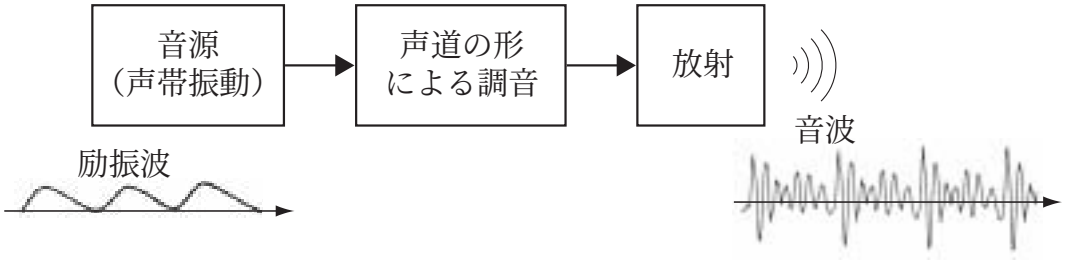

図 1.2: 有声音生成の基本形

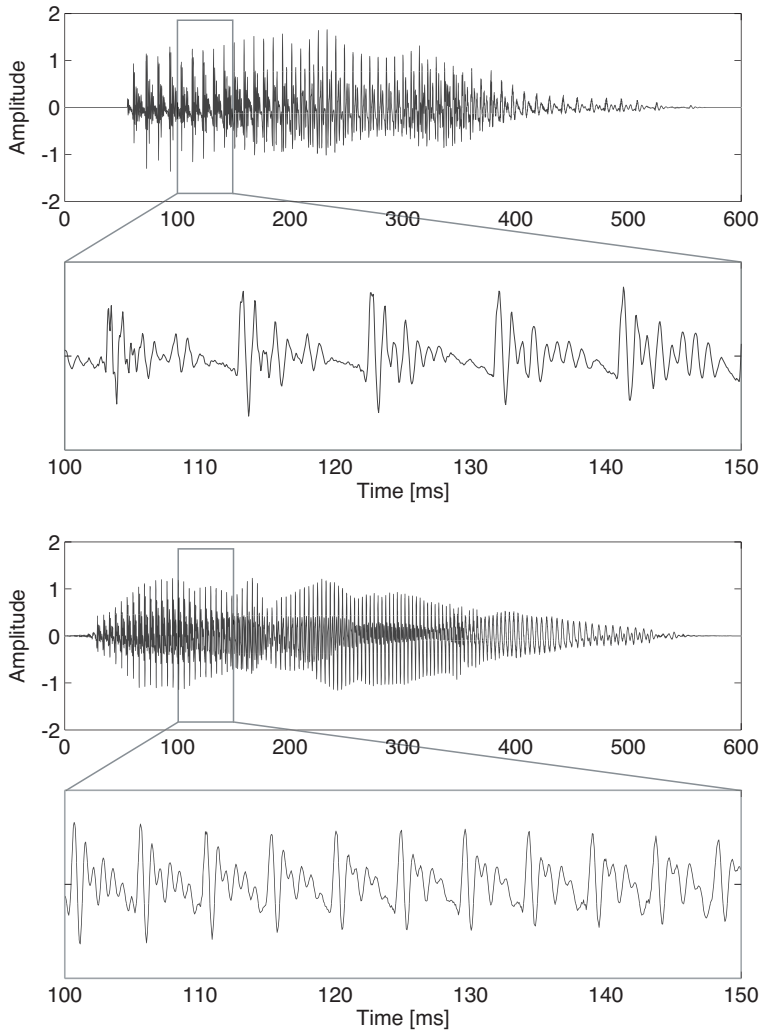

図 1.3: 有声音の音声波形: (上) 男声 /aoi/ (下) 女声 /aoi/

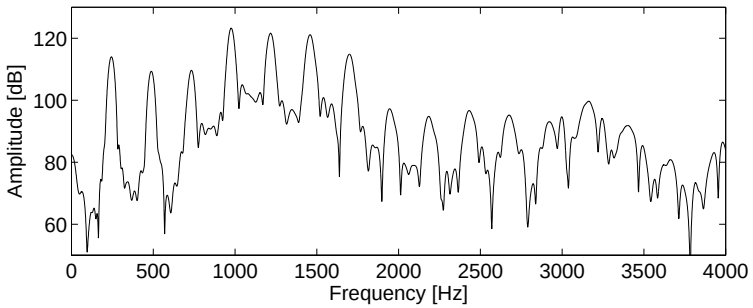

図 1.4: 女声/a/の振幅スペクトル

絡情報に分けることにより、音声波形の冗長な成分を捨てた特徴パラメータの抽出と符号化がなされている [8, 9, 10]。このとき、音源情報としての基本周波数が必要となってくるため、音声分析合成においても基本周波数抽出は重要な課題である。特に自然性の高い音声合成のためには高精度の基本周波数抽出が必要である [11]。また、Bregman は、聴覚の情景解析の研究において、人間の聴覚機構は音源の違いを特徴づけるもののひとつとして基本周波数を利用している、と結論づけている [12]。すなわち、人間が同時に発声された音や雑音を含む音の中から目的の話者の声を聞き分けることにも、基本周波数は利用されていると考えられる。聴覚の情景解析、特に音源分離においては、雑音環境において目的音声の基本周波数を利用した計算論的聴覚情景解析の研究が多くなされている [13, 14]。

1.2 基本周波数抽出の問題

基本周波数と韻律との密接な関係から音声の基本周波数は利用範囲が広く、韻律情報を利用する音声情報処理の様々な分野において基本周波数を抽出することが重要な問題となっている。しかし、次のような理由から、基本周波数抽出は容易ではない [15, 16, 17]。

1. 音声波形は完全な周期波ではなく、振幅や周期が声門の励振ごとに変化する準周期的な波である。そのため、励振波の変化に追随した抽出を行わなければならない。特に語頭・語尾では声帯振動が完全な周期性を持たず振幅の変化も大きいため、声帯振動周期を基本周波数として求めることが難しい。また、音声波形は声帯振動による励振波が声道で形作られた音響管で共振して生成されると考えられるが、この声道の伝達特性の極や零点により音声は振幅の異なる調波成分を持つようになる。基本周波数抽出のためには励振波が得られることが理想的だが、音声波から声道の影響を取り除き、励振波のみを直接取り出すことは困難である。
2. 基本周波数の変化範囲が広い。男性の平均基本周波数は 90–130 Hz、女性の平均基本周波数は 250–330 Hz であり、図 1.3 に示されるように女性の基本周波数は男性のほぼ 2 倍となる。また、一般に基本周波数の変動範囲は平均基

本周波数に比例して大きくなる傾向にあり、図 1.5 に示すように発声者ごとの基本周波数変動の標準偏差も女声は男声に比べて 2 倍程度になる [18]。すなわち、男声の基本周波数の 2 倍に対応する調波成分及びその整数倍に対応する調波成分が、女声の調波成分とほぼ等しくなる。このような幅広い周波数に対応することを考慮しなければならず、基本周波数の 2 倍の周波数を抽出してしまう誤り (倍ピッチ) や基本周波数の 1/2 を抽出してしまう誤り (半ピッチ) が起こりやすい。

3. 発声される状況により様々な雑音が音声に含まれる。雑音により音声の調波構造が歪むため、雑音を含む音声から基本周波数を抽出することは困難である。喉頭部に電極を貼り付けて声帯振動を電気信号で観測した Electroglottograph(EGG) や、ヘッドセットなどにより口に近接させたマイクロホンを用いて記録した音波ならば、周囲の雑音の影響を小さくすることができるが、実際的な問題として常にこのような環境で音声を観測することは容易ではない。また、音源分離問題では音声の他になんらかの妨害音が存在することが前提となっているため、雑音の存在しない状況は望めない。実環境において、音声認識や音声分析合成、音源分離といった様々な音声情報処理を利用するシステムを考えてみると、システムの周囲には雑音が存在し、観測された音声波形が周囲の雑音を含むことは避け難い問題となる。

今後、音声情報処理の実用を進めていく上で、上記の問題のうち、3 番目に挙げた雑音環境への対応、特に実環境雑音への対応が重要になってくると考えられる。例えば、携帯電話などに用いられている音声分析合成システムでは、雑音の影響により基本周波数の推定誤差が大きくなると音質が低下してしまうため、雑音を含む音声からでも基本周波数が推定できなければならない。また、周囲に雑音が存在する環境下では音声認識の認識精度が大幅に低下してしまうが、雑音中の音声から基本周波数を推定できれば、基本周波数を特徴量のひとつとして認識器に用いることで認識精度を向上させることができる。その他にも、音源分離システムを利用して雑音環境下でも目的音声のみを強調させる補聴器が考えられているが、周囲の音から目的音声を分離する手がかりのひとつとして基本周波数が用いられる。このように、雑音を含む音声から基本周波数が推定できれば、音声情報

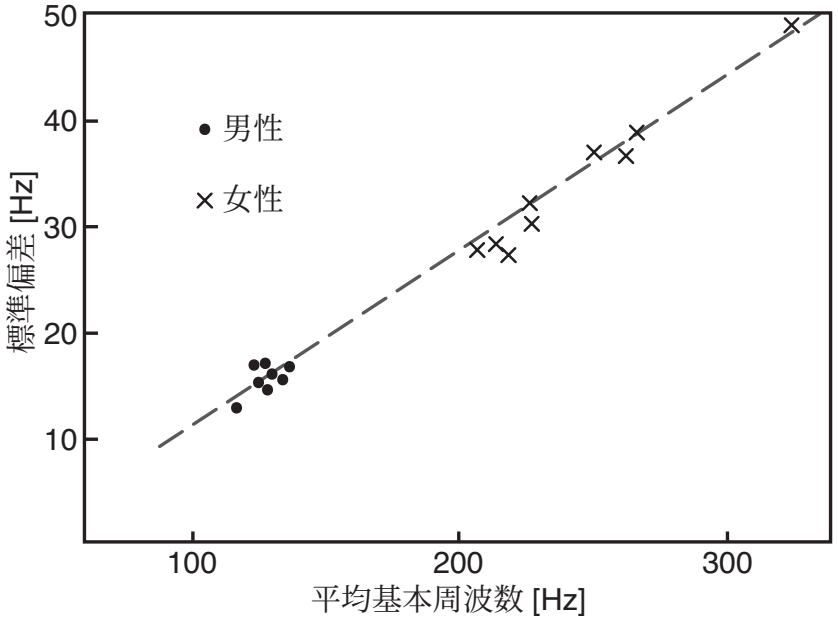

図 1.5: 基本周波数の平均値と標準偏差 (斉藤ら [18])

処理を利用した様々なシステムへ応用することが可能である。

1.3 従来の基本周波数推定法

ここで、従来の基本周波数推定法について述べる。基本周波数抽出は音声分析の初期からの課題であり、前節に記した問題から様々な種類の手法が提案されてきた [19, 20, 21]。しかし、実環境における雑音の存在に対応した決定的な方法はまだ確立されていない。

従来の基本周波数推定法は

1. 時間領域に現れる周期性の特徴 (時間情報) を利用した手法
2. 周波数領域に現れる調波性の特徴 (周波数情報) を利用した手法

に分類できる。有声音は前述のように、時間方向に基本周波数の逆数である基本周期を周期としてもつ波形となるため、この周期性の検出を対象とするものが、時間領域に現れる周期性の特徴を利用した手法である。また、図 1.4 に示したように、周波数方向には基本周波数とその整数倍の周波数帯域に調波成分が櫛の歯の

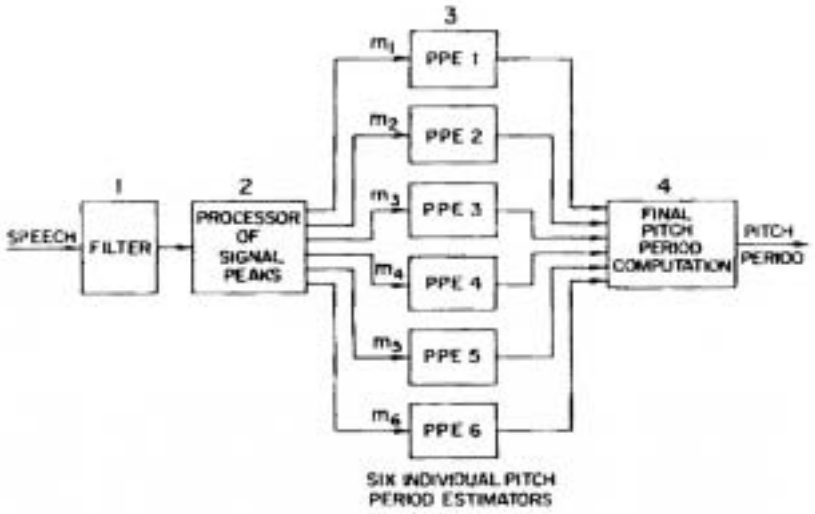

図 1.6: 並列処理法 (Gold and Rabiner[22])

ように現れるため、この調波成分のうちの最低周波数や調波成分間隔の検出を対象とするものが、周波数領域に現れる調波性の特徴を利用した手法である。

1.3.1 周期性の特徴を利用した基本周波数推定法

基本周波数抽出に関する研究の初期において、元の波形の周期性を保ちつつ基本周期に無関係な特徴を捨て去るという考えから、波形からピーク位置や零交差点を検出してその時間間隔を利用する方法が提案された。これは声門閉鎖の時刻を求め、その時間間隔を検出することに等しい。図 1.6 に示す Gold and Rabiner によるピーク位置検出による並列処理法 [22] や Geçkinli and Yavuz による零交差間隔による手法 [23] がその代表であり、これらの手法は実装が容易であるという利点がある。しかし、音声波形は声道特性により大きく変形し、また観測時に雑音が含まれることが多いが、波形のピーク位置や零交差点はそれらの影響を受けやすいために高い精度は期待できず、特に実環境における利用には適していない。

周期性の特徴を抽出する方法としては他に、音声波形の自己相関処理も古くから用いられている。自己相関関数の係数の極大値からその波形の周期間隔を求める手法は実装が容易であるうえにピーク位置検出よりも雑音の影響を受け難い。そのため、音声波形をセンタクリッピングすることにより声道特性の影響を軽減す

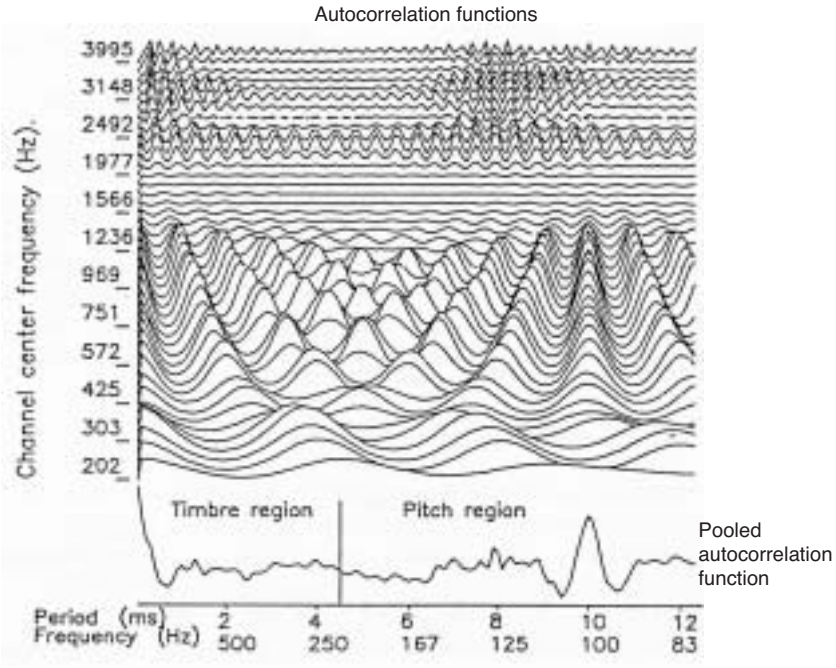

図 1.7: 自己相関関数によるピッチ知覚の聴覚モデル (Meddis[35])

る方法 [24] や自己相関関数の乗法計算を減らすために平均振幅差関数 (AMDF) を用いる方法 [25] などの初期における手法をはじめ、多数の改良法が提案されてきた。近年では複数の異なる幅の分析窓を用いて自己相関関数を計算し基本周波数の候補の中から最適値を選択する手法 [26] や、AMDF の逆数で自己相関係数の重み付けをすることにより耐雑音性を向上させる手法 [27]、波形の振幅差関数を重み付けする手法 (YIN)[28] などが提案されている。

音声波形をそのまま利用するのではなく、LPC 分析の残差信号の自己相関処理による手法も考えられている [29]。LPC 分析の残差信号は励振波の近似と考えられ、基本周期の間隔でピークを持つ。そのため、ピーク検出や自己相関処理によって基本周波数の抽出が可能となる。また、Markel はこれを発展させ、LPC 分析による逆フィルタを用いて声道特性を取り除いた信号を得た後に自己相関処理による基本周波数推定を行う手法 (SIFT) を提案した [30]。しかし、LPC 分析自体が雑音の影響を受けやすいという問題があり、実環境での基本周波数抽出には適しているとはいえない。

そのほかには、ピッチ知覚を説明する聴覚モデルとして、図 1.7 に示すような聴覚フィルタと自己相関関数を組み合わせたモデルも提案されている [31]。このよう

表 1.1: 基本周波数推定法 (時間領域の周期性の特徴を利用)

推定法	特徴	参考文献
ピーク検出		[22] [36]
零交差点検出		[23] [37]
自己相関処理		[24] [38] [39] [40] [41]
	正規化相互相関	[42] [43] [44]
	複数窓による分析	[26]
	AMDF で重み付け	[27]
	LPC 残差の相関	[29]
	SIFT	[30]
	聴覚モデル	[31] [35] [32] [33]
波形差分	平均振幅差関数 (AMDF)	[25] [45] [46]
	振幅差関数に重み付け (YIN)	[28]
	聴覚モデル	[34] [47] [48]

なピッチ知覚と自己相関処理の関連を考慮した聴覚モデルは多い [32, 33, 34]。

表 1.1 に周期性の特徴を利用した基本周波数推定法の一覧を示す。

1.3.2 調波性の特徴を利用した基本周波数推定法

短時間フーリエ変換などにより音声を時間-周波数分析することによって得られる時間-周波数表現は音声の調波構造を表すことができる [49]。周波数領域の調波性の特徴から基本周波数を抽出する手法として、対数スペクトルの自己相関関数を利用する手法 [50] や、comb filter の通過量が最大となるように comb filter の中心周波数を求めることで調波の存在する周波数を決定する振幅スペクトルの comb filtering による手法 [51]、基本波と高調波に対応する周波数成分の和を求める手法 [52] などがある。フィルタバンクで分析され時間-周波数領域で表された音声信号の瞬時振幅は雑音を含む音声であっても調波構造をよく表すことができるため、Unoki and Akagi は音源分離モデルを構築するためにこの瞬時振幅の comb filtering による基本周波数推定法を用いている [14]。しかし、この手法は、Unoki らが考慮

していた雑音を含む単母音に対して基本周波数を推定することはできるが、連続発話音声の推定基本周波数に対しては十分な精度が得られない。

ケプストラムは音声スペクトルをスペクトル包絡と微細構造に分離できる特徴がある。対数振幅スペクトルの逆フーリエ変換を求めることにより、音声波形はケプストラム係数へと変換される。ケプストラム係数はケフレンシと呼ばれる時間領域の値であるが、スペクトル上の微細構造が高ケフレンシ部のピークとして現れ、スペクトル包絡が 0–4 ms 程度の低ケフレンシ部に集中する。この高ケフレンシ部のピークから基本周期が求められる [53, 54]。この処理は声道の影響を取り除き励振波の情報を取り出すことに等しい。ケプストラムを利用する手法としては、移動平均と帯域制限を用いた手法 [55]、音声波形や対数スペクトルをクリッピングしてからケプストラムを求める方法 [56, 57]、ハフ変換によりケプストラムの時間連続性を考慮する方法 [7] などがある。

また、近年、音声の瞬時周波数が高精度な基本周波数抽出のために用いられている [58, 59, 60, 61]。これらはフィルタバンク出力の瞬時周波数とフィルタの中心周波数が一致する周波数が調波の存在する周波数と考えられることを利用している。例えば、Kawahara らは音声分析合成システムを構築するために瞬時周波数を利用した基本周波数推定法 (STRAIGHT-TEMPO) を提案している [62]。この手法はクリーン音声から高精度な基本周波数を抽出することができるが、雑音に弱い。そのため、雑音環境を考慮して、基本波だけではなく調波成分も利用した方法 [63] や、各調波成分がその近傍の周波数帯域を占有している度合を利用した手法 [64] も提案されている。これらの手法は STRAIGHT-TEMPO よりもわずかながら耐雑音性能が向上している。

表 1.2 に調波性の特徴を利用した基本周波数推定法の一覧を示す。

1.3.3 耐雑音性能と推定精度

1.3.1 節及び 1.3.2 説で述べたようにこれまでに提案された基本周波数推定法は多数にのぼるため、そのすべてについて検証を行うことは困難である。そこで、本節では、周期性の特徴を利用する基本周波数推定法から、

- 複数窓幅から得られた自己相関関数を用いる推定法 (AC)[26]

表 1.2: 基本周波数推定法 (周波数領域の調波性の特徴を利用)

推定法	特徴	参考文献
スペクトル分析	自己相関	[65] [50]
	櫛形フィルタ	[66] [67] [14] [68] [51]
	ラグ窓利用	[69] [70]
	調波成分検出	[71] [72] [73]
	基本波と高調波の和を利用 (SHS)	[52]
	基本波フィルタリング	[74] [75]
ケプストラム		[53][54]
	音声波形のクリッピング	[56]
	対数スペクトルのクリッピング	[57]
	移動平均と帯域制限を利用	[55]
	ハフ変換	[7]
瞬時周波数		[58] [59] [76] [60] [61] [77]
	STRAIGHT-TEMPO	[62]
	調波成分を利用	[63]
	調波成分の占有度を利用	[64]

- 振幅差関数に重み付けを行う推定法 (YIN)[28]

を取り上げる。複数窓幅から得られた自己相関関数を用いる推定法は自己相関処理を利用する手法のなかで、予備実験において最も雑音に対する頑健性を示した手法であり、YIN は周期性の特徴を利用する手法のなかで最新の手法である。また、調波性の特徴を利用する基本周波数推定法からは、

- 移動平均と帯域制限を用いたケプストラムによる推定法 (CEP)[55]

- 瞬時周波数の不動点を利用した推定法 (STRAIGHT-TEMPO)[62]

を取り上げる。移動平均と帯域制限を用いたケプストラムによる推定法は、ケプストラムを利用する手法のなかで予備実験において最も雑音に対する頑健性を示した手法であり、STRAIGHT-TEMPO は調波性の特徴を利用する手法のなかで

最も高精度な推定が可能な手法である。以後、複数窓幅から得られた自己相関関数を用いる推定法を自己相関法と呼び、移動平均と帯域制限を用いたケプストラムによる推定法をケプストラム法と呼ぶこととする。4 つの手法の概要については付録に記す。

本節では、各手法の耐雑音性能と推定精度についての検証を行う。ここで、「耐雑音性能 (頑健性)」とは基本周波数が存在する音声区間でその基本周波数に近い値を抽出できるかどうかを表し、「推定精度」とは正しい基本周波数と推定基本周波数がどれ程異なるかを表すものとする。

正解基本周波数の抽出

- 目的

推定精度の検証を行うためには、推定結果との比較に用いる正しい基本周波数を得なければならない。しかし、基本周波数の絶対的な正解値を得ることは困難である。そこで、正解基本周波数を求めるために用いる手法について調査した。上記の 4 つの基本周波数推定法を用い、それぞれ雑音を含まないクリーンな音声波形と Electroglottograph(EGG)[1]から推定基本周波数を求め、次の 2 つの比較を行なった。

比較 1 各推定法により EGG 波形から得られた基本周波数を比較する。

比較 2 各推定法について、音声波形から得られた基本周波数と EGG 波形から得られた基本周波数を比較する。

これは EGG 波形が声道の影響を受けず、声帯振動をよく表しているためである。比較 1 において他の推定法との差が小さく、かつ比較 2 において音声波形から得られた基本周波数と EGG 波形から得られた基本周波数の差が小さいならば、その推定法は高い精度で正しい音源情報を抽出することができていると考えられる。

- 実験条件

[1]前頸部側壁に電極を置き、電気的インピーダンスの変化を測定したもの。声門を横切る電気的インピーダンスは声門の開閉状態によって変化するので、EGG 波形は声帯振動を記録しているといえる。

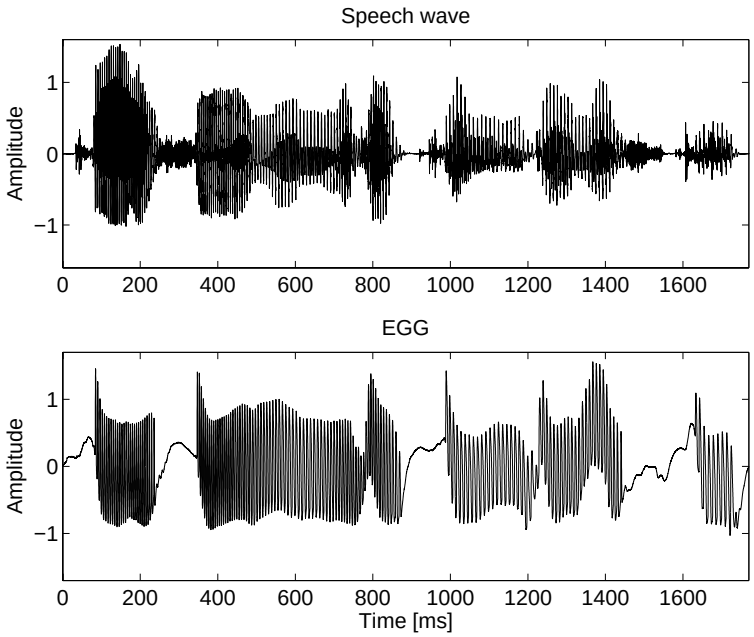

図 1.8: 基本周波数推定実験に用いる信号: (上) 音声波形、(下)EGG 波形

実験に用いた音声データは阿竹らによる音声と EGG が同時収録されたデータベース [63] を用いた。データ数は男女各 14 名の発話による 30 文章 (計 840 文) である。音声波形と EGG 波形の例を図 1.8 に示す。音声データ及び EGG 波形のサンプリング周波数は 16 kHz である。

- 評価尺度

有声/無声判定は考慮しないため、評価は有声区間のみで行なった。推定誤差 $e(n)$ は次式から得られる。

$$e(n) = F_0(n) - F_e(n) \tag{1.1}$$

ここで、$F_0(n)$ は正解基本周波数であり、$F_e(n)$ は推定基本周波数である。評価尺度として、比較 1 については、有声区間 N_v 内の誤差の平均値

$$\bar{e} = \frac{1}{N_v} \sum_{n}^{N_v} |e(n)| \tag{1.2}$$

を用いた。

比較 2 については、次の 2 種類の評価尺度を用いた。

Gross error : 有声区間 N_v において、$|e(n)| \geq 0.2 \times F_0(n)$ である区間 N_e の割合

$$\text{Gross error} = \frac{N_e}{N_v} \times 100 \quad [\%] \tag{1.3}$$

表 1.3: EGG 波形から得られた基本周波数の差 [Hz]

推定法	AC	YIN	CEP	TEMPO
AC	—	4.5	3.8	2.8
YIN	4.5	—	6.1	4.9
CEP	3.8	6.1	—	2.9
TEMPO	2.8	4.9	2.9	—

これは推定誤差が ±20%以上である区間の量を表す。

Fine error : $|e(n)| < 0.2 \times F_0(n)$ である区間 N_c における推定誤差の平均

$$\text{Fine error} = \frac{1}{N_c} \sum_{n}^{N_c} \frac{|e(n)|}{F_0(n)} \times 100 \quad [\%] \tag{1.4}$$

これは推定誤差が ±20%未満である区間内での誤差の平均を表す。

Gross error は基本周波数が推定できなかった区間を表すことから頑健性を示すと考えられる。また、Fine error は基本周波数が推定できた区間内における推定基本周波数の誤差を表すことから、推定精度を示すと考えられる。比較 2 においては音声波形から得た基本周波数を推定基本周波数とし、EGG 波形から各推定法によって得た基本周波数をその推定法に対する正解基本周波数とした。また、周波数差の絶対値 (Hz) ではなく比率 (%) を評価尺度に用いたのは、図 1.5 に示されるように基本周波数の平均と標準偏差が比例関係にあるためである。

- 実験結果 (比較 1)

EGG 波形から得られた基本周波数について各推定法間の差の平均を表 1.3 に示す。自己相関法とケプストラム法については STRAIGHT-TEMPO との差が最も小さく、それぞれ 2.8 Hz と 2.9 Hz である。YIN については、自己相関法との差が 4.5 Hz と最も小さくなっているが、STRAIGHT-TEMPO との差はそれについで小さい 4.9 Hz であり、ケプストラム法との差である 6.1 Hz と比べると大差ない。したがって、4 つの基本周波数推定法のなかでは、STRAIGHT-TEMPO が最も他の手法との差が少ない基本周波数推定を行っていると考えられる。

- 実験結果 (比較 2)

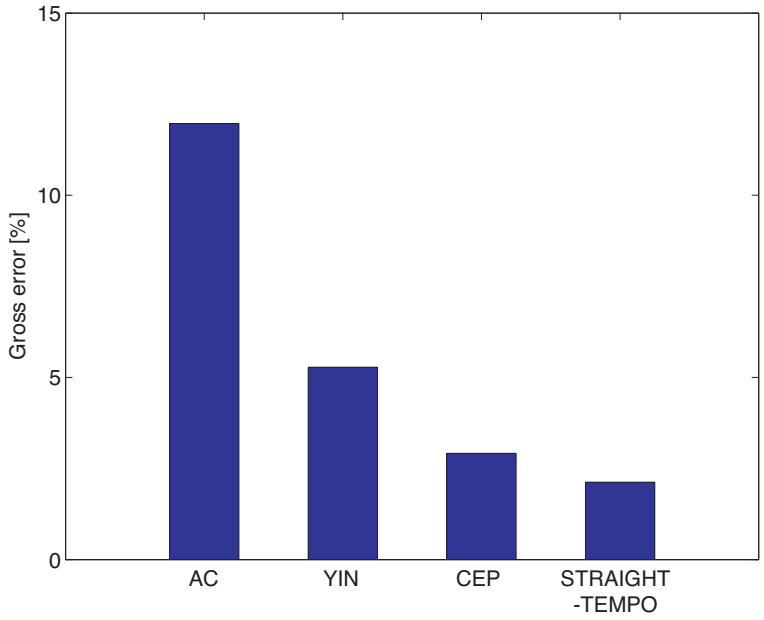

図 1.9: クリーンな音声に対する従来の基本周波数推定法の Gross error

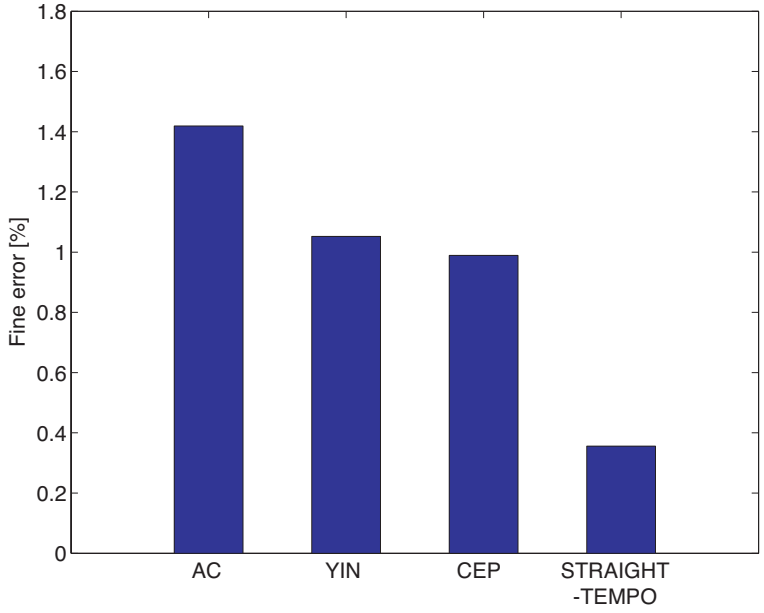

図 1.10: クリーンな音声に対する従来の基本周波数推定法の Fine error

従来の基本周波数推定法における音声波形と EGG 波形から得られた基本周波数による Gross error を図 1.9 に示す。これより音声波形と EGG 波形それぞれから得られた基本周波数において、推定区間の差が最も少ないのは STRAIGHT-TEMPO であることがわかる。次に、Fine error を図 1.10 に示す。この結果から、±20%未満の推定誤差で推定できた区間のうち、推定誤差が最も小さい手法もまた STRAIGHT-TEMPO であることがわかる。従って、4 つの基本周波数推定法において、STRAIGHT-TEMPO が最も推定精度が高いといえる。

- 考察

比較 1 の結果から、他の推定法との推定値の差が最も小さくなるのは STRAIGHT-TEMPO であることがわかり、STRAIGHT-TEMPO は他の手法と比較して偏った値を推定していないことが確かめられた。また、比較 2 の結果では、音声波形と EGG 波形から得られた基本周波数の差が最も小さい推定法は STRAIGHT-TEMPO であった。以上の結果から、STRAIGHT-TEMPO によって推定された基本周波数が真の基本周波数に最も近いと考えられる。

よって、以後の実験において EGG 波形から STRAIGHT-TEMPO によって抽出された基本周波数を正解基本周波数として評価に用いることとする。

耐雑音性能及び推定精度の検証

次に従来の基本周波数推定法の耐雑音性能及び推定精度の検証を行う。

音声データは前述の正解基本周波数の抽出と同じものを用い、雑音として次の 2 種類を用いた。

白色雑音 : 全周波数帯域に等しい雑音エネルギーが存在する。

ピンク雑音 : 全周波数帯域に雑音エネルギーがあるが、低周波数から高周波数へ向けて傾斜したパワースペクトルをもつ。パワースペクトルは高域へ向けて 1 オクターブあたり 3 dB 減少しており、低域にエネルギーが偏っている。

この 2 種類の雑音は従来から耐雑音性能の評価に用いられてきたものである。

雑音は信号対雑音比 (Signal to Noise Ratio; SNR) が 0–10 dB になるように音声データに加えた。評価尺度も前節と同じく Gross error と Fine error を用いる。前

述の通り、正解基本周波数はEGG波形からSTRAIGHT-TEMPOによって抽出された基本周波数とする。雑音付加音声に対しては、Gross errorは雑音に対する頑健性を示し、Fine errorは雑音環境における精度を示している。

- 白色雑音に対する結果

白色雑音付加音声に対する従来の基本周波数推定法のGross errorを図1.11に、File errorを図1.12に示す。

STRAIGHT-TEMPOはクリーンな音声に対しては高精度な基本周波数推定が可能であるが、雑音の影響を受けやすく雑音量が増大するにつれてGross errorが大幅に増加する。よって、瞬時周波数が雑音の影響を受けやすい特徴量であることがわかる。自己相関法とケプストラム法のGross errorはほぼ同じ値であり、雑音量が増加してもGross errorの増加は比較的少ないため、雑音に対して頑健であるといえる。また、自己相関法は時間領域の特徴を利用しており、ケプストラム法は周波数領域の特徴を利用していることから、時間領域・周波数領域のどちらの特徴も頑健な基本周波数推定には有用であることがわかる。

Fine errorをみると、STRAIGHT-TEMPOは誤差が±20%未満の区間であれば、クリーンな音声に対する場合と同じくほぼ正解に近い基本周波数を推定できている。しかし、自己相関法やケプストラム法は雑音の増大とともにFine errorが増加してしまうため、雑音が存在するときはその推定値は大まかな値であると考えなければならない。

- ピンク雑音に対する結果

ピンク雑音付加音声に対する従来の基本周波数推定法のGross errorを図1.13に、File errorを図1.14に示す。

ピンク雑音は、低域のエネルギーが大きく、高域のエネルギーが小さいという音声のスペクトルに似たエネルギー分布となっている。そのため、全体的に白色雑音よりも強く雑音の影響を受けているが、ピンク雑音の場合も白色雑音の場合とほぼ同様の傾向が見られる。すなわち、STRAIGHT-TEMPOは雑音の影響を受けやすいが推定できる区間では高精度であり、自己相関法やケプストラム法は雑音に対して頑健ではあるが雑音の増加によって推定精度が低下してしまう。

以上の結果から、

1. 雑音を含まないクリーンな音声からは高精度の基本周波数を抽出することが

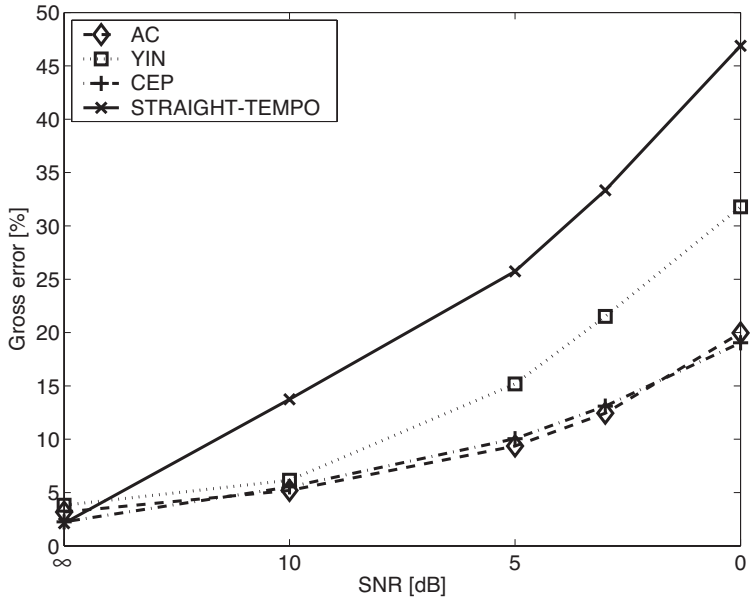

図 1.11: 白色雑音付加音声に対する従来の基本周波数推定法の Gross error

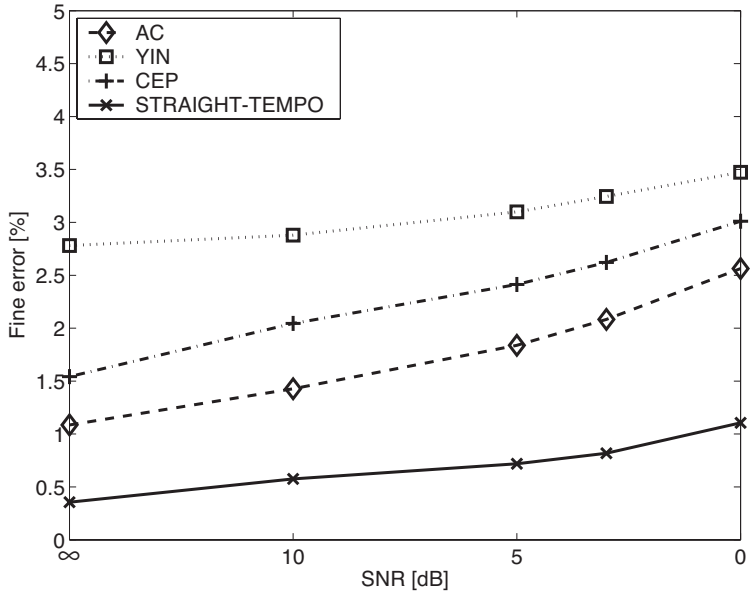

図 1.12: 白色雑音付加音声に対する従来の基本周波数推定法の Fine error

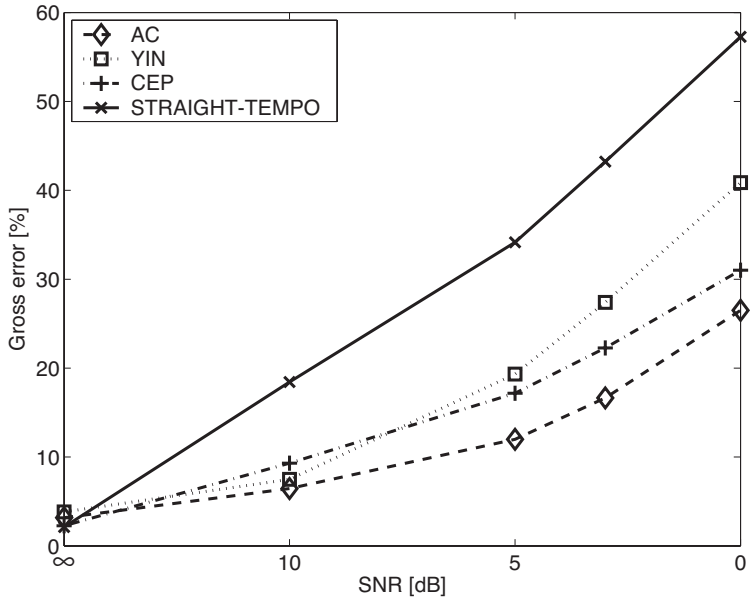

図 1.13: ピンク雑音付加音声に対する従来の基本周波数推定法の Gross error

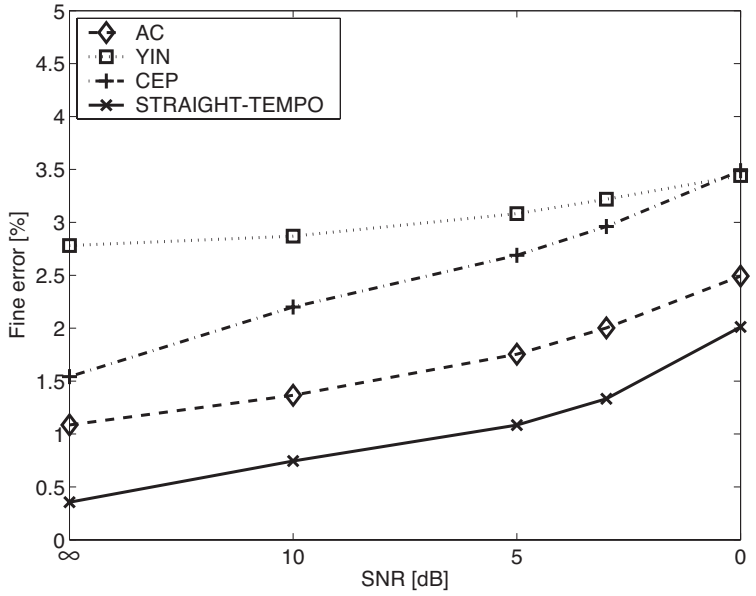

図 1.14: ピンク雑音付加音声に対する従来の基本周波数推定法の Fine error

できるが、雑音の影響を受けやすく、雑音を含む音声に対しては推定精度が急激に低下する (STRAIGHT-TEMPO)。
これは、高精度な推定を行う手法は基本周波数のわずかな変化への対応を考慮しているため、雑音の影響を受けて起こる基本周波数情報の変化にも追随してしまうためであると考えられる。

2. 大まかな基本周波数を推定できる程度には雑音に対して頑健であるが、雑音の増加とともに推定精度が低下していく (自己相関法, ケプストラム法) 。
これは、これらの手法が基本周波数情報のわずかな変化を重視するのではなく、基本周波数情報に歪みがあっても基本周波数に近い値を抽出することを重視しているためであると考えられる。

このように、雑音によって歪んだ音声から高精度の基本周波数を得ることは現有の方法では困難である。

1.4 ヒトのピッチ知覚

ヒトは雑音が存在する環境でも、目的の音の高さ (ピッチ) を知覚することができる。本節では、ヒトが音の高さをどのような処理により知覚しているのかについて説明する。

ヒトの聴覚器官の構造を図 1.15 に示す。音は聴覚器官において、外耳、中耳、内耳を経て中枢へと伝達される。外耳では音の情報は空気の圧力振動として外耳道を伝わった後に鼓膜の振動となる。中耳では鼓膜の振動は槌骨、砧骨、鐙骨の順番に耳小骨連鎖を介して伝わっていく。内耳で音受容器として働くのがラセン状の蝸牛である。蝸牛内部は前庭階、鼓室階、中央階の 3 つの管からなり、リンパ液で満たされている。中央階は前庭階とはライスネル膜で、鼓室階とは基底膜で仕切られている。前庭階には前庭窓、鼓室階には鼓室窓と呼ばれる窓があり、前庭窓には鐙骨の底板が密着している。鐙骨の振動で前庭窓が振動すると、前庭階内のリンパ液も振動し、リンパ液の振動は進行波となって蝸牛の基部から先端部に向かって伝播する。それにつれて中央階が振動し、基底膜も基部から先端部へと振動する。基底膜の上には、有毛細胞が並ぶコルチ器がある。基底膜の振動が有毛細胞の毛を曲げ、その曲がり具合に応じて有毛細胞内に受容器電位が発生す

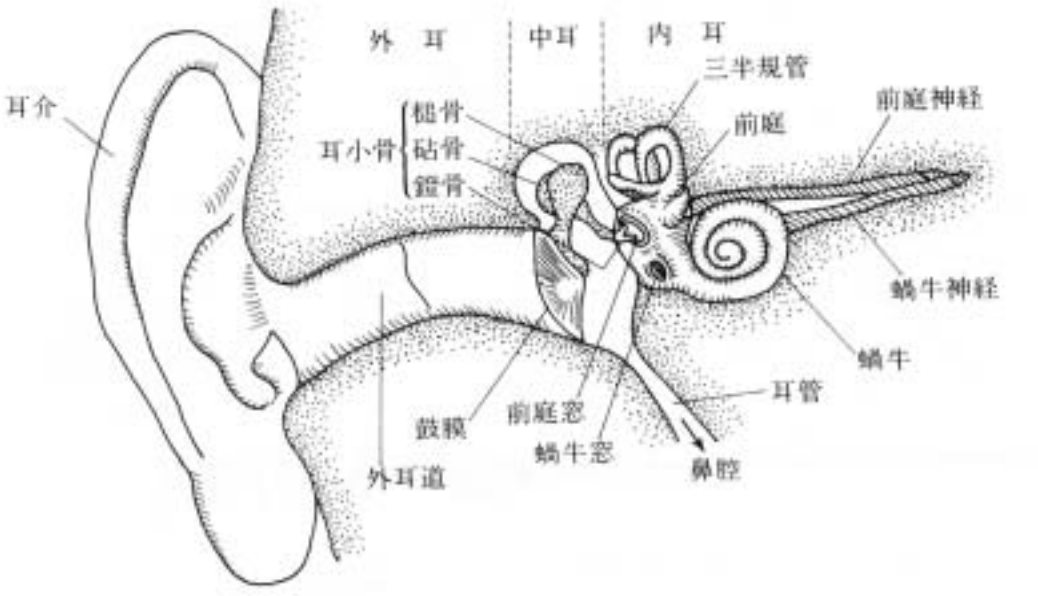

図 1.15: ヒトの聴覚器官の構造 (古井 [78])

る [79]。すなわち、ここで音響信号は聴神経発火による電気信号へと変えられ聴覚中枢へと送られる。

基底膜に沿って進む進行波の振動振幅は最初増大していき、その後急激に減少するが、その振動パターン中のピークの位置は音の周波数により変化する。これは、前庭窓が位置する基底膜の基部では膜の幅は狭くぶ厚く、基底膜の先端部は幅が広く薄いという基底膜の構造によるものである。進行波による基底膜の振動パターンを図 1.16 に示す。低い周波数の音の場合は振動パターンは基底膜全体に広がり、膜の終端の前で振幅が最大に達する。高い周波数の音の場合は前庭窓に近い位置で基底膜が最大に変化し、膜のそれ以外の部分はあまり動かない。すなわち、基底膜上で最大の変位が生じる位置が周波数によって異なることから、蝸牛において音の周波数分析がされていると考えられる [3, 33]。

ヒトのピッチ知覚に関しては、「場所」説と「時間」説の 2 つの理論が考えられている。「場所」説とは、上記のように異なる周波数では基底膜振動のピーク位置が異なることから、基底膜上のどの位置に由来する聴神経の発火頻度が高いかによって音の高さを知覚するという考えである。すなわち、「場所」説においてはピッチ知覚のために周波数情報が用いられている。一方、「時間」説とは、音の高さはそ

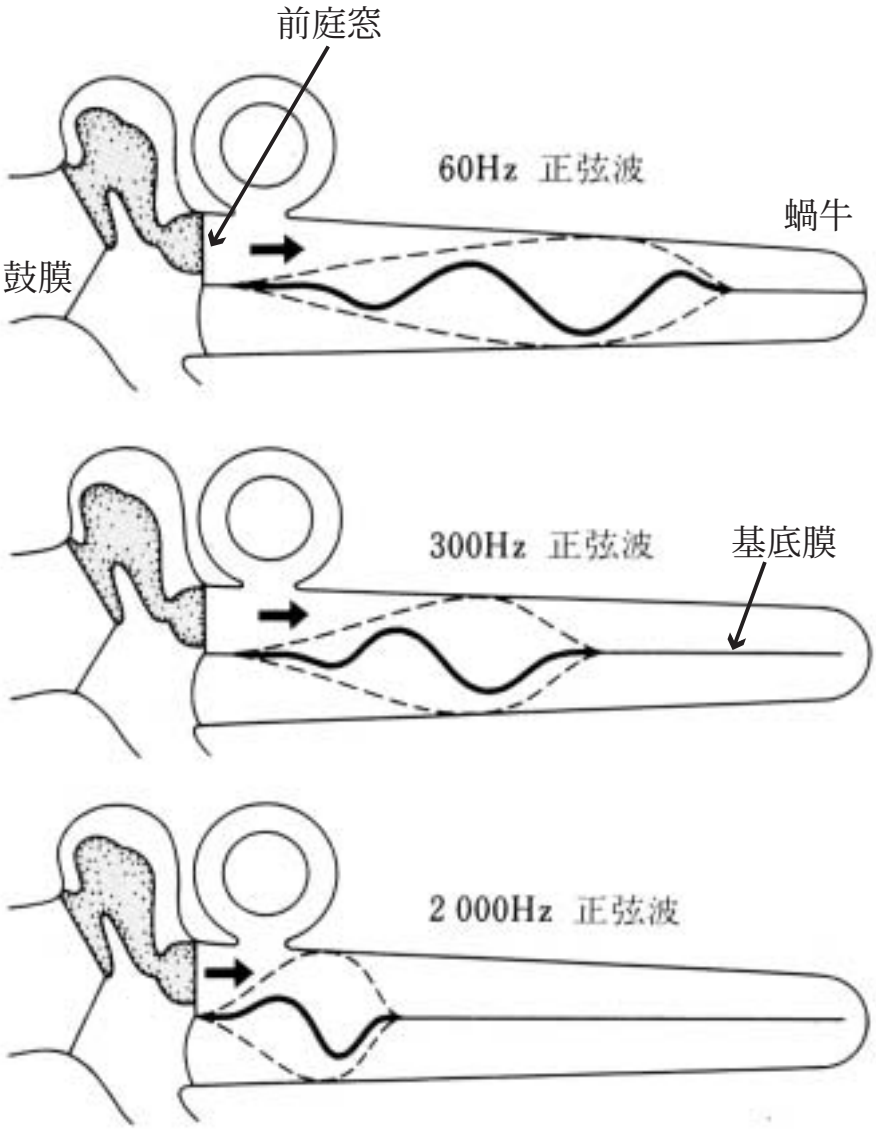

図 1.16: 基底膜上の進行波 (甘利ら [4])

の音によってひき起こされた聴神経発火の時間パターンに関係しているという考えである。聴神経は基底膜振動の特定位相に同期してインパルスを出すため、神経インパルスの間隔は音声波形の周期の整数倍に近いものとなる (これは位相固定 *phase locking* と呼ばれる)。すなわち、「時間」説ではピッチ知覚に時間情報が用いられている。長年の間、ヒトが「場所」説と「時間」説のどちらに基づいて音の高さを知覚しているかについて議論されてきた。「場所」説だけでは、複合音の知覚を説明することが困難である。これは、複合音によって生じる基底膜上の振動パターンには単一の最大点がなく、多くの極大点が分布しており、その中の最大値が基本周波数と一致しているとは限らないためである。また、「時間」説だけでは、位相固定が約 5 kHz 以上の周波数では見られないことから、非常に高い周波数の音に対するピッチ知覚を説明できない。そのため近年では、ヒトのピッチ知覚の説明には「場所」説と「時間」説の両方が取り入れられている。この周波数情報と時間情報の相対的な重要さは周波数範囲や音の種類によって異なるが、中枢においてはそれぞれから得られたピッチ感覚は同一のものとして扱われていると考えられている [33, 80]。つまり、ヒトの聴覚機構は時間情報と周波数情報の両方を利用することにより音の高さを識別している。

1.5 実環境雑音の特性

実環境にはさまざまな音源が存在しており、各音源から発生した音が雑音として目的音声に付加されて観測されることになる。このとき、各音源から発生した音はその発生原因によって、それぞれ異なる特性を持つ周波数スペクトルで表される。

例えば、自動車が走行しているときに車内で観測される雑音は、自動車のエンジン音やロードノイズなどにより発生しているため、図 1.17 に示すように低い周波数帯域に強いエネルギーを持つ。また、ガススプレーの噴射音に関しては図 1.18 のように低い周波数帯域にはほとんどエネルギーが無く、高い周波数帯域に強いエネルギーが存在する。このように実環境雑音では、白色雑音のような全周波数帯域にエネルギーが存在する人工的な雑音とは異なり、特定の周波数帯域に偏ったエネルギーを持つような特性になっている。

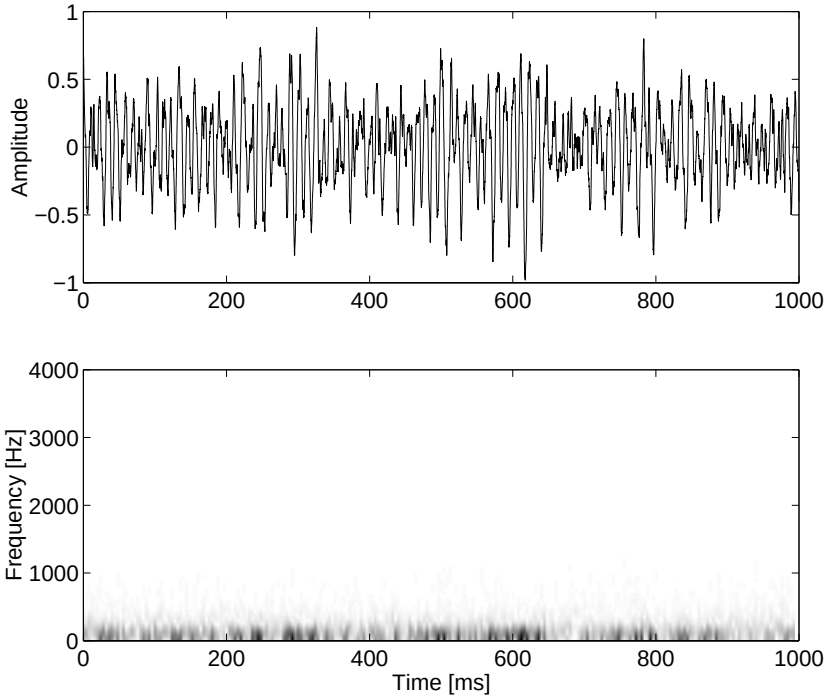

図 1.17: 走行自動車内雑音 [81] の信号波形とスペクトログラム

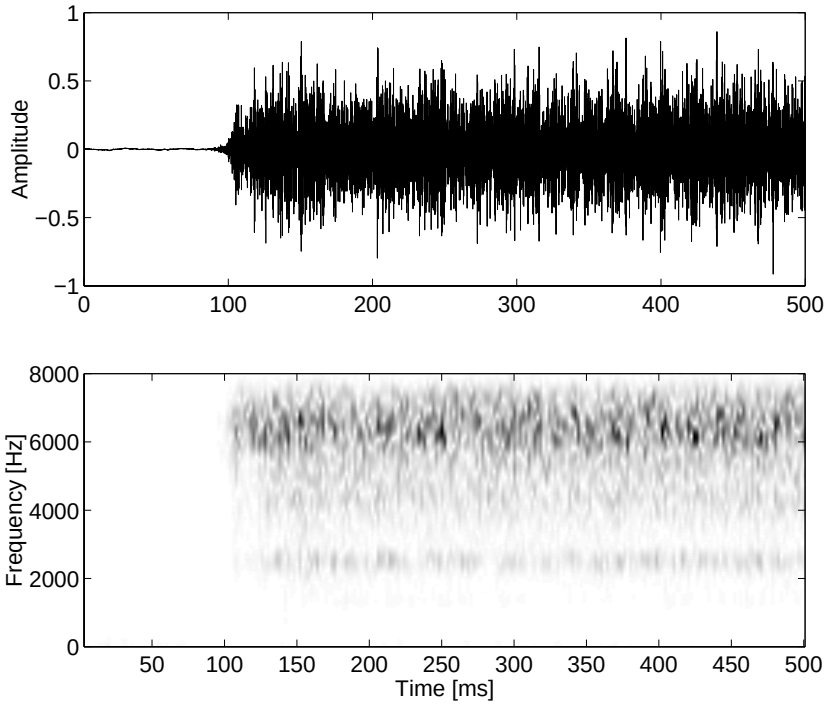

図 1.18: ガススプレーの噴射音 [82] の信号波形とスペクトログラム

実環境では雑音によって目的音声が歪むが、その歪み方は実環境雑音の特性により変わる。雑音のエネルギーが低域に偏っている場合は、音声の時間波形の概形が大きく変わる。すなわち、音声波形のピーク位置の間隔がその音声の基本周期とは異なるようになり、音声波形の周期性が歪むことになる。そのため、自己相関法のような波形の周期性検出を利用する方法は、低域に強いエネルギーを持つ雑音の影響を強く受ける。一方、雑音のエネルギーが高域に偏っている場合は、周波数軸上に櫛状に並んだ音声スペクトルが雑音の影響を受けて、明瞭な調波性が見られなくなる。すなわち、調波成分のピーク間隔から基本周波数を読み取ることが困難になり、ケプストラム法のような音声の調波成分の検出を利用する方法は雑音の影響を強く受けることになる。このように、実環境では雑音の特性によって有効に利用できる音声の特徴や周波数帯域が異なる。

また、実環境には暗騒音のように常に背後に存在する雑音だけではなく、ドアの開閉音や目覚まし時計のベル音のように突然に発生する雑音も存在する。そのため、時々刻々と変化する実環境雑音の特性をあらかじめ予測することは容易ではないことから、時間情報と周波数情報のどちらが明瞭に基本周波数情報を表わしているのか、または、どの周波数帯域が相対的に雑音エネルギーが小さくなっているかを予測することは困難である。

1.6 実環境雑音に対する基本周波数推定法の構築

実環境に存在する雑音は、これまでに基本周波数抽出法の評価に用いられることが多かった白色雑音やピンク雑音のような人為的な定常雑音と異なり、時間-周波数領域にエネルギーが偏在している。この実環境雑音の特性を考慮して、雑音環境における基本周波数推定法をどのように構築するかについて考える。

1.6.1 時間-周波数領域の基本周波数情報

従来の基本周波数推定法は、1.3 節で述べたように音声の周期性の特徴 (時間情報) と調波性の特徴 (周波数情報) のどちらかを利用して、基本周波数抽出を行なっている。しかし、1.5 節で述べたように実環境雑音は特定の時間-周波数帯域にエ

ネルギーが偏っているため、雑音の種類によって有効に利用できる特徴が異なる。つまり、実環境では、時間情報と周波数情報のどちらを利用することで頑健な基本周波数推定ができるのかは、その時々の雑音の特性によって異なることになり、時間情報と周波数情報の一方のみを利用している従来の基本周波数推定法では常に頑健な推定を行うことは望めない。

ヒトが実環境において雑音が存在する状況でも音の高さを知覚できることを考慮すると、様々な特性の雑音に対応するためには、1.4 節で述べたヒトの聴覚機構のように、時間情報と周波数情報の両方を利用することが適していると考えられる。

そこで、基本周波数を表す時間情報と周波数情報をどのように利用するのがよいかを考える。例えば、時間情報を利用する手法のうち単に音声波形の自己相関処理により周期間隔を測る方法では、低周波数帯域に雑音のエネルギーが強く存在するときに、大きな推定誤差となることが予想される、また、周波数情報を利用する手法のうち単に周波数軸上での調波成分間隔を測るだけでは、調波成分の存在する周波数帯域の一部に強い雑音エネルギーが存在する場合に、推定誤差が大きくなる。時間情報と周波数情報からこのような手法により得られた推定基本周波数をただまとめるだけでは、それぞれの手法の推定誤差が正しい値の抽出を妨害してしまう。

ここで、実環境雑音には音声のエネルギーに対して雑音エネルギーが相対的に小さい周波数帯域が存在することを考慮し、その帯域に現れる基本周波数情報 (時間情報と周波数情報) を利用することにより推定誤差を小さくする方策を検討する。各帯域における時間情報と周波数情報を表す方法として次のものが考えられる。音声を時間-周波数解析することにより、音声のエネルギー分布を時間-周波数平面上に表すことができる。そのエネルギー分布には基本周波数に対応した特徴が現れるが、時間-周波数解析をする際の分析窓長によって現れる特徴が変わる。時間-周波数解析において分析窓長が短いときには、各周波数帯域に時間方向に基本周期に対応した振幅変動が現れる。すなわち、これは基本周波数に関する情報が含まれた時間情報である。一方、分析窓長が長いときには、各時刻に周波数方向に基本周波数に対応したスペクトルの振幅変動が現れる。これは基本周波数の周波数情報である。このとき、雑音エネルギーが相対的に小さい帯域の時間情報と周波数情報を利用することにより、実環境雑音が存在していても基本周波数が

推定できると考えられる。

ただし、従来の高精度な手法によってクリーンな音声から推定した基本周波数と比較すると、雑音を含む音声から推定した基本周波数は精度の点で劣る。すなわち、音声の時間-周波数解析による時間情報と周波数情報を利用した手法であっても、雑音によって基本周波数情報が歪んでいる以上は雑音付加音声から推定できるのはある程度の精度の基本周波数であり、クリーンな音声から得られる基本周波数と同精度の推定は困難である。

1.6.2 基本周波数を利用した雑音抑圧

実環境雑音下において基本周波数を推定するもうひとつの手法として、雑音を抑圧した後に基本周波数推定を行うことを考える。雑音抑圧音声に対してならば、雑音には弱くとも高精度な基本周波数推定法を用いることにより、雑音中の音声の基本周波数を高い精度で抽出できる。

雑音抑圧は、受音点の個数により用いられる手法が異なることが多く、一般に受音点の数が多いほど雑音抑圧は容易となる。マイクロホンアレイを利用し複数の受音点から得られた信号から目的の音声を抽出する手法は数多く提案されているが、実環境における応用を考えると、マイクロホンの位置や距離を考慮する必要のない、受音点がひとつの場合の雑音抑圧が理想的である。そこで、ここでは受音点がひとつであると仮定して雑音抑圧を行うこととする。

受音点がひとつの場合における雑音抑圧の手法のなかには、スペクトルサブトラクションやカルマンフィルタによる方法があるが、これらは雑音の特性を仮定しなければならない [83, 84, 85]。そのために、これらの雑音抑圧法では、扱う雑音を限定するか、無音区間における観測から雑音の特性を推測することが行われる。しかし、実環境雑音では暗騒音として常に存在する雑音だけではなく、突然発生する雑音も考えられる。音声の途中で発生した雑音に対しては、その雑音の特性を推測することは困難である。すなわち、実環境雑音に対する雑音抑圧では雑音の特性を仮定した雑音抑圧手法は適していない。

そこで、雑音の特性を仮定しない雑音抑圧法として、櫛形フィルタにより音声の調波成分のみを残す手法を取り上げる。この手法では周波数軸上における音声の

調波成分の位置の情報が必要となる。ここで、1.6.1 節で述べたように音声の時間-周波数解析から得られる時間情報と周波数情報を利用することにより、雑音環境下でもある程度の精度ならば基本周波数抽出が期待できることを利用する。音声の調波成分は基本周波数とその整数倍の周波数に存在することから、ある程度の基本周波数が求められれば音声の調波成分の位置情報が得られる。すなわち、1.6.1 節で述べた手法によって得られた基本周波数に合わせて櫛形フィルタを構築し、櫛形フィルタの帯域幅を適切に設定できれば、音声の調波成分を取り出すような雑音抑圧が可能である。

このようにして得られた雑音抑圧音声を用いることにより、実雑音環境下においても、雑音のないクリーンな音声に対する推定と同様の高精度な基本周波数推定を行うことができると考えられる。

1.7 本研究の目的

今後、音声認識や音声分析合成、音源分離などの音声情報処理の研究は実環境における応用へと進められていくが、実環境下には周囲に常になんらかの雑音が存在すると考えなければならない。雑音を含む音声から精度の高い基本周波数を抽出することができれば、音声情報処理の様々な分野において応用可能となる。例えば、雑音環境下の音声認識は雑音の影響により認識精度が大幅に低下するが、基本周波数を句や単語の区分化に利用したり、基本周波数の概形 (パターン) を認識器に入力する特徴量として利用することで、認識精度を向上させることができる。接話マイクロホンを用いて雑音の影響を減らす方法も考えられるが、実環境での運用では常にそのような理想的な状況で音声を観測することが期待できない以上、雑音を含む音声に対しても音声の特徴を抽出できなければならない。また、音声分析合成システムにおいては、接話マイクロホンのようなわずらわしさのない、マイクロホンとの距離を意識する必要のない状況へと応用することができる。音源分離システムに対しては推定基本周波数を音源の違いの手がかりとして与えることができ、聴覚情景解析の研究に貢献できる。

1.3 節で述べたように、基本周波数の抽出法に関する研究は古くから行われており、これまでに多数の基本周波数推定法が提案されている。これは、今もなお決

定的な推定法が存在せず、目的により手法を選ばざるをえないことを意味している [16, 17]。特に、実環境での基本周波数抽出を考慮すると、1.2 節で挙げた抽出を困難にする点のうち、雑音に対する対応が重要となる。雑音環境下において従来の基本周波数推定法には

- 基本周波数の時間的に滑かな変化を高い精度で抽出する目的の手法は、雑音を含まないクリーンな音声からは高精度の基本周波数を抽出することができるが、雑音の影響を受けやすくなり雑音を含む音声に対しては推定精度が急激に低下する

- 雑音に対して頑健な手法は、ある程度の雑音が存在しても大まかな基本周波数を推定することができるが、推定精度は雑音の増加とともに低下してしまい、クリーンな音声から推定した基本周波数ほど精度は良くない

という傾向が見られる。すなわち、これまでに提案されたほとんどの基本周波数推定法は雑音環境における抽出には欠点を持ち、「雑音に対して頑健」と「高精度」の両方を満たす推定法は存在しない。また、実環境に存在する雑音は人工的に作成した白色雑音やピンク雑音とは異なり、特定の周波数帯域に強いエネルギーをもつものが多い。例えば、走行自動車内雑音では音声の基本周波数近傍の低周波数帯域に強いエネルギーが存在し、高周波数帯域にはほとんど存在しない。実環境下における基本周波数推定に関しては、このような特性をもつ実環境雑音への対応も考慮しなければならない。

本論文では、雑音環境においても頑健で高精度な基本周波数が推定可能な方法を構築することを目的とする。そこで、1.6 節で述べた実環境雑音に対応するための方策により基本周波数推定を行う。提案法は次のような処理により構築される。上述のように雑音を含む音声から高精度の基本周波数を抽出することは、雑音により基本周波数情報を持つ特徴量が歪んでしまうために困難である。しかし、雑音を含む音声からでも大まかな基本周波数を推定することはできる。そこで、提案法は最初に、雑音に頑健な基本周波数推定を行い、ある程度の精度の基本周波数を推定する。次に、ここで推定された基本周波数に櫛形フィルタの中心周波数を合わせることにより、音声の調波に合わせた櫛形フィルタを構成し雑音抑圧を行う。これで、雑音を含む音声を基本周波数を推定しやすい信号へと変えること

ができる。最後に、雑音抑圧された信号に対して、高精度な推定が可能な方法で最終的な基本周波数を推定する。

提案法は、最初に行う雑音に対して頑健な基本周波数推定においては、実環境に存在する雑音の特性に着目し、時間-周波数解析によって現れる瞬時振幅の周期性の特徴と調波性の特徴の両方を利用する。実環境に存在する雑音は白色雑音のような人為的な雑音とは異なり、エネルギーが特定の周波数帯域に集中しているため、雑音の影響の少ない周波数帯域が存在する。そこで、時間-周波数平面上の様々な時間周波数帯域から基本周波数の情報を得ることを考える。1.3 節で従来の基本周波数推定法が時間領域に現れる周期性の特徴 (時間情報) か周波数領域に現れる調波性の特徴 (周波数情報) のどちらかを利用していることを述べたが、本手法では基本周波数の情報を含む特徴を時間と周波数の両方の面から様々な帯域において分析することで、より雑音に対して頑健な推定が可能となる。

また、次の櫛形フィルタによる雑音抑圧においては音声の調波成分を誤って除去しないようにしなければならない。そのため、初期推定基本周波数にはある程度の精度が求められる。また、仮に誤った推定値が与えられても調波成分を除去しないために、初期推定基本周波数の確からしさに応じて通過帯域幅を変えることができる櫛形フィルタを構成する。提案法の最終段階である高精度な基本周波数推定には瞬時周波数を利用した手法を用いる。

なお、本論文では有声/無声判定は考慮しないものとする。雑音を含む音声の有声無声判定もまた困難を伴う研究課題である。

1.8 本論文の構成

本論文は全 6 章により構成されている。各章の相互関係を図 1.19 に示し、各章の概要を以下に述べる。

第 1 章では、音声の基本周波数における特徴及び既存の基本周波数推定法について概説する。また、ヒトが雑音中の音の高さを知覚できることから、ヒトのピッチ知覚の仕組みについて説明する。最後に実環境雑音の特性とそれに対応した基本周波数推定の方策について考察し、本論文の目的を述べる。

第 2 章では、提案法の全体の構成について述べ、提案法の処理の流れと第 3 章、

第 4 章との対応について説明する。

第 3 章では、提案法の初期推定部である雑音に対して頑健な基本周波数推定として、時間方向に現れる周期性の特徴 (時間情報) と周波数方向に現れる調波性の特徴 (周波数情報) を利用する推定法を構築する。また、周期性と調波性を両方とも用いることの有効性について検討する。

第 4 章では、提案法の雑音抑圧部である櫛形フィルタによる雑音抑圧について述べる。初期推定部で得られた初期推定基本周波数を用いて雑音抑圧を行うための帯域幅可変櫛形フィルタを定式化する。また、初期推定基本周波数の確からしさと櫛形フィルタの通過帯域幅の関連を調べ、提案法のおける雑音抑圧に適した実装を行う。

第 5 章では、様々な雑音を含む音声に対する提案法の有効性について検証する。また 1.3.3 節で取り上げた従来の基本周波数推定法との性能比較を行う。

第 6 章では、本論文で得られた結果を要約し、今後の課題について述べる。

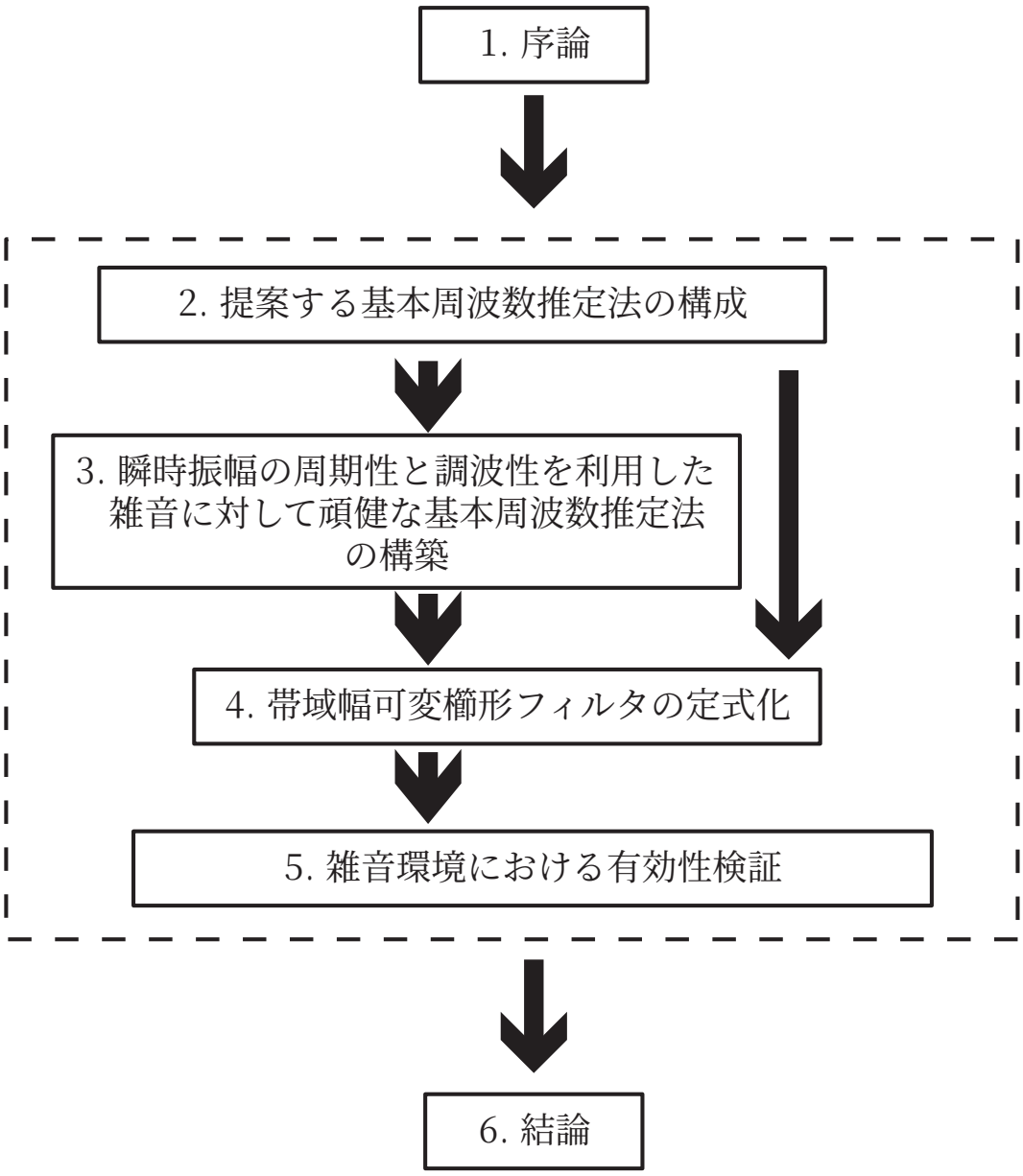

図 1.19: 各章の相互関係

第 2章

雑音環境における基本周波数推定法の構成

2.1 はじめに

1.3.3 節で述べたように従来の基本周波数推定法は

1. クリーン音声に対して高い精度で基本周波数を推定できるが、雑音の影響を受けやすく、雑音を含む音声に対しては推定精度が大幅に低下し基本周波数の大まかな値を推定することもできない
2. 雑音を含む音声に対して頑健であり、雑音が存在していても基本周波数の大まかな値は推定できるが、推定できた音声区間における精度はクリーン音声に対する基本周波数と比較すると低い

のどちらかの傾向にある。実環境における音声情報処理の応用を考えると基本周波数推定法には、クリーンな音声に対して高精度な基本周波数が抽出できるだけではなく、雑音が存在する環境においてもクリーンな音声から得られる基本周波数と同精度の推定が可能である手法が望まれる。また、実環境雑音はエネルギーが特定の周波数帯域に集中しているため、雑音の影響の少ない周波数帯域が存在するという点で、白色雑音やピンク雑音のような人工的な雑音とは異なる。

本研究では、上記の問題点及び実環境雑音の特性に対応し

- 雑音に対して頑健
- 雑音が存在する環境でも高精度

の両方を満たす基本周波数推定法を提案する。1.6 節で述べた方策に基づいた提案法として、瞬時振幅に現れる周期性と調波性という時間情報及び周波数情報を利用した雑音に対して頑健な基本周波数推定と、それによって得られた基本周波数を利用した櫛形フィルタによる雑音抑圧、そして雑音の影響を受けやすい瞬時周波数を用いるがクリーン音声に対しては高精度な基本周波数推定を利用することにより、雑音に対して頑健でかつ高精度な推定が可能な基本周波数推定法を構築する。

本章では提案法の概要について述べ、提案法の詳細は第 3 章及び第 4 章で説明する。

2.2 雑音環境における基本周波数推定法

提案法のブロック図を図 2.1 に示す。1.6 節の考察に基づき、提案法は次の 3 つのブロックから構成される。

初期推定部 雑音を含む音声に対して、瞬時振幅の時間-周波数表現に現れる周期性と調波性を利用した、雑音に対して頑健な基本周波数推定を行う。

雑音抑圧部 初期推定部において得られた基本周波数に中心周波数を合わせた帯域幅可変櫛形フィルタによって、雑音を含む音声の雑音抑圧を行う。

最終推定部 雑音抑圧された音声に対して、瞬時周波数の不動点を利用した高精度な基本周波数推定を行う。

雑音を含む音声から高精度の基本周波数を直接抽出することは困難である。しかし、高精度の基本周波数ではなくある程度の大まかな基本周波数の値ならば、雑音を含む音声からも求めることは可能である。そこで、提案法では、まず、初期推定部において高精度ではなくともある程度の精度の基本周波数を推定する。ここで推定された基本周波数を利用し、次の雑音抑圧部で櫛形フィルタによる雑音抑圧を行う。この雑音抑圧によって音声の調波成分が残され、調波成分とは異なる周波数の雑音成分が除去される。雑音除去音声に対して提案法の最終推定部で、雑音の影響は受けやすくとも高精度の推定が可能な手法で基本周波数推定を行なう。

2.2.1 雑音に頑健な基本周波数推定 (初期推定部)

Unoki and Akagi は音源分離モデルを構築するために、雑音を付加された単母音から瞬時振幅の comb filtering を用いて基本周波数を推定している [14]。この手法は定 Q フィルタバンクを用いて音声を分析し、瞬時振幅の時間-周波数表現で調波構造を表示することによって実現されている。このとき、雑音が存在していても単母音の瞬時振幅から基本周波数を推定することができる。すなわち、瞬時振幅の時間-周波数表現は雑音を含む音声に対しても基本周波数の情報を表すことができる特徴量であるといえる。しかし、Unoki らの手法では、基本周波数が大きく変化する連続音声に対しては、正しい基本周波数を推定することができなかった。

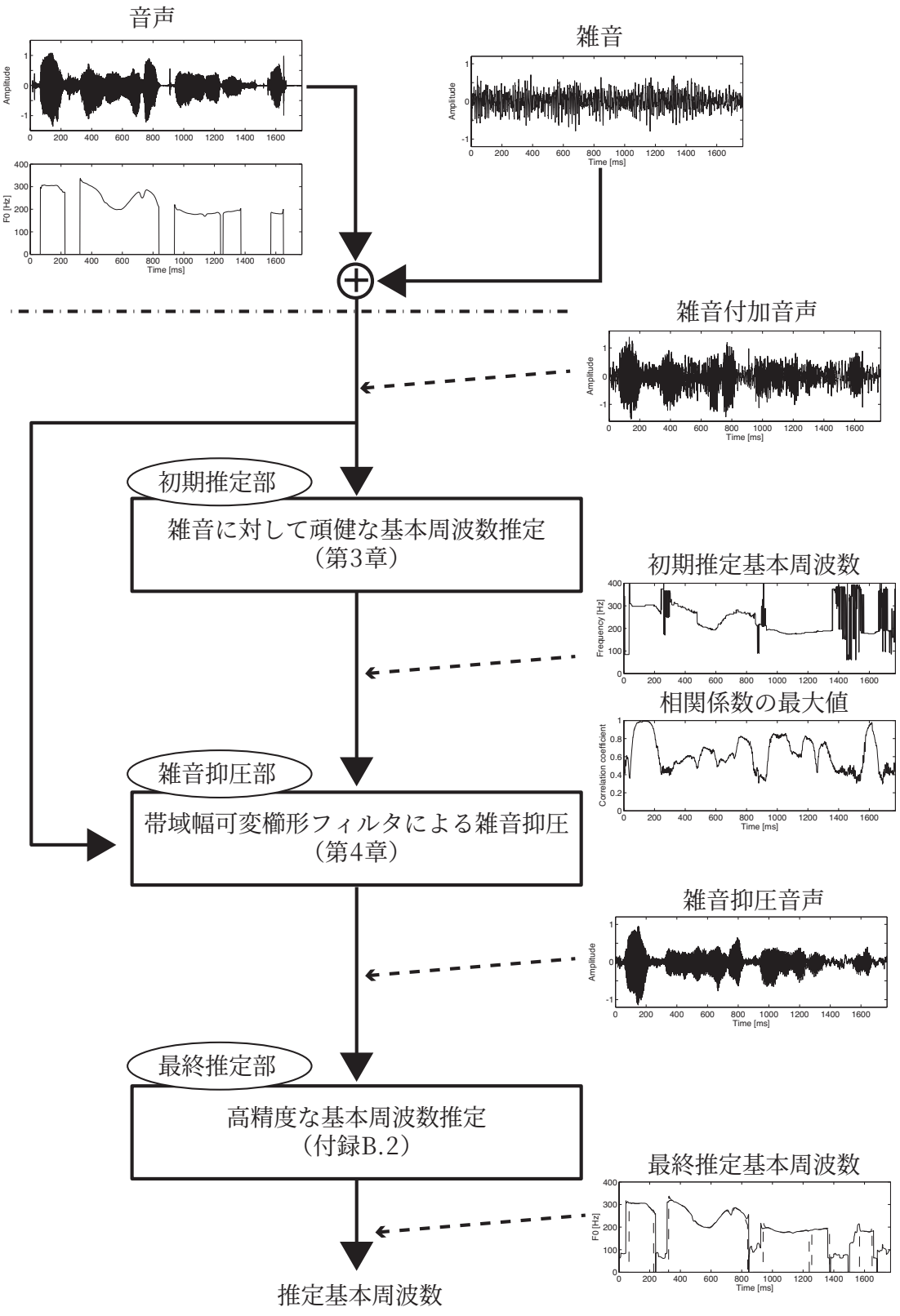

図 2.1: 提案法のブロック図

ところで、瞬時振幅の時間-周波数表現にはUnokiらが用いた基本周波数に対応した周波数方向の振幅変動以外に、時間方向に基本周期に対応する振幅変動が現れる。また、これまで提案されてきた基本周波数推定法のほとんどには、時間領域の特徴量と周波数領域の特徴量のどちらかが用いられており、そのどちらも基本周波数推定に有用であることは1.3.3節及び1.4節に示した通りである。

ここで、実環境に存在する雑音の特徴を考えると、そのエネルギーは白色雑音のような人為的な雑音とは異なり特定の周波数帯域に集中して存在しており、雑音のエネルギーがほとんどない周波数帯域が存在することが推測される。そこで、提案法の初期推定部においては、時間領域の特徴量と周波数領域の特徴量、すなわち時間情報と周波数情報の両方を利用することにより雑音に対する頑健性を向上させることを考え、瞬時振幅の時間-周波数表現に現れる時間方向の振幅変動(周期性)と周波数方向の振幅変動(調波性)から基本周波数を推定する。特に雑音エネルギーが特定の帯域に偏った雑音であるならば、この時間-周波数平面上の様々な帯域において、時間及び周波数の両領域に対して基本周波数情報を集めることによって、雑音の影響の少ない帯域からより正しい基本周波数情報を得ることが期待できる。

まず、入力信号を時間分解能の高い定Qフィルタバンクと周波数分解能の高い定帯域フィルタバンクの2種類のフィルタバンクを用いて解析し、時間-周波数領域で瞬時振幅を表す。入力信号が音声のような調波複合音であれば、定Qフィルタバンクによる分析で得られた瞬時振幅には時間方向に基本周期に対応した振幅変動(周期性)が現れる。一方、通過帯域幅の狭い定帯域フィルタバンクによる分析で得られた瞬時振幅には周波数方向に基本周波数に対応した振幅変動(調波性)が現れる。周期性を示す瞬時振幅に対してはフィルタバンクのチャネルごとに自己相関処理を行い、調波性を示す瞬時振幅に対しては周波数帯域を変えて複数の自己相関処理を行なう。それぞれの自己相関係数は基本周期(基本周波数)に対応する点でピークを持つと考えられるが、雑音を含む音声では雑音の影響によりピークが正しい基本周波数の位置からずれる。そこで、基本周波数が時間的に滑らかに変化することを利用し、短区間内の自己相関係数について、各係数ごとの和を求めて平均化し、基本周波数を示すピークを強調させるとともに、雑音による歪みを抑圧する。周期性と調波性から得られた2つの統合自己相関係数はDempster

の結合規則でさらに統合され、最も高い係数が示す周波数が基本周波数として抽出される。

雑音の影響により基本周波数情報が歪んでしまっていることから、この初期推定部では「高精度」な推定は望めないが、正しい基本周波数に近い値を推定するという「雑音に対する頑健性」は期待できる。

初期推定部の詳細については第 3 章で述べる。

2.2.2 基本周期を利用した雑音抑圧 (雑音抑圧部)

つぎに、雑音抑圧を行い雑音の影響を小さくした音声を得る。一般に、実環境雑音は時々刻々と変化しており、雑音の特性を予測することが困難である。雑音抑圧法にはスペクトルサブトラクションやカルマンフィルタを利用した方法などがあるが、雑音の特性を仮定したこれらの方法では実環境雑音には適していない。そこで、初期推定部では実環境雑音下でも正解に近い基本周波数を推定できることを利用し、初期推定部で推定された基本周波数に中心周波数を合わせて音声の調波成分を残すよう設計された櫛形フィルタにより雑音抑圧を行う。この櫛形フィルタによる雑音抑圧は時間領域での雑音波形のキャンセレーションによって実現されており、容易に櫛形フィルタの帯域幅を変更できるという点で優れている。ここで問題となるのが、初期推定部の基本周波数が必ずしも正しい基本周波数ではなく、ある程度の誤差が含まれており、雑音による影響が非常に大きい場合にはまったく誤った値となりうる点である。もし、初期推定基本周波数が大きく誤っているときに、推定値に忠実に櫛形フィルタを構成し推定基本周波数とその高調周波数に対応する周波数帯域以外の成分を除去してしまうと、音声の調波成分を誤って除去してしまうことにつながる。従って、音声の調波成分を誤って除去しないような櫛形フィルタを設計する必要がある。

そこで、提案法で用いる櫛形フィルタがひとつのパラメータの値により容易に帯域幅を変えることができる構成になっていることを利用し、初期推定部による基本周波数の確からしさによって通過帯域幅を決定することとする。すなわち、初期推定基本周波数の誤差が小さいと判断できれば、なるべく音声の調波成分以外の成分を除去するように櫛形フィルタの通過帯域幅を狭くする。また、初期推定

基本周波数の誤差が大きいと判断できれば、通過帯域幅を広くして音声の調波成分を除去しないようにする。ここで、初期推定基本周波数の誤差を推測できる指標が必要となるが、初期推定部で用いられる相関係数の値を推定基本周波数の確からしさとして利用する。

雑音抑圧部の詳細については第 4 章で述べる。

2.2.3 高精度な基本周波数推定 (最終推定部)

最終推定部では、雑音抑圧された音声に対して基本周波数推定を行う。雑音抑圧部からの出力は音声波形であるので、既存のどの基本周波数推定法もこの最終推定部として用いることができる。しかし、この最終推定部における推定精度が提案法全体の推定精度となるため、高精度な推定が可能な手法を用いることが望ましい。1.3.3 節に示すように瞬時周波数の不動点を利用した基本周波数推定 (STRAIGHT-TEMPO)[62] は高精度な基本周波数を推定することができることから、STRAIGHT-TEMPO を最終推定部として用いる。瞬時周波数は雑音の影響を受けやすい特徴量であるが、前段の雑音抑圧部で調波成分付近以外の雑音を抑圧できるため、入力信号内の雑音の存在にかかわらず高い精度を保ったままの基本周波数推定が可能となる。

2.3 まとめ

本章では、本研究で提案する雑音環境における基本周波数推定法の概要について述べた。雑音を含む音声から直接、精度の高い基本周波数を推定することは困難であるという考えから、雑音を含む音声に対して雑音抑圧を行い、雑音抑圧音声から精度の高い基本周波数を推定するという方策をとる。提案法は、初期推定部、雑音抑圧部、最終推定部の 3 つのブロックからなり、初期推定部では雑音に対して頑健な基本周波数推定を行う。初期推定部で推定された基本周波数を用いて、雑音抑圧部で櫛形フィルタにより雑音抑圧音声を得る。この雑音抑圧音声を用いて最終推定部で高精度な基本周波数推定を行う。このような方策により、雑音を含む音声に対しても頑健で高精度な基本周波数推定が可能となる。

初期推定部である雑音に頑健な基本周波数推定については第 3 章で述べ、雑音抑圧部で用いる帯域幅可変櫛形フィルタについては第 4 章で述べる。

第 3 章

瞬時振幅の周期性・調波性を利用した基本周波数推定

3.1 はじめに

本章では、提案法の初期推定部である雑音に対して頑健な基本周波数推定法について述べる。

この初期推定部で推定された基本周波数が次の雑音抑圧部における櫛形フィルタの中心周波数として用いられることを考慮すると、雑音環境においてもある程度の精度で推定できる頑健性が必要とされる。また、櫛形フィルタの帯域幅を決定するためには、初期推定基本周波数が信頼できる値であるかどうかの情報も必要となる。

音声の瞬時振幅の時間-周波数表現は雑音環境においても基本周波数情報を表しうる。そこで、時間方向に基本周期に対応するピーク間隔をもつ瞬時振幅として現れる周期性の特徴と、周波数方向に基本周波数に対応するピーク間隔をもつ瞬時振幅として現れる調波性の特徴を利用し、時間情報と周波数情報の両方を用いて雑音を含む音声から基本周波数を推定する。

3.2 瞬時振幅の周期性・調波性を利用した基本周波数推定の概要

図 3.1 に提案法の初期推定部である瞬時振幅の周期性と調波性を基にした基本周波数推定の概要について示す。

まず、入力信号を時間分解能の高い定 Q フィルタバンクと周波数分解能の高い定帯域フィルタバンクの 2 種類のフィルタバンクを用いて解析し、時間-周波数領域で瞬時振幅を表す。入力信号が音声のような調波複合音であれば、定 Q フィルタバンクによる分析で得られた瞬時振幅には時間方向に基本周期に対応した振幅変動（周期性）が現われる。一方、通過帯域幅の狭い定帯域フィルタバンクによる分析で得られた瞬時振幅には周波数方向に基本周波数に対応した振幅変動（調波性）が現われる。この周期性・調波性の特徴を両方とも利用することにより、推定の信頼性を向上させることができる。

周期性を示す瞬時振幅に対してはフィルタバンクのチャネルごとに自己相関処理を行い、調波性を示す瞬時振幅に対しては周波数帯域を変えて複数の自己相関

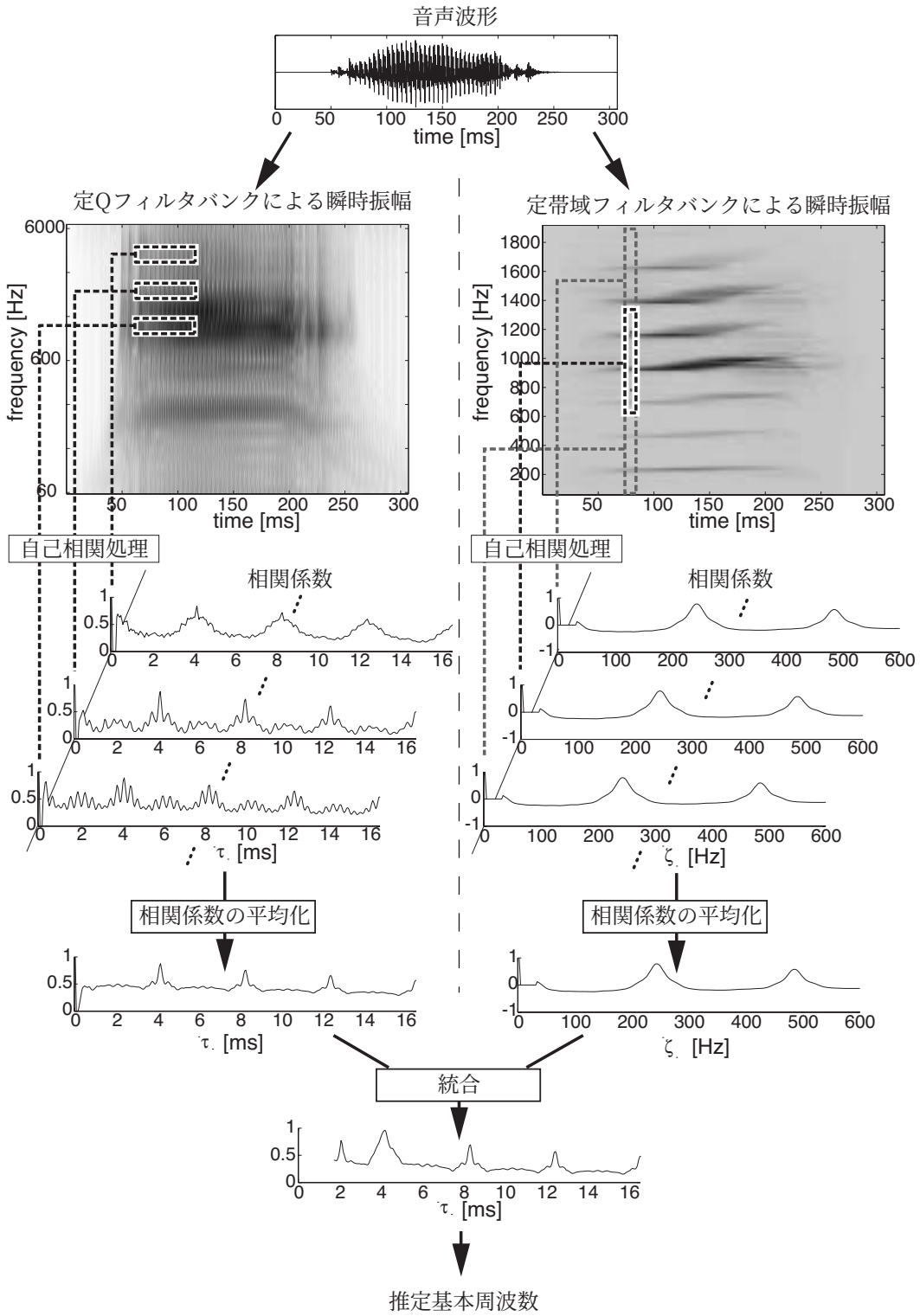

図 3.1: 瞬時振幅の周期性と調波性を基にした基本周波数推定の概要

処理を行なう。それぞれの自己相関係数は基本周期 (基本周波数) に対応する点でピークを持つと考えられるが、雑音を含む音声では雑音の影響によりピークが正しい基本周波数の位置からずれる。そこで、基本周波数が時間的に滑らかに変化することから、短区間内の自己相関係数を各係数ごとの和を求めて統合し、基本周波数を示すピークを強調させる。

周期性と調波性から得られた 2 つの統合自己相関係数は Dempster の結合規則 [86, 87] でさらに統合され、最も高い係数が示す周波数が初期推定基本周波数として抽出される。

3.3 節で瞬時振幅の時間-周波数表現についての詳細を説明する。また、3.4 節で瞬時振幅の自己相関処理について述べ、3.5 節で周期性の特徴、調波性の特徴それぞれから得られた相関係数に対して、Dempster の結合規則によって統合する方法について述べる。

3.3 瞬時振幅の時間-周波数表現

音声信号の時間-周波数表現は次のように得られる。

入力信号 $x(t)$ は帯域通過フィルタ群 $h_k(t)$ からなるフィルタバンクで分析される。このフィルタバンクの出力 $y_k(t)$ は

$$y_k(t) = x(t) * h_k(t), \tag{3.1}$$

で与えられる。ここで、k はフィルタのチャネル番号であり、演算子 $*$ は畳み込みである。フィルタバンク出力から、解析信号 $\tilde{y}_k(t)$ は

$$\tilde{y}_k(t) = \mathcal{F}^{-1}\left[2Y_k(\omega)U(\omega)\right], \tag{3.2}$$

$$U(\omega) = \begin{cases} 1, & \omega > 0 \\ 1/2, & \omega = 0 \\ 0, & \omega < 0 \end{cases} \tag{3.3}$$

として得られる。$Y_k(\omega)$ は $y_k(t)$ のフーリエスペクトルであり、$\mathcal{F}^{-1}[\cdot]$ は逆フーリエ変換である。このとき、瞬時振幅 $s_k(t)$ と瞬時周波数 $\lambda_k(t)$ は式 (3.4) と式 (3.5) で得られる。

$$s_k(t) = |\tilde{y}_k(t)|, \tag{3.4}$$

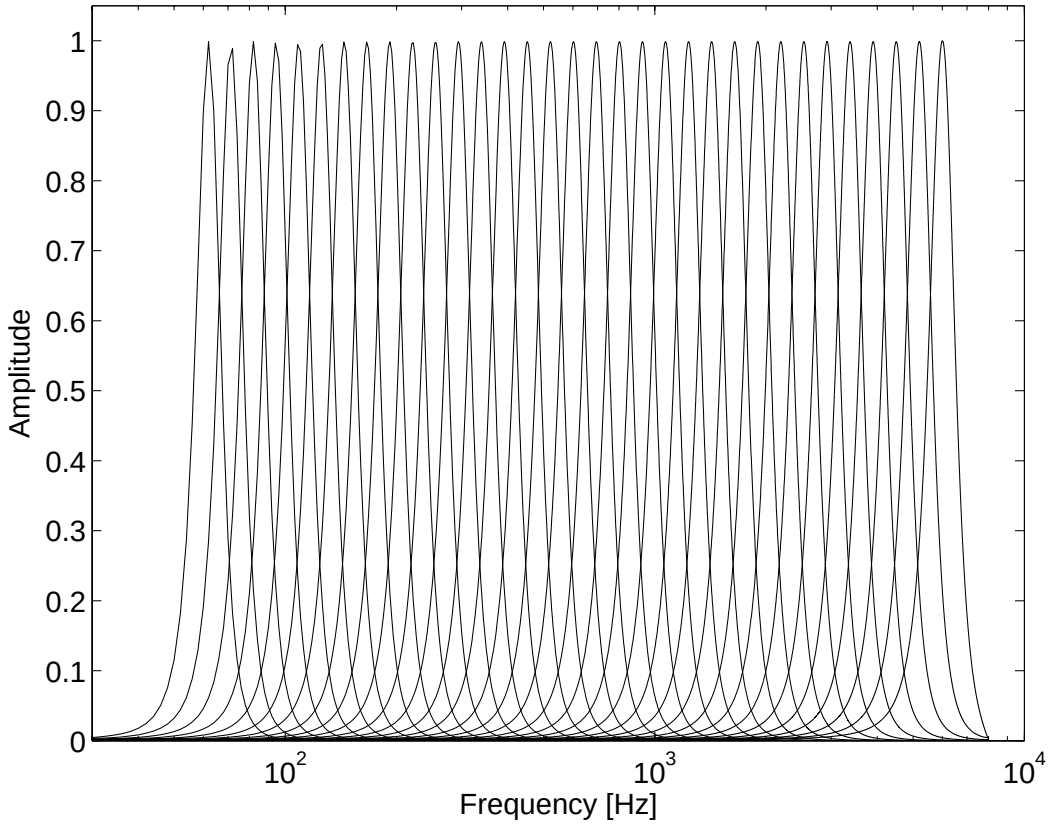

図 3.2: ガンマトーンフィルタによる定 Q フィルタバンクの振幅特性

$$\lambda_k(t) \quad = \quad \frac{\partial}{\partial t} \arg \tilde{y}_k(t). \tag{3.5}$$

瞬時振幅は瞬時のエネルギーの時間変化を表している。

この時間-周波数解析において、提案法は、定 Q フィルタバンクと、狭い帯域幅をもつ定帯域フィルタバンクの 2 種類のフィルタバンクを用いる。これらのフィルタバンクは、式 (3.6) で表されるガンマトーンフィルタ [88] を用いて構築される。

$$gt(t) = At^{N-1} \exp(-2\pi b_f ERB(f_c)t) \cos(2\pi f_c t), \qquad (t > 0) \tag{3.6}$$

ここで、N と b_f は帯域幅に関連するパラメータであり、f_c は中心周波数、$ERB(f_c)$ は Equivalent Rectangular Bandwidth[89] である。

定 Q フィルタバンクの構築においては、[90] に記されているフィルタバンク同様、

$$h_k(t) \quad = \quad \frac{1}{\sqrt{a}} gt\left(\frac{t}{a}\right), \tag{3.7}$$

$$a \quad = \quad 10^{(2/K)(k-1)-1}, \tag{3.8}$$

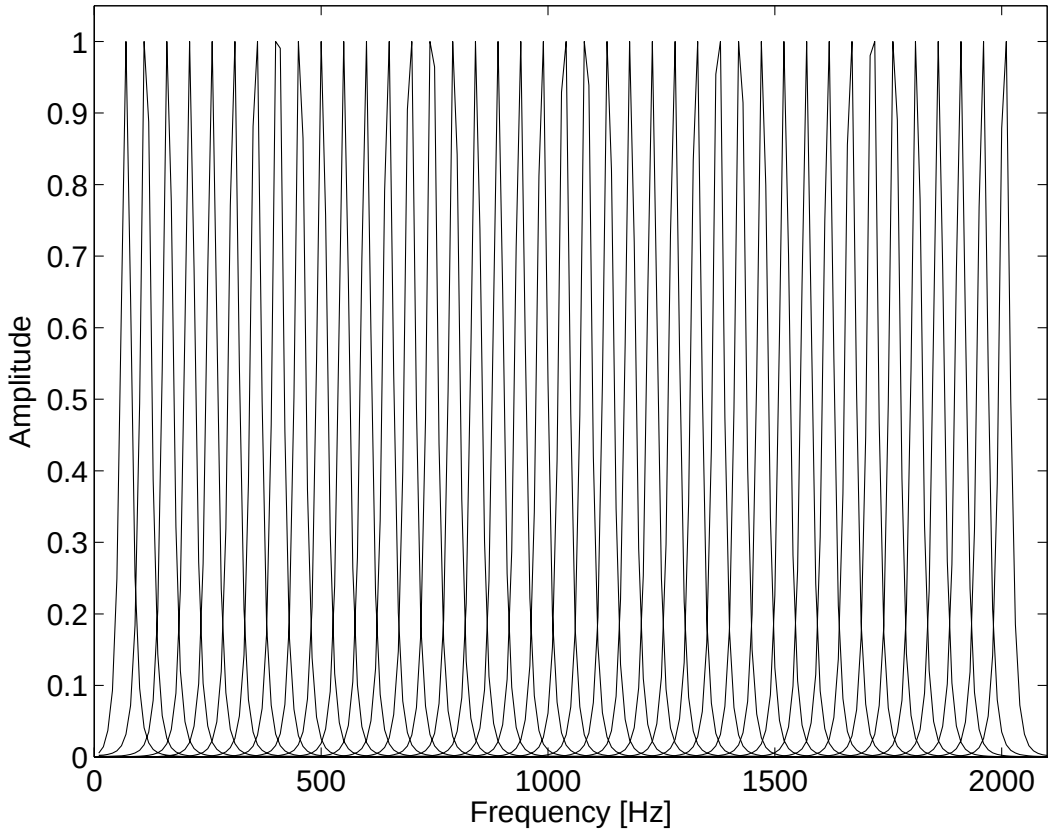

図 3.3: ガンマトーンフィルタによる定帯域フィルタバンクの振幅特性 (チャネル番号 1, 10, 20, ..., 400 のフィルタを表示)

とした。ここで、$N = 4$, $f_c = 600$ Hz, $b_f = 1$, $1 \leq k \leq K + 1$, $K = 64$ である。このとき、式 (3.1) はウェーブレット変換となる。フィルタの中心周波数は 60–6000 Hz の範囲となり、チャネル数は 33 である。この定 Q フィルタバンクを図 3.2 に示す。

定帯域フィルタバンクの構築においては、$h_k(t)$ は

$$h_k(t) = At^{N-1} \exp(-2\pi b_f t) \cos(2\pi f_k t), \qquad (t > 0) \tag{3.9}$$

$$f_k = 60 + 5(k-1) \text{ Hz}, \tag{3.10}$$

で与えられる。ここで、$N = 4$, $b_f = 20$ Hz, $1 \leq k \leq 400$ である。フィルタの中心周波数は 60–2000 Hz で 400 チャネルが等間隔となるように定めた。2 kHz 以上を考慮しないのは、音声の調波構造が 2 kHz 以上では乱れてしまい、明瞭な調波構造が現れないことが多いためである。この定帯域フィルタバンクを図 3.3 に示す。

$s_k(t)$ が定 Q フィルタバンクによる時間-周波数平面におけるチャネル番号によって配置されると、周期的な振動が時間方向の瞬時振幅に現れる。その振動のピーク間隔は基本周波数の逆数である基本周期と等しい。この振幅変動を瞬時振幅の周期性と呼ぶこととする。定 Q フィルタバンクは高周波数領域で高い時間分解能をもつため、周期性は特に高周波数領域に顕著に現れる。

同様に定帯域フィルタバンクを用いたとき、瞬時振幅の振動は周波数方向に現れる。この振幅変動を瞬時振幅の調波性と呼ぶこととする。調波性によるピークの間隔は基本周波数に等しい。音声の調波は高周波数領域では基本周波数の整数倍からそれることが多く、調波性は特に低周波数帯域で明瞭である。定帯域フィルタバンクから得られた瞬時振幅は高速フーリエ変換 (FFT) によって得ることも可能であるが、予備実験によって式 (3.9) を用いることで FFT を用いるよりも明瞭な調波性がみられることがわかった。

図 3.4 の中段は定 Q フィルタバンクによって算出された男性話者の母音/a/(図 3.4 上段) の瞬時振幅の時間-周波数表現である。図 3.4 の下段は、中段の点線で示されたフィルタバンクのチャネル番号 31 における瞬時振幅である。周期性が時間方向に明瞭に現れていることがわかる。

図 3.5 の中段は、図 3.4 と同じ母音に対して、定帯域フィルタバンクによって得られた瞬時振幅の時間-周波数表示である。図 3.5 の下段は、中段の点線で示された時刻 80 ms における対数瞬時振幅である。調波性が周波数方向に明瞭に現れていることがわかる。

3.4 周期性・調波性に対する自己相関処理

3.3 節で得られた周期性を表わす瞬時振幅の時間-周波数表現において、各チャネルごとに時間方向に自己相関処理を行い自己相関係数を求める。時刻 t、チャネル番号 k における遅れ τ の自己相関関数 $a_{k,t}(\tau)$ は

$$a_{k,t}(\tau) = \sum_{i=t}^{t+w_t} \bar{s}_k(i)\bar{s}_k(i+\tau), \tag{3.11}$$

$$\bar{s}_k(t) = C[s_k(t)] \tag{3.12}$$

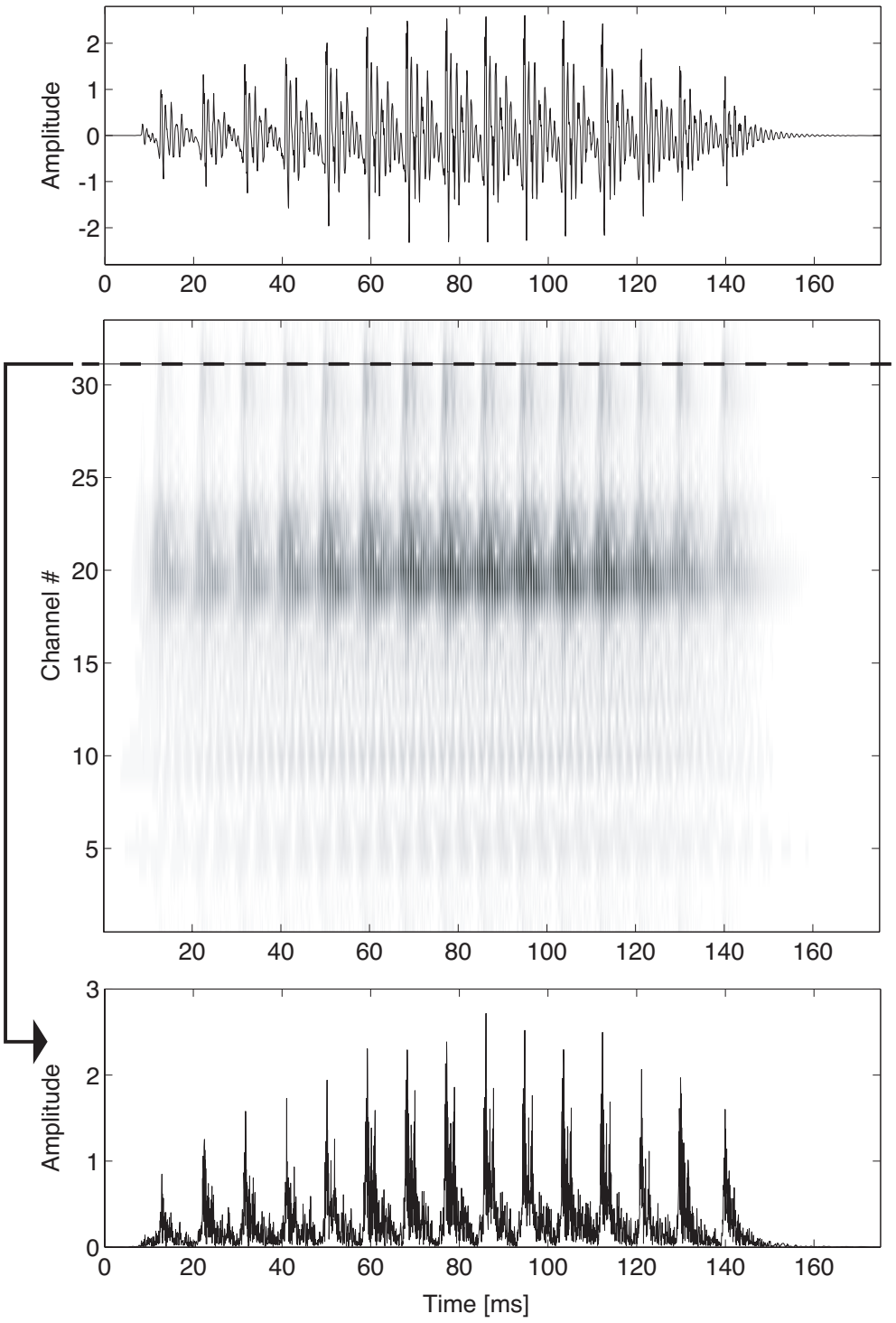

図 3.4: 時間-周波数領域における男声母音/a/の瞬時振幅に現れる周期性: (上) 音声波形/a/、(中) 定 Q フィルタバンクによる瞬時振幅の時間-周波数表現、(下) チャネル番号における瞬時振幅

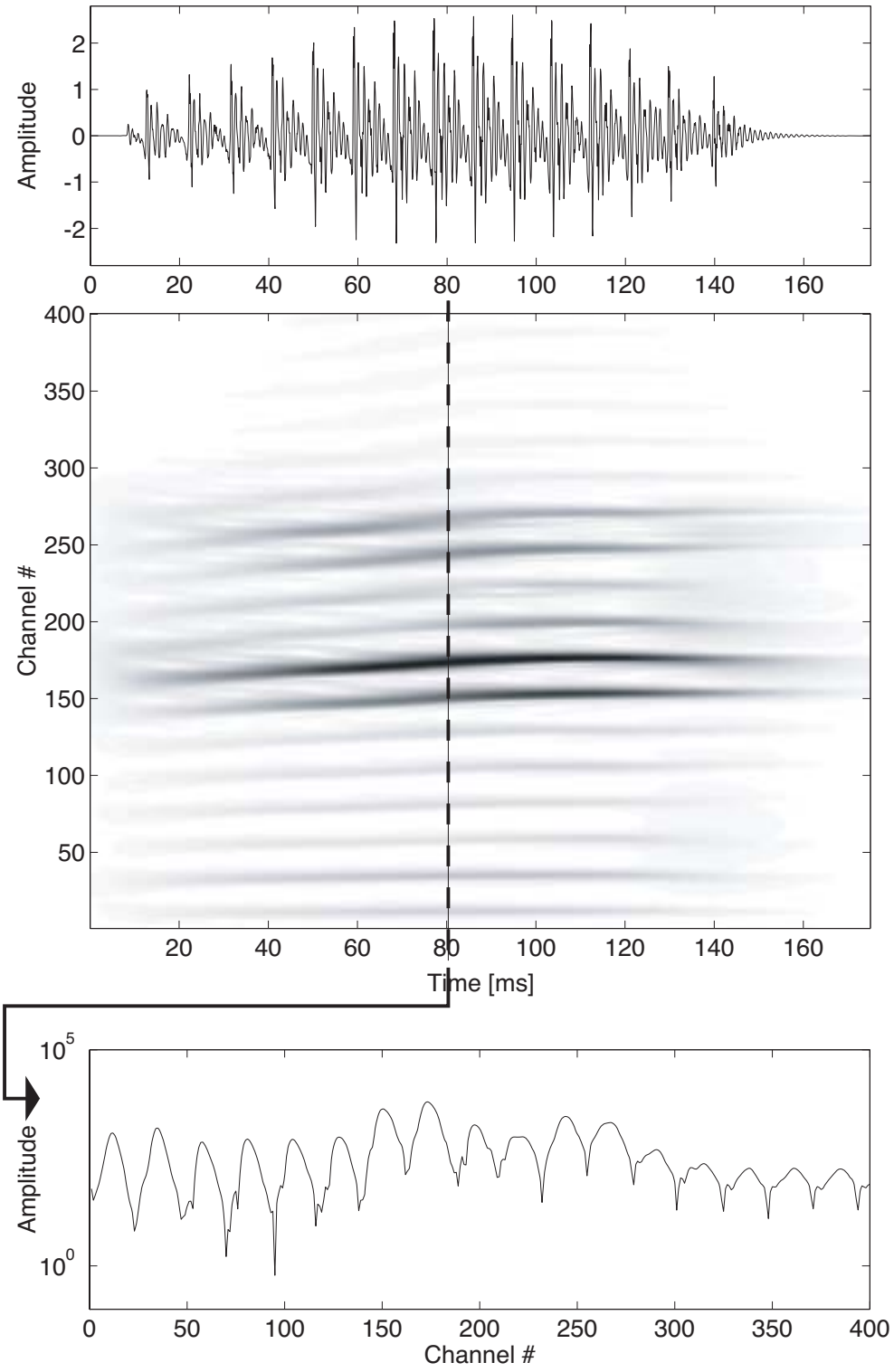

図 3.5: 時間-周波数領域における男声母音/a/の瞬時振幅に現れる調波性: (上) 音声波形/a/、(中) 定帯域フィルタバンクによる瞬時振幅の時間-周波数表現、(下) 時刻 80 ms における瞬時振幅

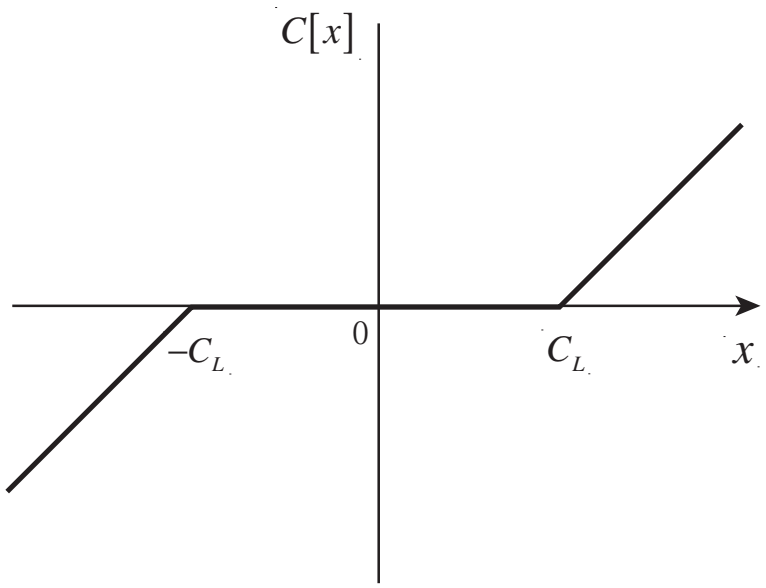

図 3.6: センタクリッピング関数

で与えられる。w_t は 35 ms とし、$\bar{s}_k(t)$ は平均振幅を利用して $s_k(t)$ をセンタクリッピングした信号である [24]。$C[\cdot]$ は図 3.6 に示すようなセンタクリッピング関数であり、センタクリッピングすることにより周期性がより強調されることになる。ここで得られた自己相関係数を全チャネル (全周波数帯域) で平均化することにより雑音の影響を抑圧する。

$$a_t(\tau) = \frac{1}{N_p} \sum_{k=1}^{N_p} a_{k,t}(\tau) \tag{3.13}$$

ここで、N_p は定 Q フィルタバンクのチャネル数 ($N_p = 33$) である。この定 Q フィルタバンクによって展開された信号に対する各チャネルごとの自己相関処理とその統合処理は、同時発生音声のピッチ差検出処理を説明する Meddis の聴覚モデル [31] とほぼ等しい処理となっている。また、複数の自己相関処理によって得られた係数をまとめるという点では、Gold and Rabiner による並列処理法 [22] の基本的な構成にも近い。

同様に調波性を表わす瞬時振幅の時間-周波数表現から各周波数帯域ごとに周波数方向に自己相関処理を行い、時刻 t、周波数帯域 f の自己相関係数 $b_{f,t}(\zeta)$

$$b_{f,t}(\zeta) = \sum_{j=w_f}^{w_e} \tilde{s}_j(t)\tilde{s}_{j+\zeta}(t) \tag{3.14}$$

を求める。ここで、w_f, w_e は自己相関処理を行う周波数帯域のチャネル番号の最

小値と最大値であり、ζ は自己相関処理におけるチャネル番号差 (周波数差) を表わす。雑音の影響を抑圧するために $b_{t,f}(\zeta)$ を全周波数帯域で平均化すると、

$$b_t(\zeta) = \frac{1}{N_h} \sum_{f}^{N_h} b_{f,t}(\zeta) \tag{3.15}$$

ここで、N_h は周波数帯域の分割数である。

平均化された自己相関係数 $a_t(\tau), b_t(\zeta)$ は基本周期 (基本周波数) に対応する時間差 τ、周波数差 ζ でピークを示すようになる。男性音声に対する平均化された自己相関係数 $a_t(\tau)$ の時刻-時間差表示および $b_t(\zeta)$ の時刻-周波数差表示を図 3.7 に示す。また、女性音声に対する平均化された自己相関係数 $a_t(\tau)$ の時刻-時間差表示および $b_t(\zeta)$ の時刻-周波数差表示を図 3.8 に示す。どちらも基本周期に対応する τ および基本周波数に対応する ζ で明瞭なピークが現れていることがわかる。

3.5 自己相関係数の統合方法

3.4 節で得られた各時刻 t における相関係数を統合する。$a_t(\tau)$ は時間差、$b_t(\zeta)$ は周波数差の関数であるので、まず、$b_t(\zeta)$ を時間差の関数 $\tilde{b}_t(\tau)$ へと変換する。

自己相関係数の値を各時間差に対応する基本周波数の確からしさであるとする。ここで、$a_t(\tau)$ と $\tilde{b}_t(\tau)$ を統合する際に以下の点に注意しなければならない。まず、$a_t(\tau)$ と $\tilde{b}_t(\tau)$ はそれぞれ、時間領域における周期性と周波数領域における調波性という異なる情報から得られたものであり、その値に対して絶対的な比較を行うことは不適切であると考えられる。例えば、$a_t(\tau)$ における 0.9 という値と $\tilde{b}_t(\tau)$ における 0.9 という値が同程度の確からしさを示しているとはいえない。また、$a_t(\tau) = 1.0,\ \tilde{b}_t(\tau) = 0.01$ であるときに、τ が基本周期であるかどうかは Bayes 確率では $a_t(\tau) \times \tilde{b}_t(\tau) = 1.0 \times 0.01 = 0.01$ となる。つまり、$a_t(\tau) = 1.0$ であるならば τ は正しい基本周期を表わしていると考えられるにもかかわらず、$a_t(\tau)$ と $\tilde{b}_t(\tau)$ をまとめた結果では確からしさが低くなってしまう。さらに、$a_t(\tau)$ と $\tilde{b}_t(\tau)$ のそれぞれの値に対して加法性が成り立つとはいえない。例えば、$a_t(\tau)$=0.9 のとき、基本周期が τ であるといえる根拠が 0.9 という値で示せても、加法性から得られる $1.0 - a_t(\tau) = 0.1$ という値は基本周期が τ ではない根拠とはならない。雑音の影響により正しい基本周波数に対応している τ であっても相関係数値が低くなること

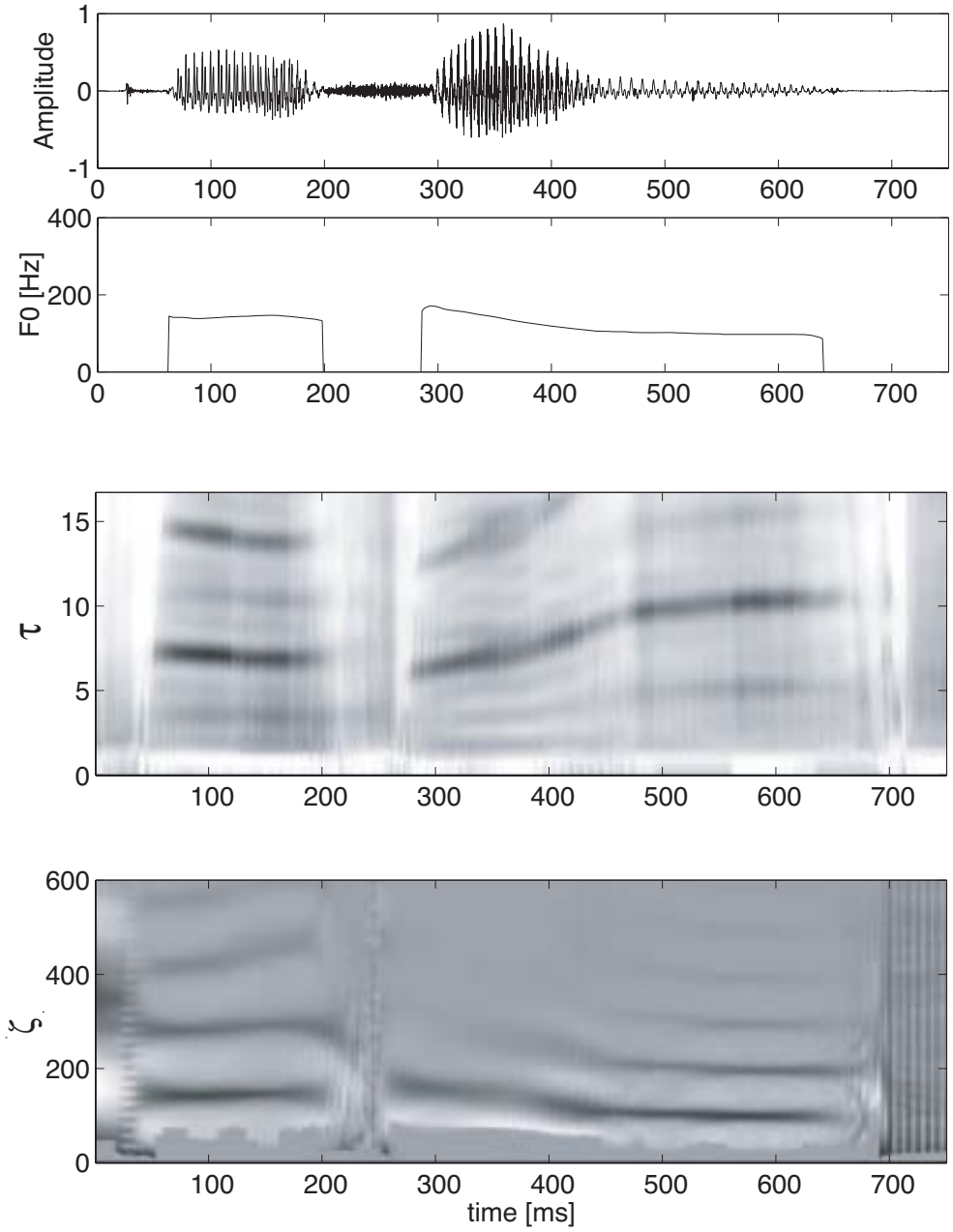

図 3.7: 男性音声に対する平均化された自己相関係数 $a_t(\tau), b_t(\zeta)$: (上から、男性音声波形、その基本周波数、$a_t(\tau)$ の時刻-時間差表示、$b_t(\zeta)$ の時刻-周波数差表示)

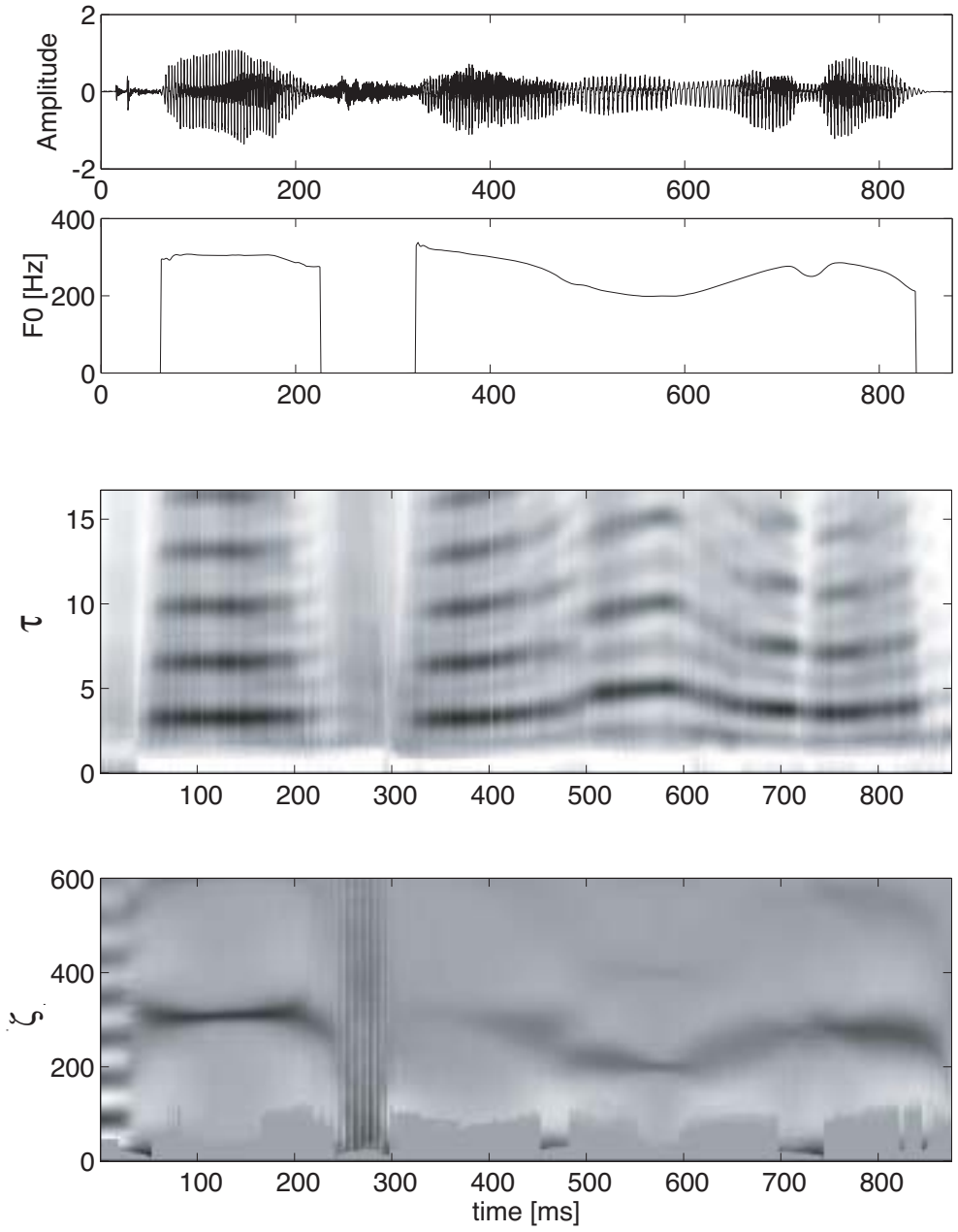

図 3.8: 女性音声に対する平均化された自己相関係数 $a_t(\tau), b_t(\zeta)$: (上から、女性音声波形、その基本周波数、$a_t(\tau)$ の時刻-時間差表示、$b_t(\zeta)$ の時刻-周波数差表示)

は十分起こりうることである。このように $a_t(\tau)$ と $\tilde{b}_t(\tau)$ の統合においては、これらの値が基本周波数らしさを表す指標に過ぎず、Bayes 確率として扱うことができないことを考慮に入れなければならない。

そこで、この問題を Dempster & Shafer の確率理論 [86] で考える。Dempster & Shafer の確率理論では、Bayes の確率のような加法性 ($p(A) + p(\bar{A}) = 1$) を取り除いており、$a_t(\tau)$ を基本周波数らしさを示す基本確率 $m_1(A_\tau)$(A_τ は τ が基本周波数に相当するという焦点要素) だとすると、$1 - a_t(\tau)$ は不信用ではなく信用性の欠如 (基本周波数かどうかはわからない) を示す基本確率 $m_1(A_\tau, \bar{A}_\tau)$ だと考えられる。同様に、$m_2(A_\tau) = \tilde{b}_t(\tau),\ m_2(A_\tau, \bar{A}_\tau) = 1 - \tilde{b}_t(\tau)$ である。よって、Dempster の結合規則

$$m(A_k) = \frac{\sum_{A_{1i} \cap A_{2j} = A_k} m_1(A_{1i}) m_2(A_{2j})}{1 - \sum_{A_{1i} \cap A_{2j} = \phi} m_1(A_{1i}) m_2(A_{2j})} \tag{3.16}$$

に当てはめると、

$$\begin{aligned} c_t(\tau) &= a_t(\tau)\tilde{b}_t(\tau) + (1 - a_t(\tau))\tilde{b}_t(\tau) + a_t(\tau)(1 - \tilde{b}_t(\tau)) \\ &= 1 - (1 - a_t(\tau))(1 - \tilde{b}_t(\tau)) \end{aligned} \tag{3.17}$$

となる [87, 91]。

このようにして得られた統合相関係数 $c_t(\tau)$ から時刻 t の基本周期 (基本周波数) を推定する。

3.6 基本周波数推定実験

本節では、

- 提案法の初期推定が周期性と調波性の両方を利用することにより、それぞれ単独で用いるよりも耐雑音性が向上すること
- 提案法の初期推定が、時間情報のみを用いた自己相関による手法や周波数情報のみを用いたケプストラム法といった従来法と比べても雑音に対して頑健であること

を検証するために、基本周波数推定実験を行う。また、初期推定の次の雑音抑圧部に与える情報として、式 (3.17) の相関係数の値が推定基本周波数の値の確からしさを表す指標となりうるか検証を行う。

3.6.1 雑音に対する頑健性に関する検証

実験条件

音声データは音声と EGG が同時収録されたデータベース [63] の男女各 14 名が発話した 5 文章 (計 140 文) を用いる。音声データのサンプリング周波数は 16 kHz である。雑音として

白色雑音 : 全周波数帯域に雑音のエネルギーがあるため、雑音の影響を受けない周波数帯域が存在しない

1 kHz 以下に帯域制限された雑音 : 1 kHz 以上に雑音の影響を受けない周波数帯域が存在する

の 2 種類を用い、それぞれ SNR が 0–10 dB になるように音声データに加える。有声/無声判定は考慮しないため、評価は有声区間のみで行なった。

評価尺度として

Gross error : 有声区間において、推定誤差が正解基本周波数の ±20%以上である区間の割合

を用いた。Gross error は雑音に対する頑健性を表わす指標となる。正解基本周波数は EGG 波形から STRAIGHT-TEMPO[62] を用いて抽出した値を用いる。

周期性と調波性の利用の効果

式 (3.17) の統合した相関係数 $c_t(\tau)$ を用いた推定結果と、式 (3.13) の周期性から得られた自己相関係数 $a_t(\tau)$ のみを用いた結果、式 (3.15) の調波性から得られた自己相関係数 $b_t(\zeta)$ のみを用いた結果を比較する。白色雑音付加音声に対する Gross error を図 3.9 に、帯域雑音に対する Gross error を図 3.10 に示す。

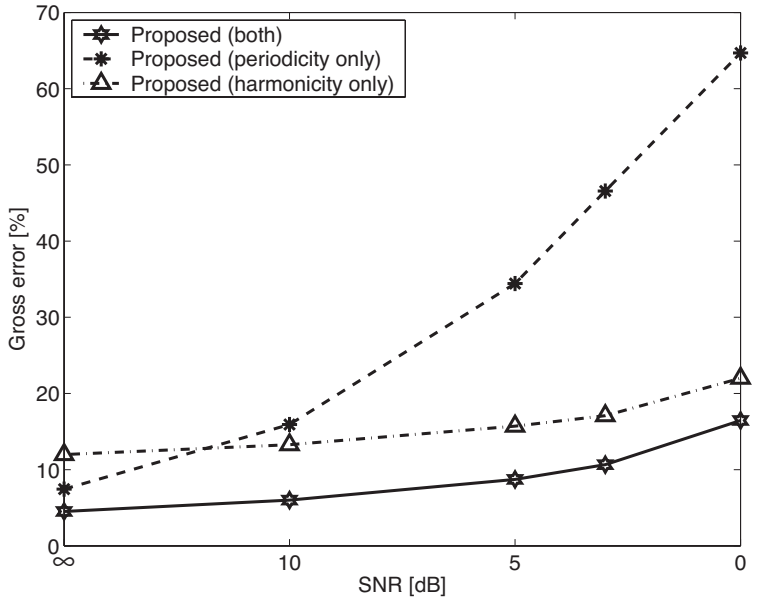

図 3.9: 白色雑音付加音声に対する提案法の初期推定の Gross error

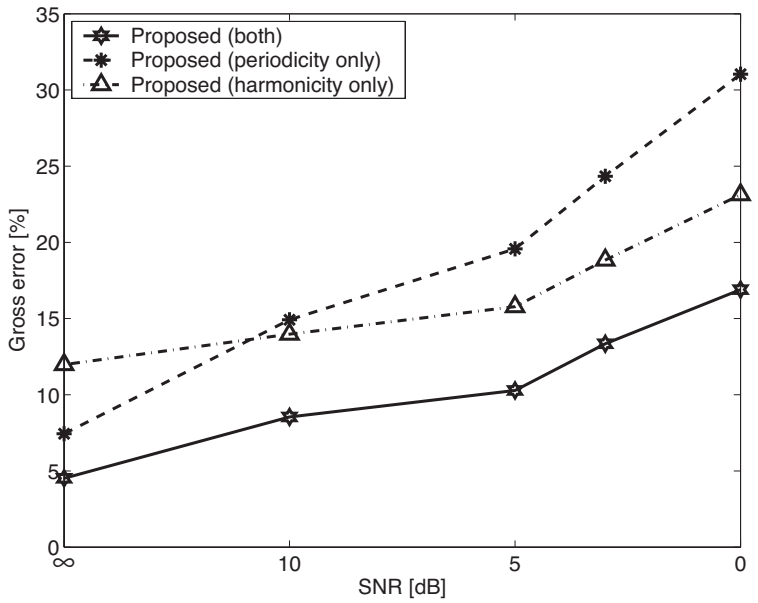

図 3.10: 1 kHz 以下に帯域制限された雑音を付加した音声に対する提案法の初期推定の Gross error

周期性から得られた自己相関係数 $a_t(\tau)$ のみを用いた場合は、雑音がないクリーンな状態では高精度の推定ができているが、白色雑音が強くなるほど Gross error が増大する。これは、白色雑音は全周波数帯域に雑音のエネルギーがあるために、周期性の特徴が歪んでしまい、利用できる帯域がほとんど存在しないためである。調波性から得られた自己相関係数 $b_t(\zeta)$ のみを用いた場合は、雑音が存在しない状態ではやや推定精度が劣るものの、SNR が小さくとも Gross error はそれほど増大しない。周期性と調波性を統合した相関係数から推定した結果は、雑音が存在していても調波性のみの場合よりも高い耐雑音性能を保っている。

1 kHz 以下に帯域制限された雑音では高周波数領域に雑音のエネルギーがないことから、周期性のみを用いた場合でも白色雑音の場合と比べて正しい基本周波数を得やすくなっている。一方、調波性のみを用いる場合では主に低域の瞬時振幅を利用することから雑音の影響を強く受けてしまい、白色雑音の場合より Gross error は増加している。統合した相関係数を用いた場合では、周期性の情報を使うことにより調波性のみの場合よりも Gross error が大きく減少している。

以上の結果から、周期性と調波性の両方を利用することによりそれぞれの利点を取り入れることができ、白色雑音と偏った帯域に強いエネルギーをもつ雑音という異なる雑音に対しても同様の耐雑音性能を得ることができていることがわかる。

従来法との比較

本章で構築した提案法の初期推定部による結果と、従来法のうち比較的耐雑音性能が高い複数窓幅から得られた自己相関関数を用いる推定法 (AC)[26] と移動平均と帯域制限を用いたケプストラムによる推定法 (CEP)[55] による推定結果を比較する。白色雑音付加音声に対する Gross error を図 3.11 に、帯域雑音に対する Gross error を図 3.12 に示す。

白色雑音に対しては、提案法の初期推定部は SNR 10 dB 以上では他の手法にやや劣る面が見られるが、SNR 5 dB 以下では Gross error が最も小さくなっている。また、帯域雑音に対しては、自己相関法の耐雑音性能が大幅に低下しているのに比べ、提案法はケプストラム法とほぼ等しい Gross error となっている。自己相関法の性能低下は低周波数帯域の雑音により音声波形が大きく歪んでいることが原因であると考えられる。ケプストラム法は音声の周波数領域における調波構造を

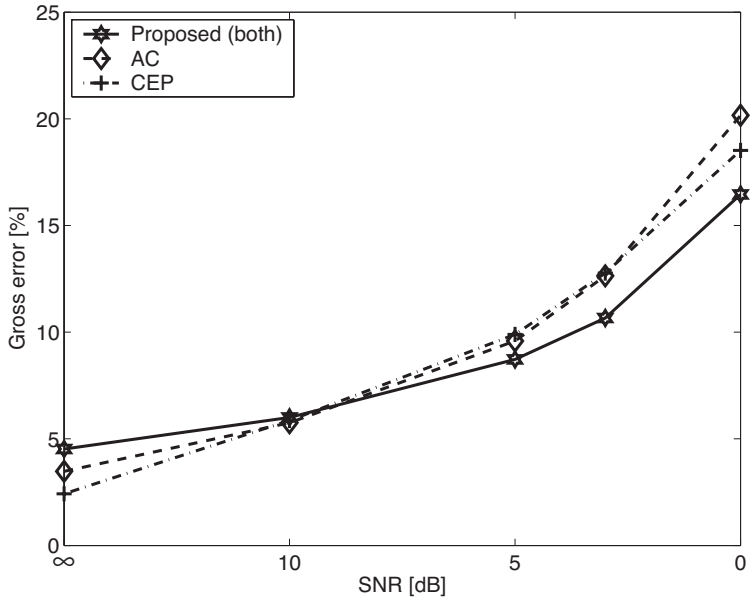

図 3.11: 白色雑音付加音声に対する耐雑音性能比較

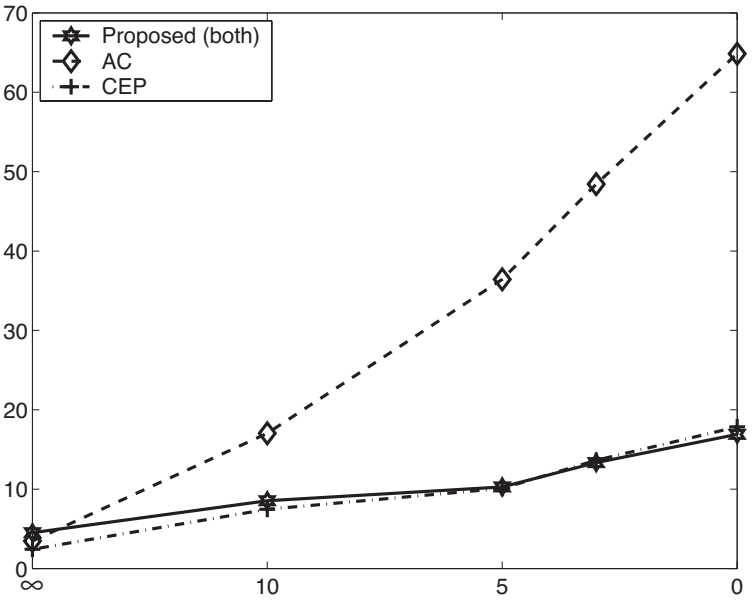

図 3.12: 1 kHz 以下に帯域制限された雑音を付加した音声に対する耐雑音性能比較

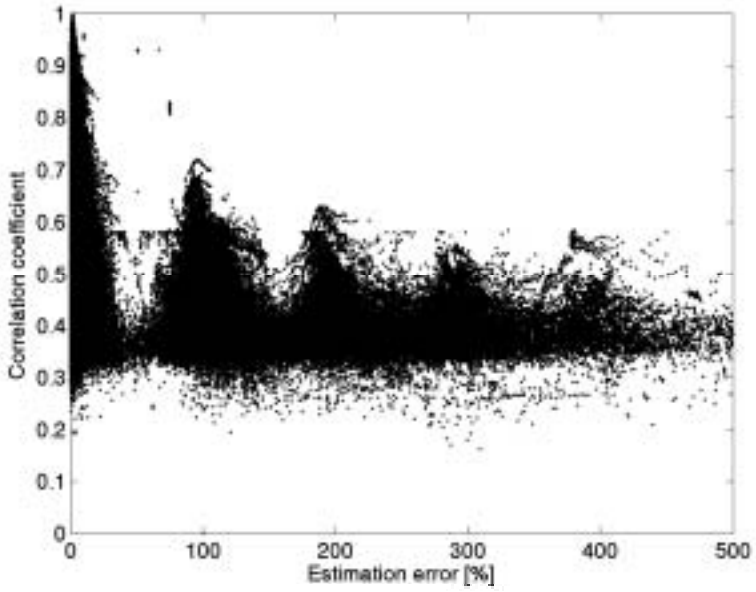

図 3.13: 推定誤差と相関係数の値との関係

利用することから、雑音エネルギーのない高周波数帯域の情報を活用することにより耐雑音性能を保っていると考えられる。提案法の初期推定部も調波性の利用により、ケプストラム法と同様の耐雑音性能を得ている。

以上の結果から、提案法の初期推定部が、雑音に対して頑健な自己相関法やケプストラム法のような基本周波数推定法と比べても、耐雑音性能において優れていることがわかる。

3.6.2 推定精度と相関係数の値の関係の検証

式 (3.17) の相関係数の値と推定誤差との関係を調べ、相関係数の値が推定基本周波数の確からしさを表わす指標となりうるか検証を行う。白色雑音付加音声に対する推定誤差と推定値を選択したときの相関係数の値 (相関係数の最大値) との関係を図 3.13 に示す。

推定誤差の大きさと相関係数の値がなんらかの相関関係になっていれば、相関係数がそのまま推定基本周波数の確からしさを表わす指標となるといえる。しかし、図 3.13 からは相関係数の値と推定誤差の間にはっきりした相関がないように思われる。従って、相関係数の値と基本周波数の確からしさを連続量として結び

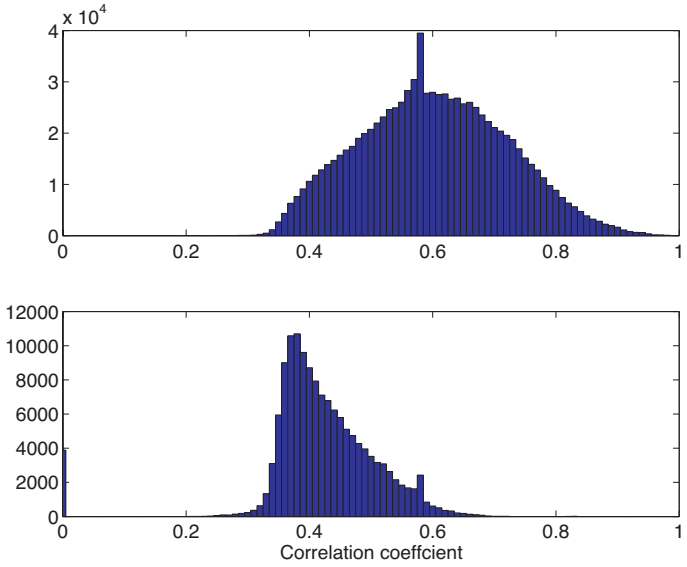

図 3.14: 白色雑音付加音声に対する相関係数の最大値と推定基本周波数誤差の関係: 推定誤差が正解基本周波数の ±20%未満の点における相関係数のヒストグラム (上)、±20%以上の点における相関係数のヒストグラム (下)

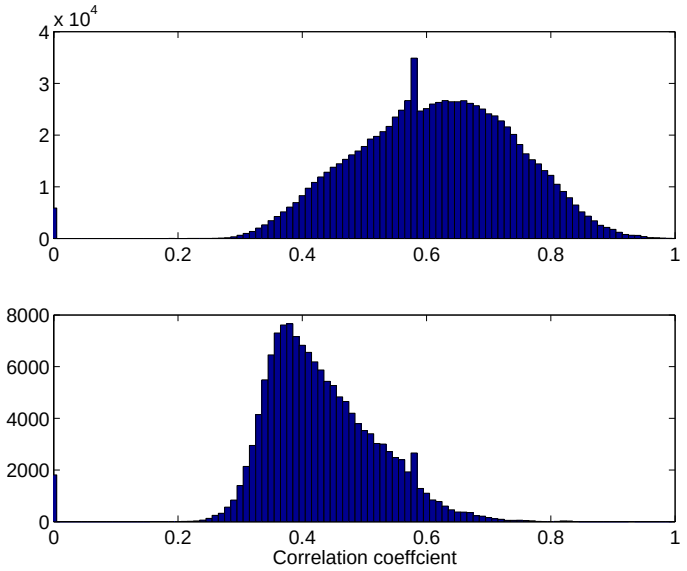

図 3.15: 1 kHz 以下に帯域制限された雑音を付加した音声に対する相関係数の最大値と推定基本周波数誤差の関係: 推定誤差が正解基本周波数の ±20%未満の点における相関係数のヒストグラム (上)、±20%以上の点における相関係数のヒストグラム (下)

つけることはできない。

そこで、推定誤差が ±20%未満のときの推定値を正しく推定できた値として考え、推定値が正しいかどうかを判定する基準を相関係数によって表すことができるかどうかを調べる。

白色雑音付加音声に対して推定誤差が正解基本周波数の ±20%未満であるときの相関係数の値のヒストグラムを図 3.14(上) に、±20%以上であるときの相関係数の値のヒストグラムを図 3.14(下) に示す。同様に、帯域雑音付加音声に対して推定誤差が正解基本周波数の ±20%未満であるときの相関係数の値のヒストグラムを図 3.15(上) に、±20%以上であるときの相関係数の値のヒストグラムを図 3.15(下) に示す。

ここで、カテゴリの代表値として図 3.14(上) の分布の中央値を求めると 0.6136 であり、図 3.14(下) の分布の中央値は 0.4311 である。従って、カテゴリを識別するためにそれぞれの中央値と等距離の値を求めると 0.5223 となる。同様に図 3.15(上) の分布の中央値は 0.6058 であり、図 3.15(下) の分布の中央値は 0.4238 であるため、各分布の代表値から等距離の値は 0.5148 となる。このような分布の特性から、雑音の種類によってある程度の分布の揺らぎがあるものの、相関係数がおよそ 0.5 以上のときは推定誤差が ±20%未満であり、0.5 以下では推定誤差が ±20%以上であると考えられる。よって、相関係数の閾値を 0.5 と定め、相関係数が 0.5 以上であるときは初期推定基本周波数は信頼できる値であると判断し、次の雑音抑圧部に初期推定基本周波数とともに確からしさの情報として相関係数も入力することとする。

3.7 まとめ

本章では、提案法の初期推定部である瞬時振幅の周期性と調波性を利用した基本周波数推定法を構築した。音声の瞬時振幅を時間-周波数表現で表すと、時間方向には基本周期に対応する周期性の振幅変動が現れ、周波数方向には基本周波数に対応する調波性の振幅変動が現れる。この振幅変動は雑音環境においても基本周波数情報を表しうる。周期性と調波性の振幅変動から各周波数帯域ごとに自己相関係数を求め、雑音の影響を抑圧するために各時刻で平均化した後、Dempster

の結合規則によりひとつの相関係数へと統合する。統合された相関係数から基本周波数を求めることにより、雑音を含む音声に対しても頑健に基本周波数を推定することができた。また、初期推定基本周波数の確からしさを表す指標として相関係数の値を用いることができることを示し、相関係数が 0.5 以上であるときは推定値は信頼できる値であり、0.5 以下では信頼できない値であると定めた。

第 4 章

帯域幅可変櫛形フィルタによる雑音抑圧

4.1 はじめに

本章では、提案法の雑音抑圧部である帯域幅可変櫛形フィルタについて述べる。

第 3 章で得られた初期推定基本周波数を櫛形フィルタの中心周波数として、音声の調波成分を残し調波成分以外の雑音を抑圧するような雑音抑圧を行う。このとき、初期推定基本周波数の誤差が大きいと誤って音声の調波成分を取り除いてしまう。そこで、初期推定基本周波数の確からしさに応じて櫛形フィルタの通過帯域幅を変えることができる帯域幅可変櫛形フィルタを構築する。

すなわち、この雑音抑圧部における櫛形フィルタでは、

- 初期推定基本周波数が信頼できる値であれば、通過帯域幅を狭くして雑音抑圧量を多くする
- 初期推定基本周波数が信頼できない値であれば、通過帯域幅を広くして音声の調波成分を誤って除去しないようにする

という処理が求められることになる。

4.2 櫛形フィルタの定式化

櫛形フィルタのブロック図を図 4.1 に示す。

目的信号 $s(t)$ が調波複合音であるとすると、非周期性雑音 $n(t)$ を含む入力信号 $x(t)$ は

$$\begin{aligned} x(t) &= s(t) + n(t) \\ &= \sum_l a_l e^{j(l\omega_0(t)t+\theta_l)} + \sum_m b_m e^{j(\omega_m t+\theta_m)} \qquad (4.1) \\ \omega_0(t) &= 2\pi/T_0(t) \qquad (4.2) \end{aligned}$$

と表せる。ここで、$T_0(t)$ は $s(t)$ の基本周期である。櫛形フィルタの構築を簡単にするために $T_0(t)$ は

$$T_0 = 2\pi/\omega_0 \qquad (4.3)$$

とし、時間によらず一定であるとする。

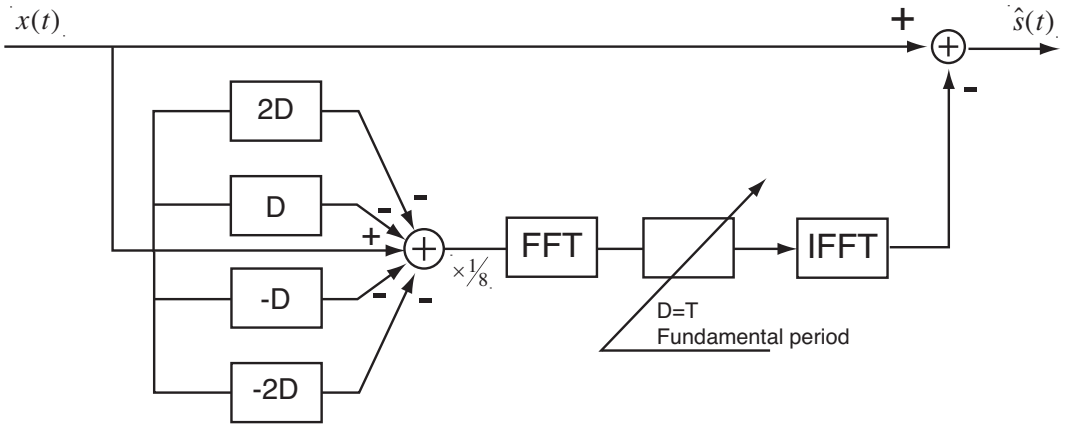

図 4.1: 帯域幅可変櫛形フィルタのブロック図

式 (4.1) と、$x(t)$ を時間方向に $\pm T_0$ および $\pm 2T_0$ ずらした信号との差分信号から次の式が得られる。

$$\begin{aligned} c(t) &= \frac{1}{8}\{4x(t) - x(t-T_0) - x(t+T_0) \\ &\quad - x(t-2T_0) - x(t+2T_0)\} \qquad (4.4) \\ &= \sum_m b_m e^{j(\omega_m t+\theta_m)} \cdot d(\omega_m) \qquad (4.5) \end{aligned}$$

ただし、

$$d(\omega_m) = \frac{1}{2} - \frac{1}{4}\left(\cos\frac{2\omega_m}{\omega_0}\pi + \cos\frac{4\omega_m}{\omega_0}\pi\right) \qquad (4.6)$$

である。$c(t)$ のフーリエ変換を $C(\omega_m)$ とし、$n(t)$ のフーリエ変換を $N(\omega_m)$ とすると、式 (4.5) より、

$$C(\omega_m) = N(\omega_m) \cdot d(\omega_m) \qquad (4.7)$$

となる。すなわち、$C(\omega_m)$ は $N(\omega_m)$ が $d(\omega_m)$ により歪んだ信号となっている。よって、雑音スペクトル $N(\omega_m)$ は

$$N(\omega_m) = C(\omega_m)/d(\omega_m) \qquad (4.8)$$

と表される。従って、$N(\omega_m)$ を逆フーリエ変換して得られる雑音信号を用いて

$$s(t) = x(t) - n(t) \qquad (4.9)$$

として、目的信号が得られることとなる。これは中心周波数が $1/T_0$ の整数倍の櫛形フィルタに等しい。しかし、ω_m/ω_0 が整数のとき式 (4.6) より $d(\omega_m) = 0$ とな

り、式 (4.8) では $N(\omega_m)$ が不定となるため、

$$\hat{N}(\omega_m) = \begin{cases} C(\omega_m)/d(\omega_m), & d(\omega_m) \geq \varepsilon \\ C(\omega_m), & d(\omega_m) < \varepsilon \end{cases} \tag{4.10}$$

として雑音スペクトルを推定することとする。推定雑音スペクトルを逆フーリエ変換した推定雑音 $\hat{n}(t)$ を用い、

$$\hat{s}(t) = x(t) - \hat{n}(t) \tag{4.11}$$

として、雑音抑圧信号 $\hat{s}(t)$ を得る。このとき、式 (4.10) は ε の値により通過帯域幅が変わる櫛形フィルタとなっている。この櫛形フィルタを帯域幅可変櫛形フィルタと呼ぶこととする。帯域幅可変櫛形フィルタによる雑音抑圧システムの周波数応答を図 4.2 に示す。帯域幅パラメータ ε の値が小さいと通過帯域幅は狭く、ε の値が大きくなるほど通過帯域幅は広くなる。

4.3 帯域幅パラメータの決定

初期推定部の推定誤差に応じて通過帯域幅を変化させることにより、雑音をできるだけ抑圧し、かつ音声の調波成分を誤って除去しない雑音抑圧を行う。ここで、どのように通過帯域幅を決定するかが問題になる。雑音抑圧による雑音抑圧音声の SNR 向上と最終推定基本周波数の誤差減少との間には必ずしも関連があるわけではなく、たとえ雑音抑圧部により SNR が低下しても推定誤差が減少するということは起こりうる。すなわち、櫛形フィルタだけの雑音抑圧性能を考えることは本提案法のシステムにおいては適さない。そこで、櫛形フィルタだけではなく、次の最終基本周波数推定部までを含めて、帯域幅の決定方法を検討する。

4.3.1 初期推定基本周波数の誤差と櫛形フィルタの帯域幅の関係

まず、初期推定基本周波数の誤差と櫛形フィルタの帯域幅、最終推定結果の関係を考える。

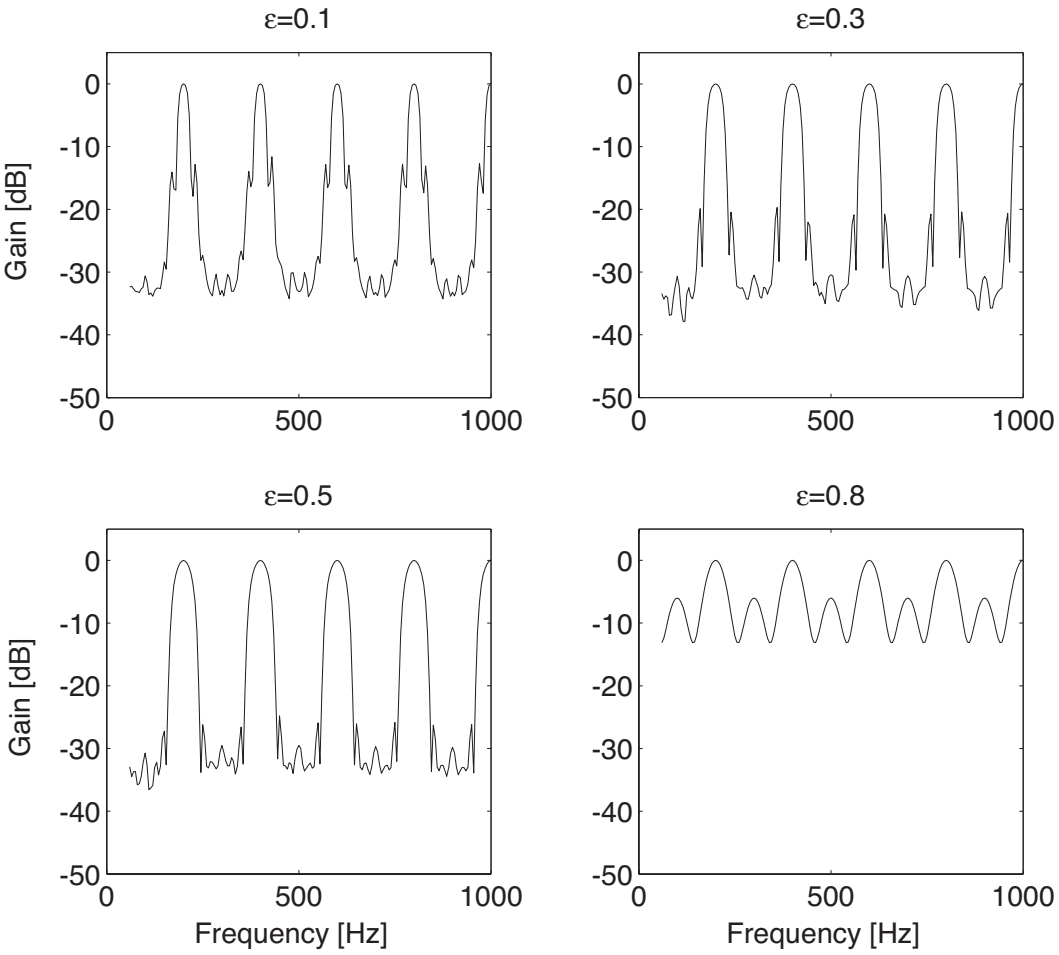

図 4.2: 帯域幅可変櫛形フィルタによる雑音抑圧システムの周波数応答

目的

櫛形フィルタに入力される基本周波数の誤差が小さい場合と大きい場合について、最適な通過帯域幅を調べる。

実験条件

音声とEGGが同時収録されているデータベース[63]から男女各7名発話の5文章を用いる。参照のための正解基本周波数はEGG波形からSTRAIGHT-TEMPO[62]を用いて求めた。音声には白色雑音をSNR 0–10 dBとなるよう付加した。

様々な帯域幅 ($\varepsilon = 0.1 - 1.0$ で0.1刻み) をもつ櫛形フィルタを用意し、正解基本周波数 $F_0(n)$ Hzに対して

- $F_0(n)$±5 Hz
- $F_0(n)$±50 Hz

とした周波数を櫛形フィルタに入力した。これは推定誤差を人為的に加えたことに相当する。$F_0(n)$±5 Hzは初期推定部における推定誤差が小さい場合を想定しており、$F_0(n)$±50 Hzは推定誤差が大きい場合を想定している。

得られた雑音抑圧音声から次の最終基本周波数推定部(STRAIGHT-TEMPO)を用いて最終推定結果を求めた。

評価尺度は、3.6.1節と同じく、有声区間で推定基本周波数と正解基本周波数の差が±20%以上である割合をGross errorとして用いた。

実験結果

図4.3に櫛形フィルタに $F_0(n)$± 5 Hzの基本周波数を入力したときの、最終推定結果に与える影響を示す。これは初期推定での基本周波数推定誤差が小さい場合と考えることができる。雑音がないクリーンな音声では ε の値が0.1以上であれば最終結果にはほとんど影響がない。一方、雑音が大きいときは、ε の値が0.3–0.4で最もGross errorが小さくなることがわかる。

図4.4に櫛形フィルタに $F_0(n)$± 50 Hzの基本周波数を入力したときの、最終推定結果に与える影響を示す。これは初期推定部における誤差が大きい場合と考え

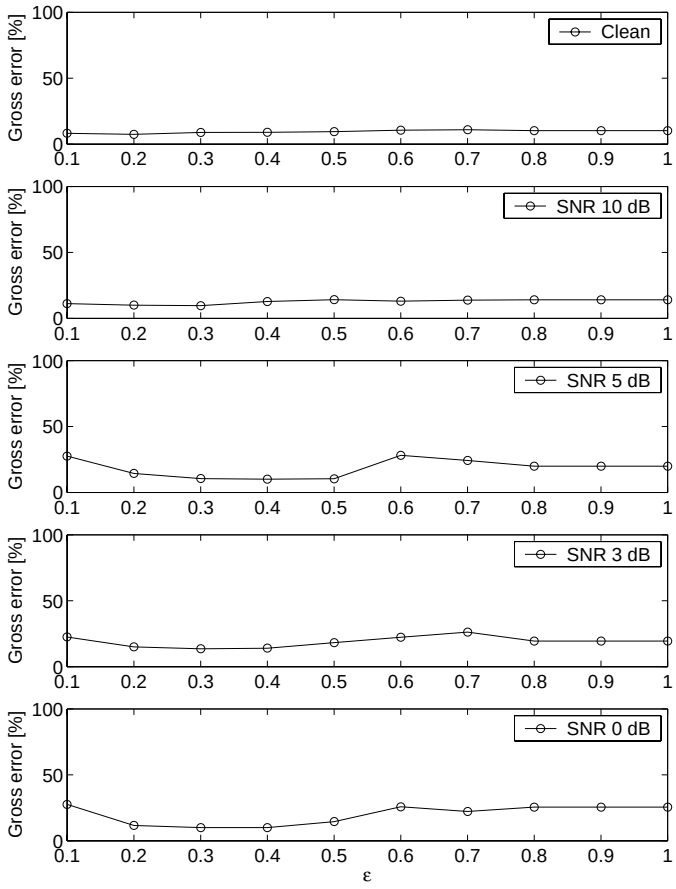

図 4.3: 櫛形フィルタの通過帯域幅パラメータ ε と最終推定結果の関係 (正解基本周波数 $F_0(n)$±5 Hz が櫛形フィルタに入力された場合)

られる。雑音がないクリーンな音声の場合には、ε の値を小さくし帯域幅を狭くすると Gross error が増大してしまう。これは櫛形フィルタの中心周波数が音声の基本周波数や高調周波数からずれているために、音声の調波成分を誤って抑圧してしまったことによると思われる。一方、ε を 0.7 以上とすると推定誤差が大きくとも最終結果にはほとんど影響がない。雑音が大きさに応じて Gross error は大きくなっていくが、ε の値が 0.8–1.0 で最も Gross error が小さくなる。

また、予備実験において、基本周波数の誤差が 5 Hz と 50 Hz の間である場合も同様に Gross error が小さくなる ε の値は 0.3–0.4 または 0.8–1.0 であった。

以上の結果から、櫛形フィルタに入力される基本周波数の誤差と通過帯域幅を決めるパラメータ ε の関係について、

- 基本周波数の誤差が小さいときは ε=0.3–0.4

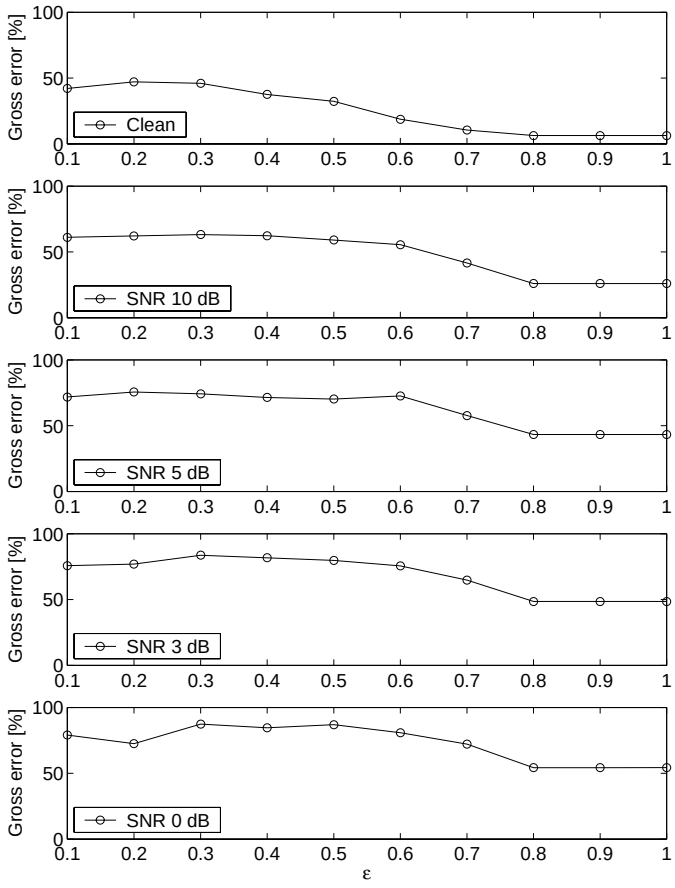

図 4.4: 櫛形フィルタの通過帯域幅パラメータ ε と最終推定結果の関係 (正解基本周波数 $F_0(n)$±50 Hz が櫛形フィルタに入力された場合)

- 基本周波数の誤差が大きいときは ε=0.8–1.0

が適していることがわかった。

4.3.2 櫛形フィルタの帯域幅の決定方法

3.6.2 節で述べたように、初期推定部において

- 相関係数の値が 0.5 以上のときは推定誤差が ± 20%未満である可能性が高い (初期推定基本周波数は信頼できる値である)
- 相関係数の値が 0.5 未満のときは推定誤差が ± 20%以上である可能性が高い (初期推定基本周波数は信頼できない)

ということがいえる。また、4.3.1 節から櫛形フィルタの帯域幅について以下のことが確認できた。

- 初期推定における誤差が小さいときは ε=0.3–0.4 が適している
- 初期推定のおける誤差が大きいときは ε=0.8–1.0 が適している

そこで、前節で帯域幅パラメータ ε の値が 0.3 と 0.4 では Gross error に大きな差がみられなかったこと、同様に 0.8 から 1.0 の間でも大きな差がないことから、

- 初期推定部において相関係数が 0.5 以上であったときは $\varepsilon = 0.3$
- 初期推定部において相関係数が 0.5 未満であったときは $\varepsilon = 0.8$

として櫛形フィルタの通過帯域幅を調節することとする。

4.4 基本周期を一定とする波形の時間伸縮

4.2 節の帯域幅可変櫛形フィルタの定式化において、式 (4.3) で式の簡単化のために基本周期を時間によらず一定と仮定した。しかし、実際の音声では基本周期は時間とともに変化するため、式 (4.3) の仮定は誤差の原因となる。そこで、櫛形フィルタによる雑音抑圧の前に、初期推定基本周波数を使って基本周期が一定となるような音声波形の時間伸縮を行なう。

基本周期を一定とする音声波形の伸縮はサンプリング間隔を変えたのちにリサンプリングすることで行う。初期推定部によって推定された基本周期 $T_0(t)$ を使って、サンプリング間隔は次の式で変えられる。

$$\tilde{T}_s(t) = \frac{\overline{T}_0}{T_0(t)} \times T_s \tag{4.12}$$

ここで、$\overline{T}$ は $T_0(t)$ の平均値であり、T_s は元の音声波形のサンプリング周期である。$\tilde{T}_s(t)$ で表された音声波形は T_s 間隔で補間によりリサンプリングされる。このとき、音声波形は一定の基本周期を持つ波形となる。図 4.5 は時間伸縮波形の例を示す。櫛形フィルタによる雑音抑圧ののちに、音声波形は初期推定基本周波数を使った式 (4.12) と逆の操作により元の基本周期を持つ音声波形へと時間伸縮される。

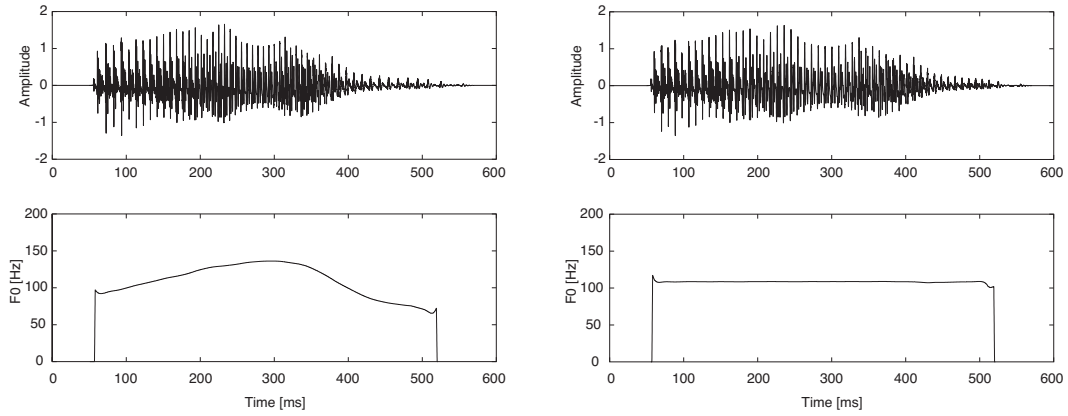

図 4.5: 基本周期を一定とする音声波形の時間伸縮: (左) 元音声波形、(右) 時間伸縮後の音声波形

4.5 雑音抑圧性能の検証

4.3.2 節で定めた通過帯域幅に関する ε の決定方法により、ε を固定した値にするよりも効果的に雑音抑圧できることを検証するために、基本周波数推定実験を行う。

4.5.1 実験条件

音声データは、4.3.1 節において帯域幅パラメータの決定に用いたデータと異なる話者の同じ文章とし、男女各 7 名の 5 文章を用いた。

雑音として白色雑音と 1 kHz 以下の帯域雑音を用い、SNR 0–10 dB となるよう音声データに付加した。評価尺度は、有声区間で推定基本周波数と正解基本周波数の差が ±20%以上である割合を Gross error として用いた。

比較として、帯域幅パラメータ ε を 0.3 と 0.8 に固定した場合の結果についても求めた。

4.5.2 実験結果

白色雑音を付加された音声に対する最終推定結果を図 4.6(上) に示す。また、帯域雑音を付加された音声に対する推定結果を図 4.6(下) に示す。

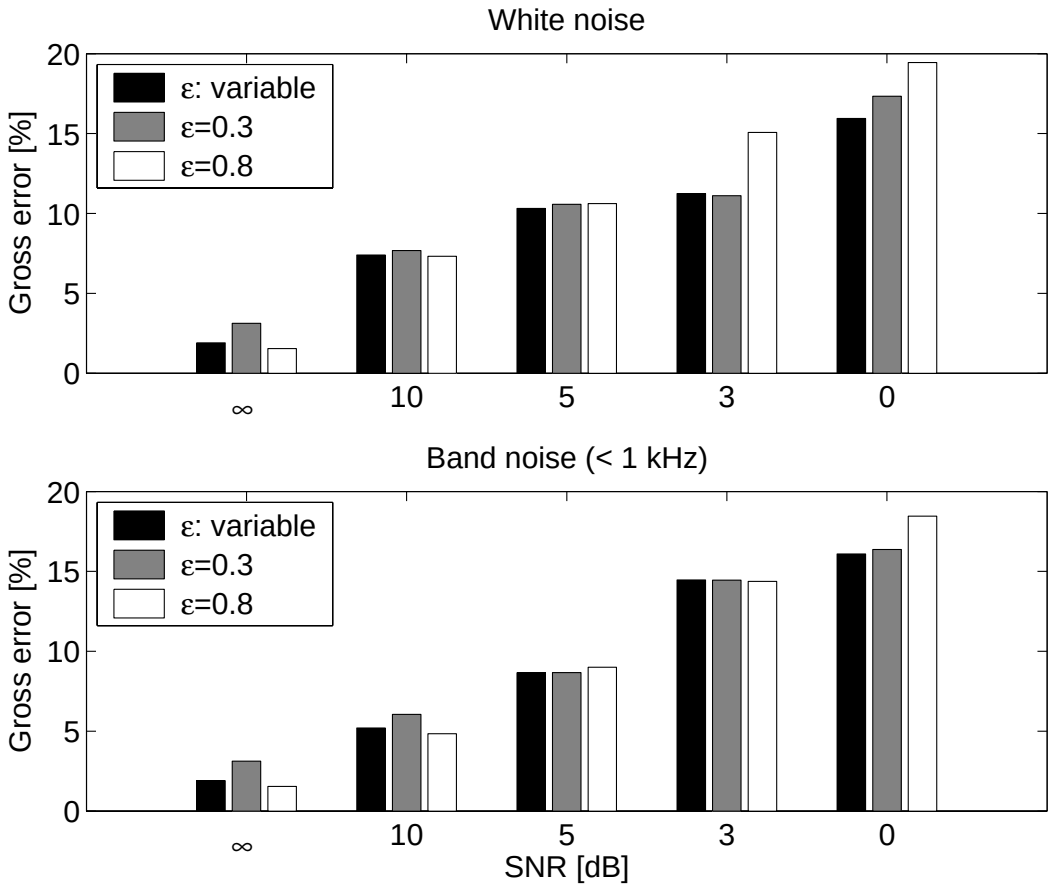

図 4.6: 帯域幅パラメータ ε の可変による効果: (上) 白色雑音付加音声に対する推定結果、(下) 帯域雑音付加音声に対する推定結果

雑音が白色雑音の場合も帯域雑音の場合もほぼ同傾向の結果がみられる。ε を 0.3 と 0.8 に固定したときの結果を比較すると、雑音が存在しないときでは $\varepsilon = 0.8$ として広い帯域幅を用いたほうが良い結果が得られた。雑音が大きくなったときには $\varepsilon = 0.3$ として狭い帯域幅で雑音を抑圧したほうが Gross error は小さくなる。ε を可変にしたときは、雑音がない状態のときこそ $\varepsilon = 0.8$ での結果よりもやや劣るものの、雑音が存在する状態では ε を固定したときよりも Gross error は小さくなる。これは 4.3.2 節で定めた通過帯域幅に関する ε の決定方法が有効に働いていることを示している。すなわち、雑音が存在する状態であっても、一律に ε の値を小さくして雑音を抑圧するだけではなく、場合によっては帯域幅を広くすることによってより効果的に推定精度を向上させることができる。

4.6　まとめ

本章では、提案法の雑音抑圧部である帯域幅可変櫛形フィルタについて述べた。

第 3 章で得られた初期推定基本周波数を櫛形フィルタの中心周波数として用いることにより、音声の調波成分を残し調波成分以外の雑音を抑圧するような雑音抑圧を行う。このとき、初期推定部の相関係数の値を推定基本周波数の確からしさとして利用し、推定誤差に応じて櫛形フィルタの通過帯域幅を変えることができ、帯域幅を固定した場合よりも効果的に雑音抑圧が可能な帯域幅可変櫛形フィルタを構築した。

本章の雑音抑圧部によって雑音抑圧された音声が最終推定部に入力され、最終基本周波数が推定される。最終基本周波数の推定精度及び提案法全体の耐雑音性能については、第 5 章において検証を行う。

第 5 章

雑音環境における有効性検証

5.1 はじめに

本章では、計算機シミュレーションにより、本研究で提案する基本周波数推定法の耐雑音性能及び推定精度を検証する。

実環境に存在する雑音は白色雑音やピンク雑音のような人工的な雑音とは異なり、その雑音エネルギーが特定の周波数帯域に集中していることが多く、雑音の影響の少ない周波数帯域が存在する。よって、白色雑音やピンク帯域雑音のような人工的な雑音の他に、実環境における有効性を検証するために、走行自動車内で収録された雑音 (走行自動車内雑音) とデパート内で収録された雑音 (デパート内雑音) に対する頑健性についても評価を行う。

5.2 計算機シミュレーション

5.2.1 実験条件

音声データ

実験に用いた音声データは阿竹らによる音声と EGG が同時収録されたデータベース [63] を用いた。データ数は男女各 14 名の発話による 30 文章 (計 840 文) である。音声データ及び EGG 波形のサンプリング周波数は 16 kHz である。

雑音データ

実験に用いた雑音は次の 4 種類である。

白色雑音 : 全周波数帯域に等しい雑音エネルギーが存在する。

ピンク雑音 : 全周波数帯域に雑音エネルギーがあるが、低周波数から高周波数へ向けて傾斜したパワースペクトルをもつ。パワースペクトルは高域へ向けて 1 オクターブあたり 3 dB 減少しており、低域にエネルギーが偏っている。

走行自動車内雑音 : エンジン音やロードノイズなどにより低周波数帯域にほとんどの雑音エネルギーが存在する。

デパート内雑音 : アナウンス放送や周りの客のざわめきなど多数の他話者の音声が入り混っているため、パワースペクトルは音声に近い。

白色雑音とピンク雑音はこれまでに耐雑音性能の評価によく用いられてきた人工的な雑音であり、走行自動車内雑音とデパート内雑音は実環境における運用を仮定するために用いる。走行自動車内雑音とデパート内雑音は電子協騒音データベース [81] に収録されている騒音を用いた。サンプリング周波数は 48 kHz から 16 kHz へダウンサンプリングしている。

雑音は各音声データに SNR が 0–10 dB となるように加えた。

評価尺度

本実験においては、1.3.3 節と同様に、次の 2 種類の評価尺度を用いた。

Gross error : 有声区間において推定誤差が ±20%以上である区間の割合

Fine error : 推定誤差が ±20%未満である区間内での誤差の平均

Gross error は基本周波数が推定できなかった区間を表わすことから雑音に対する頑健性を示す指標となる。また、Fine error は基本周波数が推定できた区間内における推定基本周波数の誤差を表わすことから、推定精度を示す指標となる。1.3.3 節で述べたように、ここで用いている「耐雑音性能 (頑健性)」とは基本周波数が存在する音声区間でその基本周波数に近い値を抽出できるかどうかを表し、「推定精度」とは正しい基本周波数と推定基本周波数がどれ程異なるかを表すものとする。

比較のために、

- 複数窓幅から得られた自己相関関数を用いる推定法 (AC)[26]
- 振幅差関数に重み付けを行う推定法 (YIN)[28]
- 移動平均と帯域制限を用いたケプストラムによる推定法 (CEP)[55]
- 瞬時周波数の不動点を利用した推定法 (STRAIGHT-TEMPO)[62]

の 4 つの基本周波数推定法における結果も合わせて示す。

5.2.2 白色雑音

白色雑音を付加された音声に対する Gross error を図 5.1 に示す。また、Fine error を図 5.2 に示す。

図 5.1 から、他の手法と比較して提案法は、雑音の影響による Gross error の増加量が少なく、SNR が 5 dB 以下で Gross error が最小となっている。このことから提案法が白色雑音に対して頑健に基本周波数が推定できていることがわかる。また、提案法の最終推定部で用いている STRAIGHT-TEMPO は雑音の影響を受けて Gross error が大幅に増大しているのに対し、提案法の Gross error は最も小さくなっていることから、提案法の雑音抑圧部において、雑音抑圧が有効にはたらいていると考えられる。

図 5.2 から、雑音が存在しない状況では提案法は STRAIGHT-TEMPO ほどの高精度な推定はできていないが、他の手法に劣らない精度で基本周波数が推定できていることがわかる。また、SNR が小さくなっていくと他の手法の Fine error は増加するのに対し、提案法の Fine error はほとんど増加していない。すなわち、提案法は白色雑音の存在する状況においても、Fine error の評価閾値である誤差 ±20%未満であれば、クリーンな音声に対する基本周波数とほぼ同精度の基本周波数が推定可能であることがわかる。

5.2.3 ピンク雑音

ピンク雑音を付加された音声に対する Gross error を図 5.3 に示す。また、Fine error を図 5.4 に示す。

音声は低周波数帯域に比較的強いエネルギーをもち、高周波数帯域のエネルギーは小さくなっているために、ピンク雑音のスペクトル構造に近い。そのため、同じ SNR 時の白色雑音に比べるとピンク雑音は音声の全周波数帯域にわたって音声情報を歪ませやすくなる。図 5.3 の Gross error の結果も強く雑音の影響を受けており、白色雑音の場合よりも Gross error の値が大きくなっている。提案法は SNR 10 dB では自己相関法の Gross error よりも大きくなっているが、SNR 3 dB 以下では最も小さい Gross error となっている。すなわち、提案法は、白色雑音の場合と同様に、他の手法に劣らない高い耐雑音性能を示している。

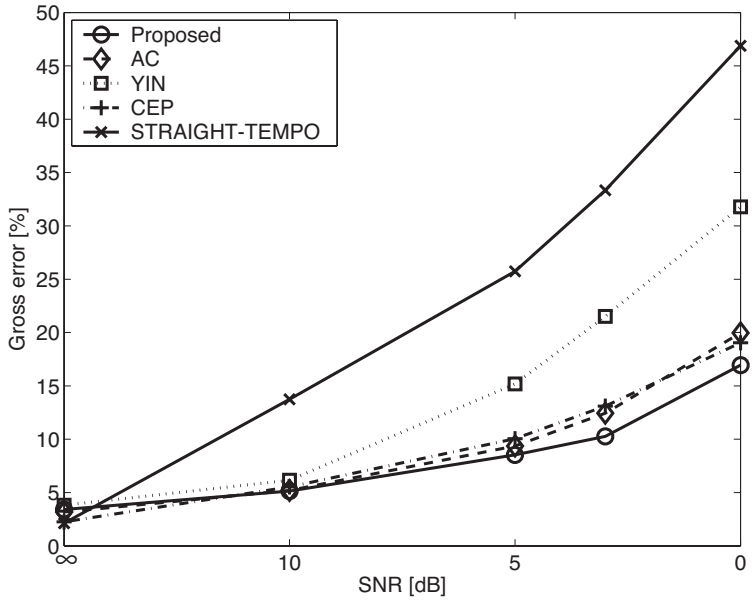

図 5.1: 白色雑音付加音声に対する基本周波数推定法の Gross error

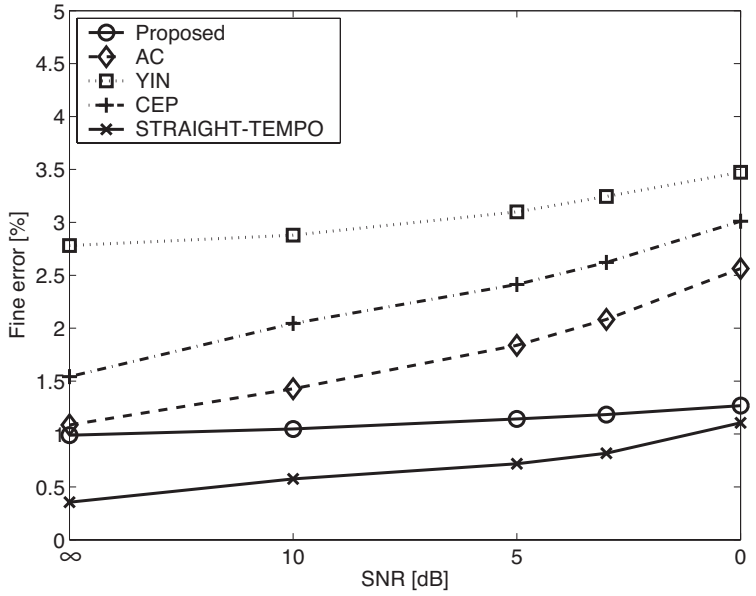

図 5.2: 白色雑音付加音声に対する基本周波数推定法の Fine error

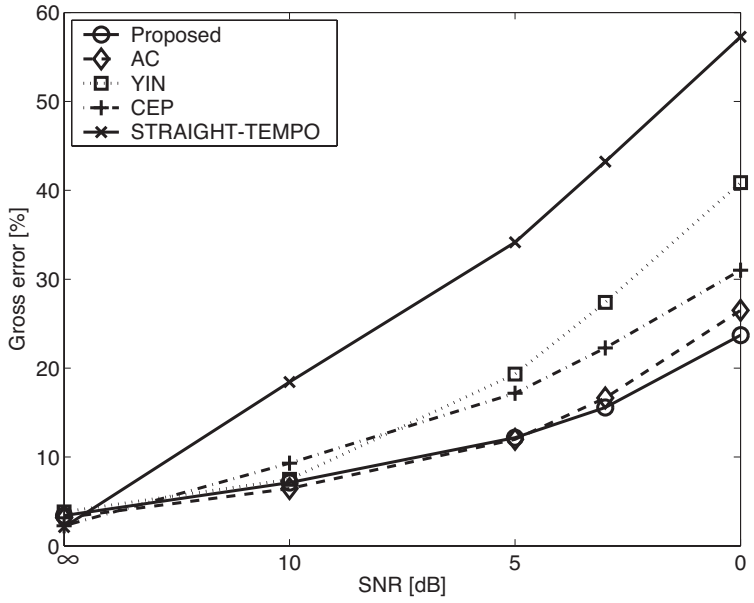

図 5.3: ピンク雑音付加音声に対する基本周波数推定法の Gross error

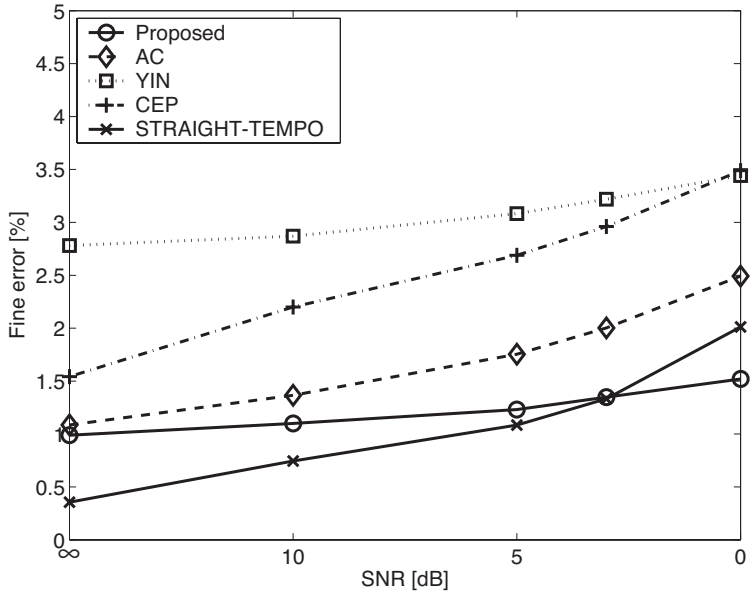

図 5.4: ピンク雑音付加音声に対する基本周波数推定法の Fine error

図 5.4 の fine error の結果も白色雑音の場合とほぼ同様の傾向を示しており、提案法はピンク雑音に対してもクリーンな音声とほぼ同精度の基本周波数推定ができている。

5.2.4 走行自動車内雑音

走行自動車内雑音を付加された音声に対する Gross error を図 5.5 に示す。また、Fine error を図 5.6 に示す。

走行自動車内雑音は白色雑音やピンク雑音と異なり、低周波数帯域にエネルギーが集中していて、高周波数帯域にはほとんど雑音エネルギーが存在しない。図 5.5 をみると、自己相関法や YIN などは走行自動車内雑音の低周波数成分によって音声波形の振幅が大きく変動するために強く雑音の影響を受けている。一方、雑音のエネルギーのない高周波数帯域の情報を利用できるケプストラム法や STRAIGHT-TEMPO は比較的雑音の影響を受けにくい。提案法も、初期推定部において瞬時振幅の時間-周波数表現の高周波数帯域に現れる周期性の特徴を利用しており、また最終推定部では STRAIGHT-TEMPO を用いているために、走行自動車内雑音による Gross error の増加量は少なく、SNR 10 dB 以下において Gross error が最も小さくなっている。

また、図 5.6 における Fine error も STRAIGHT-TEMPO が雑音の影響により推定精度が大きく低下しているにもかかわらず、提案法においてはほとんど変化がなく、雑音に関係なく高精度な推定ができていることがわかる。

5.2.5 デパート内雑音

デパート内雑音を付加された音声に対する Gross error を図 5.7 に示す。また、Fine error を図 5.8 に示す。

デパート内雑音には背景雑音とした話し声が多数含まれているために、目的音声の時間領域での周期性も周波数領域での調波性も乱されやすい。そのため、図 5.7 の Gross error において提案法はすべての SNR で自己相関法に劣る結果となった。これは、提案法が時間-周波数解析によりできるだけ基本周波数に関連した情報を

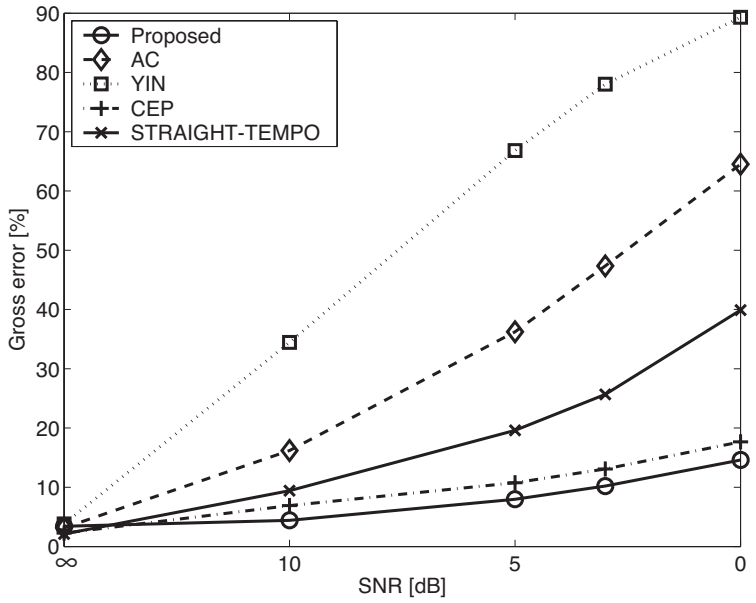

図 5.5: 走行自動車内雑音付加音声に対する基本周波数推定法の Gross error

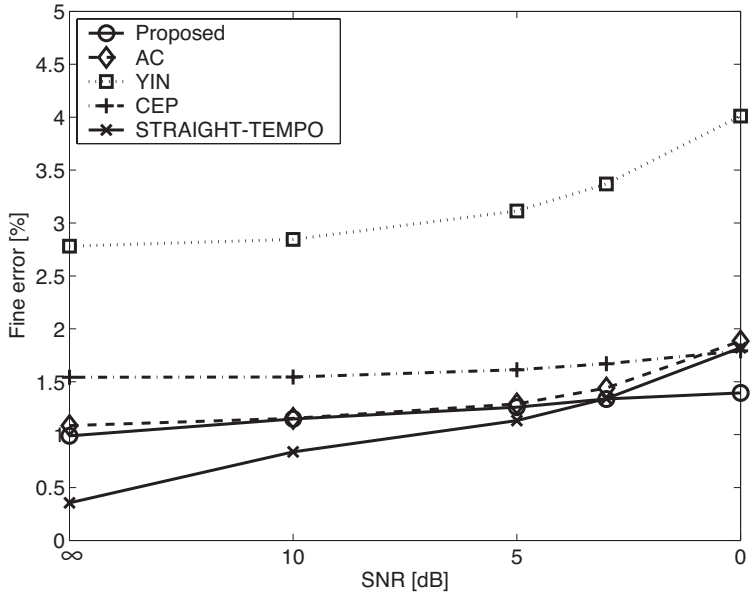

図 5.6: 走行自動車内雑音付加音声付加音声に対する基本周波数推定法の Fine error

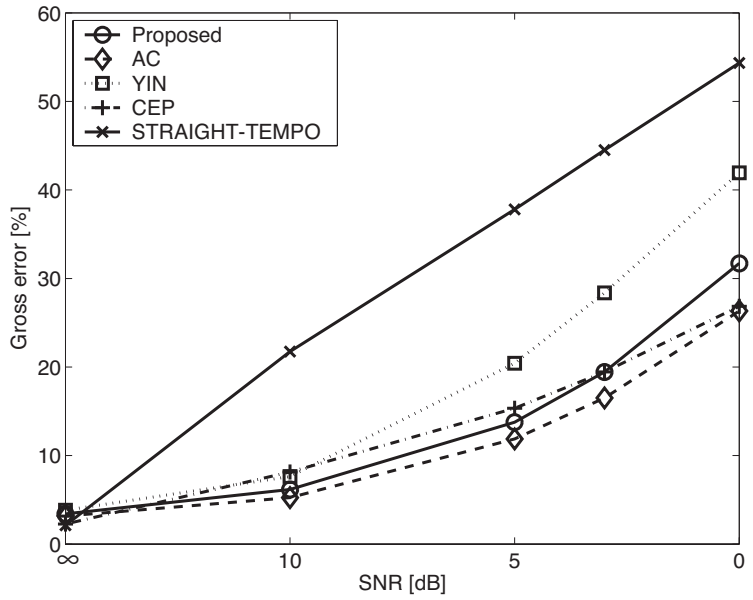

図 5.7: デパート内雑音付加音声に対する基本周波数推定法の Gross error

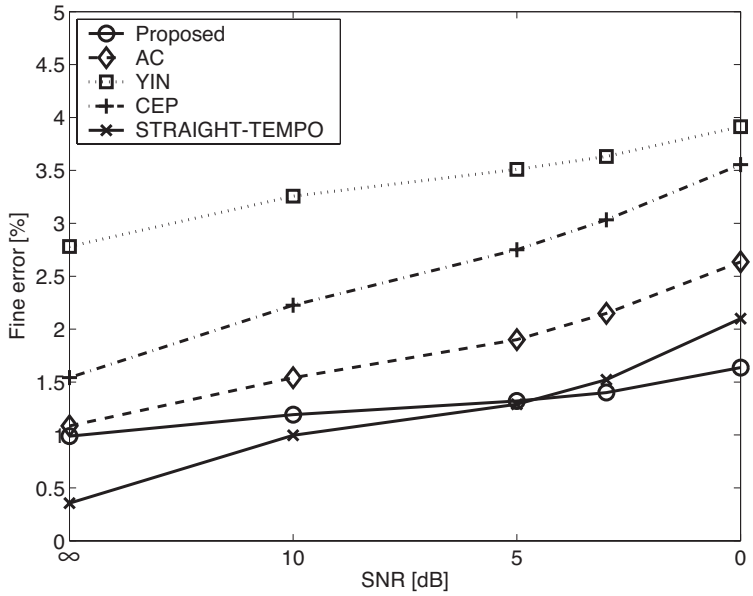

図 5.8: デパート内雑音付加音声付加音声に対する基本周波数推定法の Fine error

得るという方策をとった結果、雑音として扱われるべき音における基本周波数情報の影響を受けてしまったためであると考えられる。図 5.8 の Fine error では他の雑音と同様に提案法は雑音による Fine error の増加は少なく、基本周波数が推定できる区間では高精度な値を推定できている。

5.2.6 考察

提案法と従来の基本周波数推定法の差を明確にするため、耐雑音性能と推定精度について表 5.1 に示す。

雑音が白色雑音の場合、提案法は他の手法と比べ最も Gross error が小さく、高い耐雑音性を示した。STRAIGHT-TEMPO は雑音によって大幅に Gross error が増大しているが、その STRAIGHT-TEMPO を最終推定部に利用している提案法にはその影響は現れていない。これは、提案法の雑音抑圧部が正しく雑音抑圧できているためであると考えられる。また、Fine error について、他の手法は雑音の増加とともに Fine error が増大しているが、提案法の Fine error の値はクリーンな音声の場合とほとんど変化がない。よって、提案法は白色雑音に対して頑健で高精度な推定ができていることがわかる。

雑音がピンク雑音の場合、提案法は SNR 5 dB 以下で Gross error が他の手法と比べて最小となった。ピンク雑音のエネルギー分布は音声のエネルギー分布に近いため、白色雑音と異なり音声にとっては全周波数帯域にわたって等しく妨害を受けていることになる (白色雑音の場合は音声にとっては低周波数帯域よりも高周波数帯域のほうが強く妨害を受けていることになる)。このような場合でも提案法は他の手法と同等かそれ以上に頑健に基本周波数推定ができている。また、STRAIGHT-TEMPO の Gross error が大きいにもかかわらず提案法の Gross error が小さくなっていることから、ピンク雑音の場合も雑音抑圧が適切に働いていることがわかる。Fine error も提案法は他の手法と比べて最も雑音による増加量が少なく、クリーンな音声に対する値に近い。よって、提案法はピンク雑音に対しても頑健かつ高精度な推定となっている。

走行自動車内雑音の場合も、提案法の Gross error は他の手法と比べて最も小さくなっている。また、白色雑音やピンク雑音の場合と比較しても、走行自動車内

表 5.1: 提案法と従来の基本周波数推定法の耐雑音性能と推定精度

推定法	Gross error			
	白色雑音	ピンク雑音	走行自動車内雑音	デパート内雑音
提案法	◎	○	◎	△
自己相関法	◎	○	×	○
YIN	△	×	×	×
ケプストラム法	◎	△	◎	○
TEMPO	×	△	△	×

SNR 0 dB での Gross error: ◎ (〜20%), ○ (20〜30%), △ (30〜40%), × (40%〜)

推定法	Fine error			
	白色雑音	ピンク雑音	走行自動車内雑音	デパート内雑音
提案法	◎	◎	◎	○
自己相関法	×	○	○	△
YIN	○	○	△	△
ケプストラム法	△	×	◎	×
TEMPO	○	×	△	×

クリーン音声と雑音付加音声 (SNR 0 dB) の Fine error の差:

◎ (〜0.5 pt.), ○ (0.5〜1.0 pt.), △ (1.0〜1.5 pt.), × (1.5 pt.〜)

雑音に対する Gross error は小さい。これは、走行自動車内雑音が低周波数帯域にエネルギーが集中しており、高周波数帯域にほとんどエネルギーが存在しないことが原因であると考えられる。提案法は、その初期推定部において時間-周波数平面上の様々な周波数帯域から基本周波数情報を得ている。そのため、低 SNR 時でも高周波数帯域から正しい基本周波数を推定することができ、雑音抑圧を有効に行うことができたと考えられる。実環境における雑音の多くは白色雑音のように全周波数帯域に雑音が等しく存在するのではなく、ある周波数帯域にエネルギーが集中しており、雑音のエネルギーが非常に小さくなっている周波数帯域が存在すると考えられる。提案法の初期推定部はその雑音のエネルギーが少ない帯域から基本周波数情報を得ることができている。提案法は Fine error の値もクリーンな音声の場合とほとんど変化がなく、高精度な推定もできていることがわかる。

雑音がデパート内雑音の場合は、提案法の Gross error は他の手法よりもわずかに高くなっている。デパート内雑音は周囲に大勢の人がいる環境であり、大勢の人の話し声やアナウンス放送などが含まれている。そのため、目的音声が他の音声によって妨害されることになり、目的音声の調波成分が乱されやすい。また、音声波形の振幅は大きく変動することから、瞬間的に目的音声よりも他話者の音声のほうがエネルギーが大きくなる状態が起こる。提案法は時間-周波数平面上の様々な帯域から基本周波数情報を得るが、この処理によって、他話者の基本周波数情報を抽出してしまうことも起こりうる。よって、走行自動車内雑音のような低い Gross error とはならなかった。しかし、Fine error の値をみると、他の手法が雑音の影響により推定精度が大きく低下しているのに対し、提案法は雑音の影響をあまり受けずに大きな Fine error の増加はみられなかった。このことから、提案法はデパート内雑音環境下でも高精度な推定を実現できていることがわかる。

以上の結果から、提案法は、他の基本周波数推定法以上の耐雑音性能を有しており、その推定精度は雑音の影響を受けにくく、クリーンな音声から得られる基本周波数と同精度の基本周波数を雑音が存在する環境においても抽出できることがわかった。

5.3　まとめ

本章では、計算機シミュレーションにより、本研究で提案した基本周波数推定法の耐雑音性能及び推定精度の検証を行った。その結果、提案法は、雑音が音声のような周期性雑音を含む場合には耐雑音性がわずかに低下するものの、白色雑音やピンク雑音、走行自動車内雑音のような非周期性雑音に対しては自己相関法やケプストラム法などの他の基本周波数推定法以上の耐雑音性能を示した。また、提案法の推定精度は雑音の影響を受けにくく、クリーンな音声から得られる基本周波数と同精度の基本周波数を雑音が存在する環境においても抽出できることを確認した。

提案法は、特に、走行自動車内雑音のように雑音のエネルギーがない周波数帯域がある場合に高い耐雑音性能を示した。実際に、実環境に存在する雑音においては、白色雑音やピンク雑音のように全ての周波数帯域に雑音が存在することは少なく、ほとんどは特定の周波数帯域に雑音のエネルギーが集中している。そのため、提案法の初期推定における時間-周波数平面上から基本周波数に関連した情報を集めるという方略は、そのような実環境雑音に適した基本周波数推定法であるといえる。

しかし、実環境には周期性の雑音が数多く存在することも確かであり、提案法の実環境における応用を推し進めるためには、周期性雑音への対応を考慮しなければならない。

第 6 章

結論

6.1 本論文の要約

音声の基本周波数は、音の高さや抑揚に対応するという点において聴覚機構を考える上で重要になるだけでなく、韻律情報の利用による音声情報処理の工学的な応用においても重要な役割を担う特徴量である。それゆえに、音声研究の初期から基本周波数推定に関して様々な手法が考えられてきた。しかし、雑音環境においては、雑音によって音声の調波構造が歪むという問題があるため正確な基本周波数を得ることは困難であり、古くからの研究課題であるにもかかわらず、決定的な基本周波数推定法は確立されていない。

そこで、本論文では、雑音環境においても頑健で高精度な基本周波数推定を目指し、

- 瞬時振幅の時間-周波数表現に現れる周期性と調波性を基にした雑音に対して頑健な基本周波数推定 (初期推定部)
- 得られた基本周波数に中心周波数を合わせた帯域幅可変櫛形フィルタによる雑音抑圧 (雑音抑圧部)
- 瞬時周波数の不動点を利用した高精度な基本周波数推定 (最終推定部)

を組み合わせた基本周波数推定法を構築した。

雑音を含む音声から直接に正確な基本周波数を抽出することは困難であるため、雑音を抑圧してから精度の高い基本周波数を抽出することを考える。ここで、雑音抑圧の際に音声の調波成分を残すために櫛形フィルタを用いることとし、櫛形フィルタの中心周波数を決めるために雑音に対して頑健な手法で大まかな基本周波数を得るという方略が本研究で構築する基本周波数推定法の大きな枠組である。

第 1 章では、従来の基本周波数推定法が時間領域に現れる基本周期に対応する特徴か周波数領域に現れる基本周波数に対応する特徴のどちらかを利用していることを示し、それぞれの代表的な手法について推定精度及び耐雑音性能の検証を行った。その結果、従来の手法が「雑音の影響を受けやすいがクリーン音声に対する推定精度が高い」か「雑音環境でも大まかな基本周波数を推定できるが雑音の増加とともに推定精度は低下する」のどちらかの傾向にあることを示した。また、実環境雑音に対応するための基本周波数推定の方策について考察した。

第 2 章では、上記の方策に基づいた提案法の概要について述べ、その詳細について第 3 章と第 4 章に示した。提案法の初期推定について述べた第 3 章では、実環境の雑音の多くが時間-周波数平面上では雑音エネルギーが偏在していることを考慮し、瞬時振幅の時間-周波数表現を用いて様々な周波数帯域から基本周波数情報を集めることにより、雑音のエネルギーの少ない帯域を有効に利用する方法を構築した。また、基本周波数情報を集める際に、従来の基本周波数推定法が時間情報か周波数情報のどちらかのみを用いていたのに対し、時間情報である周期性の特徴と周波数情報である調波性の特徴を両方用いることによって、雑音に対して頑健性を高めることを目指した。その結果、この初期推定が雑音に対して頑健な推定となっていることと、周期性と調波性の両方を用いることが雑音に対する頑健性を強めることを示した。櫛形フィルタによる雑音抑圧について述べた第 4 章では、通過帯域幅を容易に変えることができる帯域幅可変櫛形フィルタを構築し、初期推定誤差と帯域幅の関係について検証を行った。また、櫛形フィルタの帯域幅の決定方法について検証し、帯域幅を初期推定誤差によって変えることにより効果的に雑音抑圧ができることを示した。

第 5 章では、計算機シミュレーションにより提案法の耐雑音性能及び推定精度の検証を行った。その結果から、提案法が、周期性雑音に対しては耐雑音性が低下するものの、非周期性雑音に対しては他の基本周波数推定法以上の耐雑音性能を示すことを確認した。また、提案法の推定精度は雑音の影響を受けにくく、提案法がクリーンな音声から得られる基本周波数と同精度の基本周波数を雑音が存在する環境においても抽出できることを示した。

雑音環境においても音声の基本周波数を推定できるという本研究の成果は、音声情報処理の幅広い分野に応用可能である。例えば、一般に雑音環境での自動音声認識では認識精度が大幅に低下することが問題とされているが、基本周波数を句や単語の境界区分の推定に利用したり、韻律辞書により基本周波数の概形を特徴量として認識器に用いることにより認識精度の向上を図ることができる。また、基本周波数推定の精度が合成音声の自然性を左右する音声分析合成符号化においては、雑音の存在する実環境でのシステムの運用に貢献できる。すなわち、これまで接話マイクロホンでなければ活用できなかった状況から、ハンズフリーホンのようにマイクとの距離を意識する必要のない状況へ応用することができる。さ

らに、様々な音が混じりあった状態から目的音声を抽出するカクテルパーティ効果の概念を基にした音源分離では、基本周波数を音源の違いを示す特徴として用いているため、本研究は聴覚情景解析の研究に対しても貢献できる。

6.2 今後の課題

本論文で提案した基本周波数推定法を実環境で用いるには、さらに以下に挙げる課題を解決することが望まれる。

1. 周期性雑音・複数音源への対応

   提案法では周期性雑音への対策を考慮していないため、周期性雑音に対しては耐雑音性能が低下してしまう。実環境には周期性の雑音も多く存在するため、提案法の実環境における応用には、周期性雑音への対応を考慮しなければならない。周期性雑音において問題となるのは、周期性雑音自体も基本周波数をもつという点である。基本周波数推定の応用範囲を考えると、周期性雑音の基本周波数を不必要な情報とする枠組みよりも複数音源への対応へと拡張したほうが良い。

   複数音源への対応とは、同時刻に存在する複数音源の基本周波数をすべて推定することを意味する。例えば、日常会話においては話者 A の発声の最後と話者 B の発声の最初が重なることはよくあることである。このとき、重なった区間においてはそれぞれの話者の基本周波数が抽出できなければならない。2つの音の基本周波数を推定する手法はいくつか提案されているが、提案法で複数音源に対応するためには、初期推定部と雑音抑圧部を用い、雑音抑圧部によって得られた推定雑音波形に対して基本周波数推定を行う手法が考えられる。

2. 残響環境への対応

   実環境には雑音のみならず残響という問題がある。残響に対しては、相関が非常に高い信号がわずかに時間遅れを伴って重畳されるという特性から、目的音声と相関の低い雑音を対象としている本論文の提案法とは違った枠組みで考えなければならないだろう。残響環境下では、本研究で用いた時間領域

の特徴量に関しては、基本周波数が時間的に変化する上に波形が準周期的であることから、相関が高いもののわずかに異なる信号が重畳されるために波形が歪んでしまう。周波数領域の特徴量に関しては、残響が時間遅れを伴っていることから調波成分が時間的に後方へ伸びたかたちとなるために、本来の音声の調波成分と干渉を起こし、調波成分がぼやけてしまうという問題がある。残響環境への対応には、残響の影響を受けずに基本周波数情報を保持できる特徴量を検討しなければならない。

3. ミッシング・ファンダメンタルへの対応

ヒトは基本周波数成分が欠けた音 (ミッシング・ファンダメンタル) に対しても、音の高さを知覚できる。例えば、アナログの電話帯域は 0.3–3.4 kHz であり、300 Hz 以下に基本周波数成分が存在する音声は、アナログ電話では基本周波数成分が伝送されていない。このような音声に対してヒトがどのように音の高さを知覚しているかについては様々なモデルが考案されているが、一般に、高調波成分の周波数間隔から音の高さを知覚していると考えられている。提案法の初期推定部では、瞬時振幅の時間-周波数表現を用い高調波成分から基本周波数推定を行っているため、ミッシング・ファンダメンタルにも対応可能であると考えられる。

付録
代表的な従来の基本周波数推定法

A　時間領域に現れる周期性の特徴を利用する基本周波数推定法

有声音の音声波形は、準周期的な信号となっている。この信号の繰返し周期が基本周期であり、基本周波数の逆数に対応する。時間領域に現れる周期性の特徴を利用する方法とは、この波形の繰返し周期を検出することに他ならない。周期性検出には自己相関関数や自己相関関数の変形が用いられることが多く、本節で取り上げる手法も音声波形の自己相関処理を基としている。

A.1　複数窓幅から得られた自己相関関数を用いる推定法

一般に、基本周波数推定において半ピッチや倍ピッチの抽出誤りを避けるには、基本周期の 3–4 倍程度の時間長に相当する長さの分析窓を用いると良い結果が得られる。しかし、音声の基本周波数の存在範囲は広く、固定した窓長では任意の入力音声を扱うことは困難である。

都木らは複数の異なる長さの分析窓で切り出した波形に対してそれぞれ自己相関処理を行い、最適なものを選択する基本周期推定法を提案した [26]。複数の窓長を用いることで得られた基本周期候補のなかには、抽出しようとする基本周期に対して最適な窓長から得られたものが存在すると考えられる。図 A.1 に処理の流れを示す。

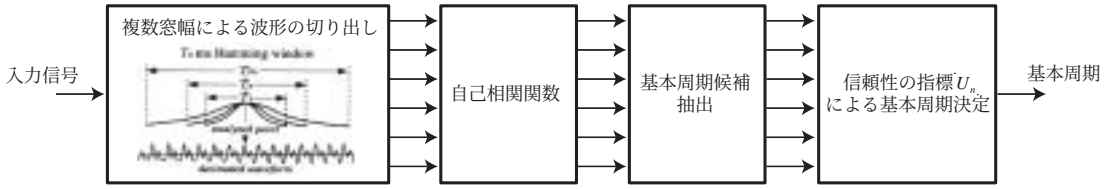

図 A.1: 複数窓幅から得られた自己相関関数を用いる基本周期推定法 (一部、都木 [26] より引用)

分析窓の数を N_w とし、窓幅 (窓の長さ)T_n (ms) が

$$T_n = 15.0 + (n-1)\{41.0/(N_w - 1) + 0.5\}, \quad (1 \leq n \leq N_w) \tag{1}$$

となるハミング窓を用いて波形 x_n を切り出す[1]。ここで、n は窓番号に対応する。都木らは予備実験から式 (1) を定めている。切り出した波形から次式の自己相関関数 $R_n(k)$ を求める。

$$R_n(k) = \sum_{i=0}^{L_n - k - 1} x_n(i) x_n(i+k)/(L_n - k) \tag{2}$$

ここで、L_n は窓の長さに対応するサンプリング点数である。次に $k_{min}(n) \leq k \leq k_{max}(n)$ の範囲で $R_n(k)$ が最大となる k を検索し、基本周期の候補 k_n とする。ここで、$k_{min}(n)$, $k_{max}(n)$ は以下のように定義される[2]。

$$k_{min}(n) = 0.15 \cdot L_n \cdot (n-1)/(N_w - 1) \tag{3}$$

$$k_{max}(n) = L_n/2 \tag{4}$$

次に k_n の周期性の程度を示す量 V_n を

$$V_n = R_n(k_n)/R_n(0) \tag{5}$$

[1]ただし、都木らは $N_w = 1$ のときは $T_1 = 35.0$ ms、$N_w = 2$ のときは $T_1 = 25.0$ ms, $T_2 = 45.0$ ms、$N_w = 3$ のときは $T_1 = 21.0$ ms, $T_2 = 36.0$ ms, $T_3 = 51.0$ ms とし、N_w が 4 以上においては T_n の下限値を 15 ms に固定している。これは極端に短い窓幅や長い窓幅では良い結果が得られないためである。

[2]ただし、$k_{min}(n)$ に関しては $N_w = 1$ では 1.25 ms に相当する値、$N_w > 1$ では 1.25 ms に相当する値を下限値としている。また、$k_{max}(n)$ に関しては $n = N_w$ のときは常に 28.6 ms に相当する値としている。

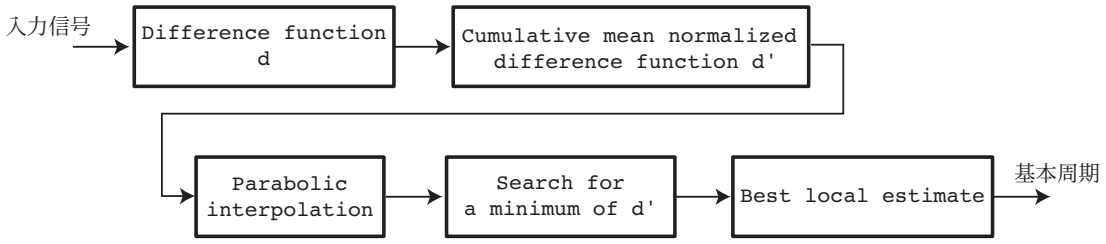

図 A.2: YIN

として求め、k_n の信頼性を示す指標として、次式に示す U_n を定義する。

$$U_n = V_n + \sum_{j=1}^{N_w} g_{nj} \cdot V_j \tag{6}$$

$$g_{nj} = \begin{cases} 1, & (r_{nj} \leq 0.1) \\ (0.25 - r_{nj})/0.15, & (0.1 < r_{nj} \leq 0.25) \\ 0, & (0.25 < r_{nj}) \end{cases} \tag{7}$$

$$r_{nj} = |k_j/k_n - 1| \tag{8}$$

このとき、U_n の最大値を与える n を n_{max} として、$k_{n_{max}}$ が基本周期として推定される。この手法では窓幅数 N_w が増加するにつれて推定精度が向上する傾向がみられる。

A.2 YIN

de Cheveigné and Kawahara は振幅差関数による基本周波数推定法 (YIN) を提案している [28]。これは、同時発声された音声の抽出に対する聴覚神経のキャンセレーションモデルが基になっている [34, 92]。図 A.2 に処理の流れを示す。

信号 $x(t)$ が周期 T の周期信号であれば、T だけシフトさせた信号との差分は

$$x(t) - x(t+T) = 0 \tag{9}$$

となる。これを利用し、周期検出のための振幅差関数 (difference function) を以下のように定義する。

$$d_t(\tau) = \sum_{j=1}^{W} \{x(j) - x(j+\tau)\}^2 \tag{10}$$

この振幅差関数を用いることは自己相関関数とほぼ等しい処理となるが、自己相関関数に比べ周期間の振幅の変化の影響を受けにくい。$d_t(\tau)$ は基本周期に対応す

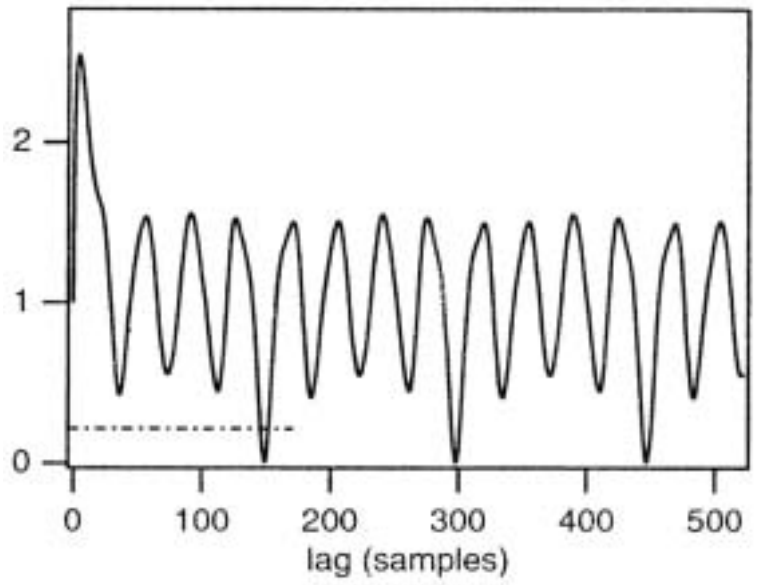

図 A.3: Cumulative mean normalized difference function (de Cheveigné, Kawahara[28])

る τ で深いディップを持つ。しかし、式 (10) は遅れ 0 では常に 0 となるが、信号が完全な周期信号ではない場合にはその遅れが周期 T に相当していても $d_t(T) = 0$ にはならない。よって、最小値探索の際に τ の下限を定めなければ、$d_t(0)$ が最小となってしまう。また下限を定めても、第一ホルマントでの強い共振により別のディップができ、時には周期を表わすディップよりも深くなるという問題は解決できない。従って、以下の式で正規化を行い、あらたに累積平均正規化振幅差関数 (cumulative mean normalized difference function) $d'_t(\tau)$ を定義する。

$$d'_t(\tau) = \begin{cases} 1, & \tau = 0 \\ d_t(\tau) / \left[(1/\tau) \sum_{j=1}^{\tau} d_t(j) \right], & \text{otherwise} \end{cases} \tag{11}$$

この正規化により、探索範囲の下限の設定が不要となる。図 A.3 に $d'_t(\tau)$ の例を示す。さらに、サンプリングの影響を小さくするために放物線による $d'_t(\tau)$ の補間を行う。また、倍ピッチの誤抽出を防ぐために閾値を設けた上で、$d'_t(\tau)$ の最小値を探索する。このとき推定された周期を基にさらに $[t - T_{max}/2, t + T_{max}/2]$ 内における $d'_t(\tau)$ の最小値を探索することにより、平滑化を行う。ここで、T_{max} は予想される周期の最大値である。

B　周波数領域に現れる調波性の特徴を利用する基本周波数推定法

有声音のスペクトルには、基本周波数とその高調波成分からなる調波構造が表われる。すなわち、音声の調波成分のうち一番低い周波数のものが検出できれば基本周波数がわかることになる。隣接した調波成分間の周波数差も基本周波数に等しいため、調波成分の周波数軸上の位置がわかれば基本周波数推定が可能となる。本節で取り上げる基本周波数推定法も周波数軸上に基本周波数の整数倍の調波成分が存在することを利用した手法である。

B.1　移動平均と帯域制限を用いたケプストラムによる推定法

音声スペクトルには基本周波数の整数倍の調波成分が櫛状に現れているが、声道の影響によりその振幅は平坦ではなく大きく変動している。そこで、ケプストラムや LPC などを用い、声道の影響を取り除く手法が考えられている。

Noll によるケプストラム法 [54] では、音声スペクトルを微細構造とスペクトル包絡に分けることができる。音声 $x(t)$ は、声帯による励振波 $g(t)$ と声道のインパルス応答 $h(t)$ との畳み込みとして

$$x(t) = \int_0^t g(t-\tau)h(\tau)d\tau \tag{12}$$

$$X(\omega) = G(\omega)H(\omega) \tag{13}$$

と与えられる。但し、$X(\omega), G(\omega), H(\omega)$ は $x(t), g(t), h(t)$ のフーリエ変換である。よって、音声の対数スペクトルは

$$\log|X(\omega)| = \log|G(\omega)| + \log|H(\omega)| \tag{14}$$

となる。これを逆フーリエ変換したものがケプストラム $c(\tau)$

$$c(\tau) = \mathcal{F}^{-1}[\log|X(\omega)|] = \mathcal{F}^{-1}[\log|G(\omega)|] + \mathcal{F}^{-1}[\log|H(\omega)|] \tag{15}$$

であり、ケフレンシと呼ばれる時間軸で表わされる。ここで、$\mathcal{F}^{-1}$ は逆フーリエ変換である。式 (15) の右辺第 1 項が音源情報である微細構造、すなわちスペクト

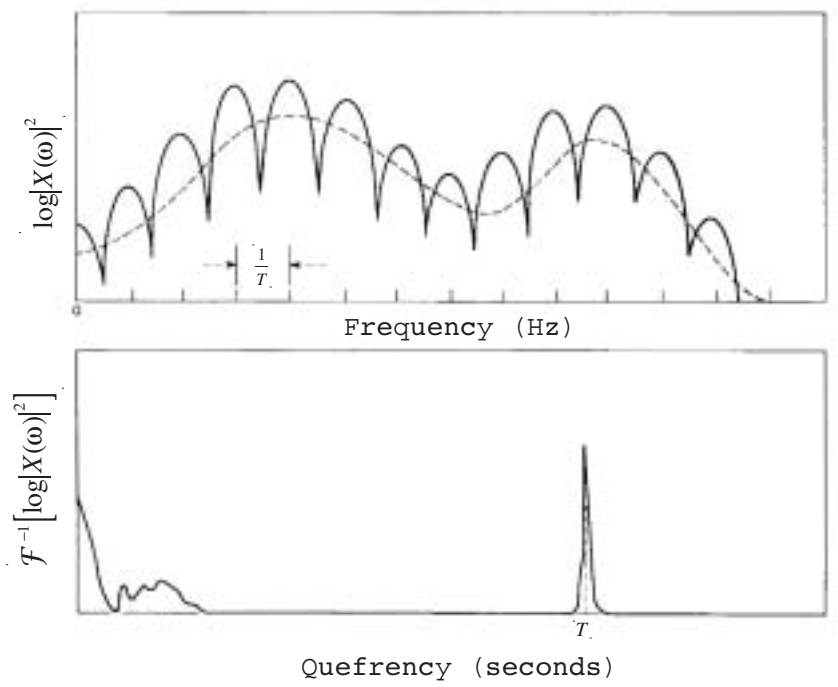

図 B.1: 有声音のケプストラム例: (上) 対数パワースペクトル (下) ケプストラム (Noll[54])

ルの細い周期のパターンであり、右辺第 2 項が声道情報であるスペクトル包絡、すなわちスペクトルの緩やかの変化のパターンとなる。第 1 項は高ケフレンシ部のピークとなり、第 2 項は低ケフレンシ部に集中する。従って、図 B.1 に示すように高ケフレンシ部のピークの位置から基本周期が得られる。

加藤と三輪は、Noll のケプストラム法を改良し、対数パワースペクトルと移動平均パワースペクトルの差分に帯域制限処理を行なう基本周波数推定法を提案した [55]。図 B.2 に処理の流れを示す。先に述べた Noll のケプストラム法では、低いホルマント周波数の影響や高周波数領域に現れる調波の乱れの影響が問題となる。そこで、低いホルマント周波数の影響を除去するために、対数パワースペクトル $\log|X(\omega)|$ と平滑化した移動平均パワースペクトルとの差分スペクトルを用いる。また、帯域制限処理により高周波数領域の調波の乱れを取り除く。この差分スペクトルを逆フーリエ変換することによるケプストラムが得られ、高ケフレンシ部よりピークを抽出して基本周波数を求める。

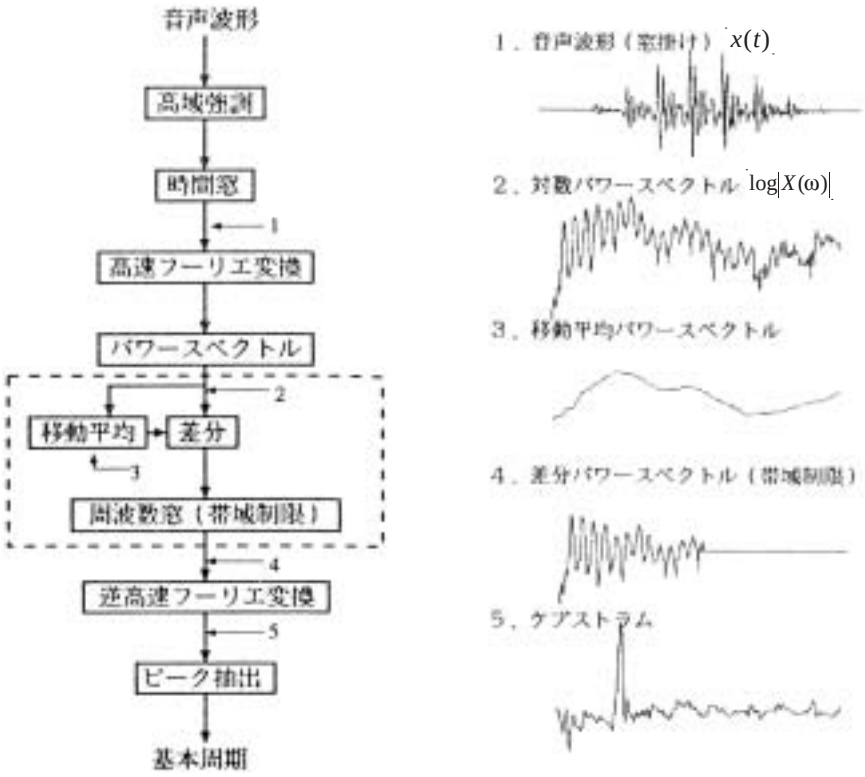

図 B.2: 移動平均と帯域制限を用いた基本周波数推定法 (加藤, 三輪 [55])

B.2 STRAIGHT-TEMPO

Kawahara らは音声分析合成システム STRAIGHT に用いる基本周波数推定法として、フィルタの中心周波数からフィルタ出力の瞬時周波数への写像の不動点を利用した手法 (STRAIGHT-TEMPO) を提案している [62]。処理の流れを図 B.3 に示す。

信号 $x(t)$ の瞬時周波数 $\omega(t)$ はヒルベルト変換 $H[x(t)]$ を用いて次のように定義される。

$$\omega(t) = \frac{d\phi(t)}{dt} \tag{16}$$

$$\phi(t) = \arg[x(t) + jH[x(t)]] \tag{17}$$

ここで、$\phi(t)$ は瞬時位相を表している。河原らの手法では、キャリア周波数 λ_c に関して時間軸方向に η 倍だけ伸長した Gabor 関数 $w(t, \lambda_c)$ とその周波数の逆数の 2 倍の寸法の 2 次のカーディナルスプライン関数 $h(t, \lambda_c)$ を畳み込んで作成した次

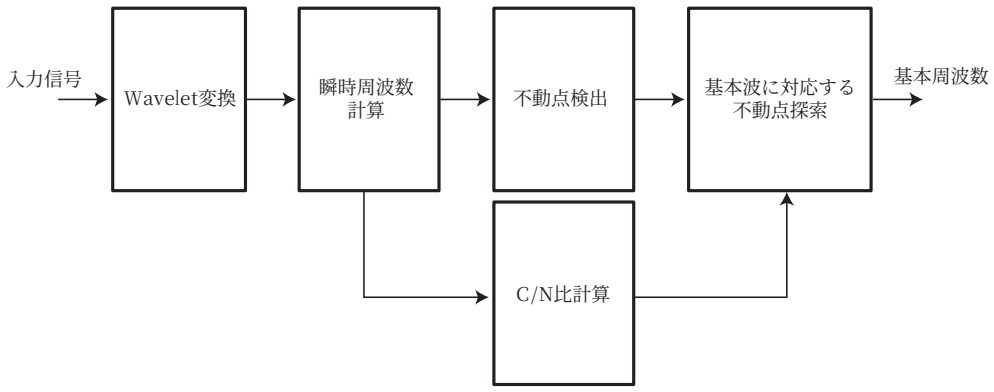

図 B.3: STRAIGHT-TEMPO (Kawahara ら [62])

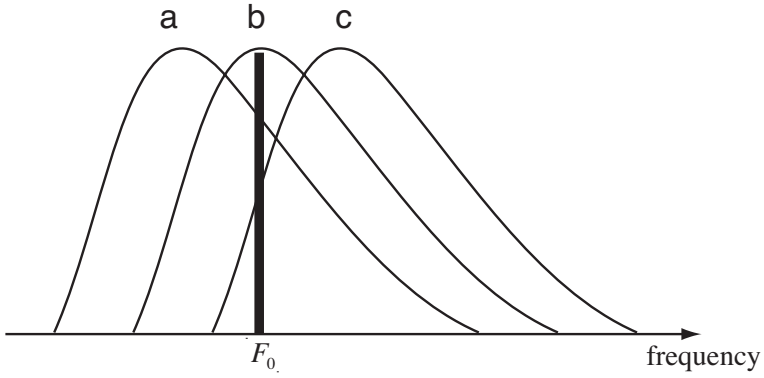

図 B.4: 複数のフィルタに入る基本波

のような関数 $w_s(t, \lambda_c)$ を用いて、$x(t)$ の wavelet 解析を行なう。

$$w_s(t, \lambda_c) = w(t, \lambda_c) * h(t, \lambda_c) \tag{18}$$

$$w(t, \lambda_c) = e^{-\frac{\lambda_c^2 t^2}{4\pi\eta^2}} e^{j\lambda_c t} \tag{19}$$

$$h(t, \lambda_c) = \max\left\{0, 1 - \left|\frac{\lambda_c t}{2\pi\eta}\right|\right\} \tag{20}$$

ここで、演算子 $*$ は畳み込みである。

音声において基本波は優勢な成分であるので、図 B.4 に示すように中心周波数が基本周波数に近い複数のフィルタに最も主要な成分として基本波の成分が入ることになる。そのため、ウェーブレットの出力の瞬時周波数 $\omega(t, \lambda)$ は基本波の値 (あるいは高調波) の値に近い値となる。また、キャリア周波数 (フィルタの中心周波数)λ_c との関係を調べると、図 B.5 に示すようにグラフは階段状になり、基本周波数やその高調波に対応する値で両者が一致する。この一致点を不動点と呼び、

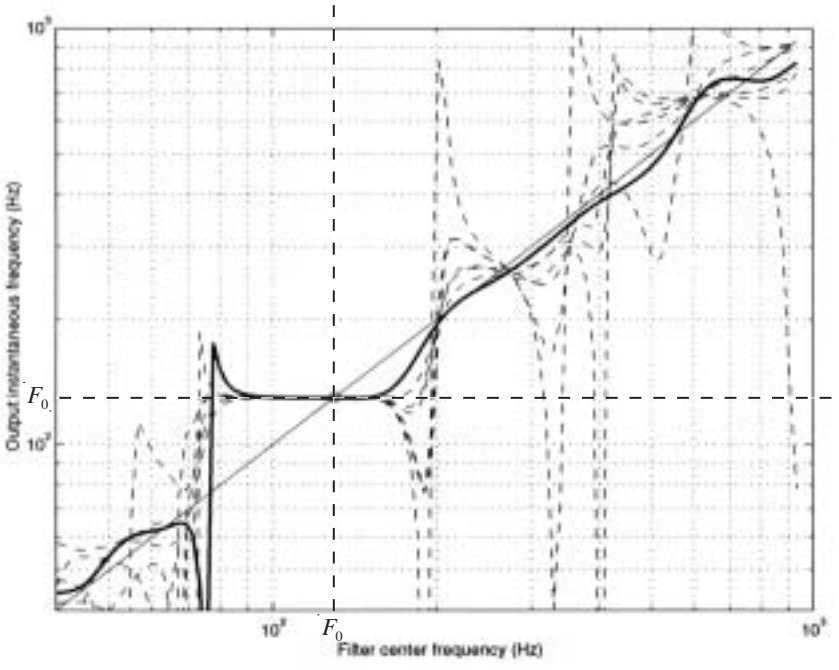

図 B.5: フィルタの中心周波数とフィルタ出力の瞬時周波数の関係 (Kawahara ら [62])

基本周波数に対応する不動点を検出できれば、基本周波数が推定できる。ここで、キャリア周波数が基本周波数にほぼ一致するスケールのウェーブレット以外では主要な正弦波成分とそれ以外の成分のレベル比として定義される C/N 比 (carrier to noise ratio) が増加することから、この C/N 比が小さいような周波数範囲に存在する不動点を基本周波数を表す点として抽出する。C/N 比は、不動点の周波数を λ_o としたとき、相対雑音エネルギ $\tilde{\sigma}^2(t)$

$$\tilde{\sigma}^2(t) = c_a \left(\frac{\partial \omega(t,\lambda)}{\partial \lambda} \right)^2 + c_b \left(\frac{\partial^2 \omega(t,\lambda)}{\partial t \partial \lambda} \right)^2 \tag{21}$$

$$c_a = \frac{1}{\displaystyle\int_{-\infty}^{\infty} \left(\lambda_o \left. \frac{dg(\lambda)}{d\lambda} \right|_{\lambda=\lambda_o} \right)^2 d\lambda_o} \tag{22}$$

$$c_b = \frac{1}{\displaystyle\int_{-\infty}^{\infty} \left(\lambda_o^2 \left. \frac{dg(\lambda)}{d\lambda} \right|_{\lambda=\lambda_o} \right)^2 d\lambda_o} \tag{23}$$

を利用して以下のように求められる $\bar{\sigma}$ を用いて C/N=$1/\bar{\sigma}$ として近似的に推定す

ることができる。

$$\bar{\sigma}^2(t,\lambda) = \int_{-T_w}^{T_w} |w(\tau,\lambda)|\tilde{\sigma}^2(t-\tau,\lambda)d\tau \tag{24}$$

ここで、T_w は関数 $|w(\tau,\lambda)|$ が 0 でない範囲を覆うことができるように設定する。

謝辞

本研究を行なうにあたり、終始御指導を賜わった北陸先端科学技術大学院大学 情報科学研究科 赤木正人 教授に深く感謝致します。

本論文をまとめるにあたり、草稿の段階から貴重な御助言と御指導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 党建武 助教授、情報科学研究科 小谷一孔 助教授、情報科学研究科 下平博 助教授、東京工科大学 メディア学部 相川清明 教授に心より感謝致します。

また、日頃から熱心な御指導ならびに御助言をいただき、多面に渡って励ましていただい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 助手に深く感謝致します。

筆者がNTTコミュニケーション科学基礎研究所に学外実習生として勤務したとき、有益な御助言を賜りましたNTTコミュニケーション科学基礎研究所 石塚健太郎 氏に心より感謝致します。

本研究の一部は、科学技術振興財団による戦略的基礎推進事業(CREST)による援助を受けて行われました。感謝の意を表します。

最後に、本論文をまとめるに当たって熱心な議論と激励をいただきました音情報処理学講座の諸兄に厚く御礼申し上げます。

参考文献

[1] 城戸 健一, 曽根 敏夫, 柴山 乾夫, 山口 公典, 中鉢 憲賢, “基礎音響工学,” コロナ社, 1990.

[2] 斎藤 収三, 中田 和男, “音声情報処理の基礎,” オーム社, 1981.

[3] 電子情報通信学会編, 新版聴覚と音声, コロナ社, 1980.

[4] 甘利 俊一, 中川 聖一, 鹿野 清宏, 東倉 洋一, “音声・聴覚と神経回路網モデル,” オーム社, 1990.

[5] H. Singer, S. Sagayama, “Pitch dependent phone modeling for HMM based speech recognition,” Proc. ICASSP92, vol.1, pp. 273–276, 1992.

[6] 川崎 真護, 中井 満, 下平 博, “F_0 生成モデルに基づくピッチパターン整合を用いた雑音重畳単語音声の認識,” 日本音響学会 平成 10 年度春期研究発表会講演論文集, pp. 109–110, 1998.

[7] 岩野 公司, 関 高浩, 古井 貞熙, “雑音に頑健な基本周波数抽出法とその音声認識への適用,” 電子情報通信学会技術報告, SP2002-13, 2002.

[8] H. Dudley, “Remaking speech,” J. Acoust. Soc. Amer, vol. 11, pp. 167–177, 1939.

[9] M. R. Schroeder, “Vocoders: Analysis and synthesis of speech,” Proc. of IEEE, vol. 54, no. 5, pp. 720-734, 1966.

[10] 板倉 文忠, “スペクトル符号化に基づく音声分析合成,” 音響学会誌, vol. 37, no. 5, pp. 197–203, 1981.

[11] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication 27, pp. 187–207, 1999.

[12] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, MA, 1990.

[13] T. Nakatani, T. Kawabata, H. G. Okuno, "A computational model of sound stream segregation with multi-agent paradigm," Proc. ICASSP95, vol.4, pp. 2671–2674, 1995.

[14] M. Unoki, M. Akagi, "Signal extraction from noisy signal based on auditory scene analysis," Proc. ICSLP98, vol.4, pp. 1515–1518, 1998.

[15] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," IEEE Trans. Acoust., Speech & Signal Process., vol. 24, no. 5, 1976.

[16] 古井 貞熙, "ディジタル音声処理," 東海大学出版会, 1985.

[17] 鈴木 久喜, "ピッチ抽出の今昔", 日本音響学会誌 56 巻 2 号, pp. 121–128, 2000.

[18] 斉藤 収三, 加藤 勝洋, 寺西 昇, "音声の基本周波数の特性について," 日本音響学会誌, vol. 14. no. 2, pp. 111–116, 1958.

[19] W. Hess, *Pitch determination of speech signals*, Springer-Verlag, Berlin, 1983.

[20] W. J. Hess, *Pitch and voicing determination*, In: Furui, S., Sondhi, M.M. (Eds.), *Advances in speech signal processing.* Marcel Dekker, New York, 1992.

[21] D. J. Hermes, D.J., *Pitch analysis*, In: Cooke, M., Beet, S., Crawford, M. (Eds.), *Visual representations of speech signals.*, John Wiley & Sons, Chichester, 1993.

[22] B. Gold, L. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," J. Acoust. Soc. Am., vol. 46, no. 2, pp. 442–448, 1969.

[23] N. C. Geçkinli, D. Yavuz, "Algorithm for pitch extraction using zero-crossing interval sequence," IEEE Trans. Acoust., Speech & Signal Process., vol. 25, no. 6, 1977.

[24] M. M. Sondhi, "New methods of pitch extraction," IEEE Trans. Audio, Electroacoust., AU-16, no. 2, pp. 262–266, 1968.

[25] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, H. J. Manley, "Average magnitude difference function pitch extraction," IEEE Trans. Acoust., Speech, Signal Process. ASSP-22, pp. 353–361, 1974.

[26] 都木 徹, 清山 信正, 宮坂 栄一, "複数の窓幅から得られた自己相関関数を用いる音声基本周期抽出法," 電子情報通信学会論文誌, vol.J80-A, no.9, pp. 1341–1350, 1997.

[27] T. Shimamura, H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," IEEE Trans. on Speech and Audio Processing, vol.9, No.7, pp. 727–730, 2001.

[28] A. de Cheveigné, H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., 111(4), pp. 1917–1930, 2002.

[29] B. S. Atal, S. 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.

[30] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio vol. AU-20, pp. 367–377, 1972.

[31] R. Meddis, M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: pitch identification," J. Acoust. Soc. Am., vol. 89, no. 6, pp. 2866–2882, 1991.

[32] P. Cariani, M. Tramo, B. Delgutte, "Neural representation of pitch through temporal autocorrelation," Proc. of Audio Engineering Society Meeting, Preprint #4583 (L-3), Sep. 1997.

[33] B. C. Moore, *An introduction to the psychology of hearing (4th ed.),* Academic press, London, 1997.

[34] A. de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am., vol. 93, no. 6, pp. 3271–3290, 1993.

[35] R. Meddis, "A physiological model of auditory selective attention," In *Advances in Speech, Hearing and Language Processing*, ed. W. A. Aintworth, vol. 3, part B, pp. 428–445, 1996.

[36] D. M. Howard, "Peak-picking fundamental period estimation for hearing prostheses," J. Acoust. Soc. Am., vol. 86, no. 3, pp. 902–910, 1989.

[37] F. Gaillard, F. Berthommier, G. Feng, J. L. Schwartz, "A modified zero-crossing method for pitch detection in presence of interfering sources," Proc. of Eurospeech97, vol. 1, pp. 445–448, 1997.

[38] J. J. Dobnowski, R. W. Schafer, L. R. Rabiner, "Real-time digital hardware pitch detector," IEEE Trans. Acoust., Speech & Signal Process., vol. 24, no. 1, pp. 2–8, 1976.

[39] M. Karjalainen, T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," Proc. of ICASSP99, vol. 2, pp. 923–932, 1999.

[40] J. di Martino, Y. Laprie, "An efficient F0 determination algorithm based on the implicit calculation of the autocorrelation of the temporal excitation signal," Proc. of Eurospeech99, vol. 6, pp. 2773–2776, 1999.

[41] T. Tolonen, M. Karjalainen, "A computationally efficient multipitch analysis model," IEEE Trans. Speech, Audio Process., vol. 8, no. 6, pp. 708–716, Nov. 2000.

[42] K. Kasi, S. A. Zahorian, "Yet another algorithm for pitch tracking," Proc. of ICASSP2002, vol. 1, pp. 361–364, 2002.

[43] S. Koval, V. Bekasova, M. Khitrov, A. Raev, "Pitch detection reliability assessment for forensic applications," Proc. of Eurospeech97, vol. 1, pp. 489–492, 1997.

[44] Y. R. Wang, I. J. Wong, T. C. Tsao, "A statistical pitch detection algorithm," Proc. of ICASSP2002, vol. 1, pp. 357–360, 2002.

[45] D. Chazan, Y. Stettiner, D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," Proc. of ICASSP93, vol. II, pp. 728–731, 1993.

[46] W. Zhang, G. Xu, Y. Wang, "Pitch estimation based on circular AMDF," Proc. of ICASSP2002, vol. 1, pp. 341–344, 2002.

[47] A. de Cheveigue, "Cancellation model of pitch perception," J. Acoust. Soc. Am., vol. 103, no. 3, pp. 1261–1271, March 1998.

[48] A. de Cheveigue, H. Kawahara, "Multiple period estimation and pitch perception model," Speech Communication, vol. 27, pp. 175–185, 1999.

[49] L. Cohen, *Time-frequency analysis*, Prentice Hall PTR, New Jersey, 1995.

[50] 国枝 伸行, 島村 徹也, 鈴木 誠史, "対数スペクトルの自己相関関数を利用したピッチ抽出法," 電子情報通信学会論文誌, vol. J80-A, no. 3, pp. 435–443, 1997.

[51] K. Nishi, S. Ando, "An optimal comb filter for time-varying harmonics extraction," IEICE Trans. Fundamentals, Vol.E81-A, NO.8, pp. 1622–1627, 1998.

[52] D. J. Hermes, "Measurement of pitch by subharmonic summation," J. Acoust. Soc. Am., vol. 83, no. 1, pp. 257–264, 1988.

[53] A. M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal-pitch detection," J. Acoust. Soc. Am., vol. 36, no. 2, pp. 226–302, 1964.

[54] A. M. Noll, "Cepstrum pitch determination," J. Acoust. Soc. Am., vol. 41, no. 2, pp. 293–309, 1966.

[55] 加藤 誠二, 三輪 譲二, "移動平均と帯域制限を用いたケプストラム型基本周波数抽出とその応用," 電子情報通信学会技術報告, SP94-95, 1995.

[56] A. M. Noll, "Clipstrum pitch determination," J. Acoust. Soc. Am., vol. 44, no. 6, pp. 1585–1591, 1968.

[57] 小林 戴, 島村 徹也, "対数スペクトルにクリッピングと帯域制限を用いる基本周波数抽出法," 電子情報通信学会論文誌, vol. J82-A, no. 7, pp. 1115–1122, 1999.

[58] F. J. Charpentier, "Pitch detection using the short-term phase spectrum," Proc. of ICASSP86, vol. 1, pp. 391–394, 1986.

[59] L. Qiu, H. Yang, S. N. Koh, "Fundamental frequency determination based on instantaneous frequency estimation," Signal Processing 44, pp. 233–241, 1995.

[60] T. Abe, T. Kobayashi, S. Imai, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," Proc. ICSLP96, Vol.2, pp. 1277–1280, 1996.

[61] T. Abe, T. Kobayashi, S. Imai, "The IF spectrogram: A new spectral representaion," Proc. ASVA97, pp. 423–430, 1997.

[62] H. Kawahara, H. Katayose, A. de Cheveigné, R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," Proc. Eurospeech99, pp. 2781–2784, 1999.

[63] 阿竹 義徳, 入野 俊夫, 河原 英紀, 陸 金林, 中村 哲, 鹿野 清宏, “調波成分の瞬時周波数を用いた基本周波数推定方法,” 電子情報通信学会論文誌, vol. J83-D-II, no. 11, pp. 2077–2086, 2000.

[64] T. Nakatani, T. Irino, “Robust fundamental frequency estimation against background noise and spectral distortion,” Proc. ICSLP2002, pp. 1733–1736, 2002.

[65] M. Lahat, R. J. Niederjohn, D. A. Krubsack, “A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech,” IEEE Trans. Acoust., Speech & Signal Process., vol. 35, no. 6, pp. 741–750, 1987.

[66] D. Chazan, M. Tzur (Zibulski), R. Hoory, G. Cohen, “Efficient periodicity extraction based on sine-wave representation and its application to pitch determination of speech signals,” Proc. of Eurospeech2001, pp. 2427–2430, 2001.

[67] 柳沢 浩一, 田中 京子, 山浦 逸雄, “スペクトル包絡の時間的連続性を利用した基本周期の検出法,” 電子情報通信学会論文誌, vol. J83-D-II, no. 11, pp. 2087–2098, 2000.

[68] 三輪 多恵子, 田所 嘉昭, 斎藤 務, “くし形フィルタを利用した採譜のための異楽器音中のピッチ推定,” 電子情報通信学会論文誌, vol. J81-D-II, no. 9, pp. 1965–1974, 1998.

[69] 嵯峨山 茂樹, 古井 貞熙, “ラグ窓を用いたピッチ抽出の一方法,” 電子情報通信学会全国大会予稿集, vol. 5, p. 263, 1978.

[70] E. Geoffrois, “The multi-lag-window method for robust extended-range F0 determination,” Proc. of ICSLP96, vol. 4, pp. 2239–2242, 1996.

[71] D. J. Liu, C. T. Lin, “Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure,” IEEE Trans. Speech, Audio Process., vol. 9, no. 6, pp. 609–621, 2001.

[72] R. C. Maher, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," J. Acoust. Soc. Am., vol. 95, no. 4, pp. 2254–2263, 1994.

[73] H. S. Pang, S. J. Baek, K. M. Sung, "Improved fundamental frequency estimation using parametric cubic convolution," IEICE Trans. Fundamentals, vol. E83-A, no. 12, pp. 2747–2750, 2000.

[74] 大村 浩, 田中 和世, "基本波フィルタリング法による精細ピッチパターンの抽出," 日本音響学会誌, vol. 51, no. 7, pp. 509–518, 1995.

[75] 佐宗 晃, 中村 尚五, "ウェーブレット変換を用いたピッチ抽出の一方法," 電子情報通信学会論文誌, vol. J80-A, no. 11, pp. 1848–1856, 1997.

[76] J. D. Brown, M. S. Puckette, "A high resolution fundamental frequency determination based on phase changes of the Fourier transform," J. Acoust. Soc. Am., vol. 94, no. 2, pp. 662–667, MIT Press, Cambridge, MA, 1993.

[77] T. Tanaka, T. Kobayashi, D. Arifianto, T. Masuko, "Fundamental frequency estimation based on instantaneous frequency amplitude spectrum," Proc. of ICASSP2002, vol. 1, pp. 329–332, 2002.

[78] 古井 貞熙, "音響・音声工学," 近代科学社, 1992.

[79] 川人 光男, 行場 次朗, 藤田 一郎, 乾 敏郎, 力丸 裕, "認知科学 3 視覚と聴覚," 岩波書店, 1994.

[80] 藤崎 和香, 柏野 牧夫, "絶対音感保持者の音高知覚特性," 日本音響学会誌, vol.57, no.12, pp. 759–767, 2001.

[81] 板橋 秀一, "騒音データベースと日本語共通音声データ DAT 版," 日本音響学会誌, vol.47, no.12, pp. 951–953, 1991.

[82] S. Nakamura, K. Hiyane, F. Asano, T. Endo, "Sound scene data collection in real acoustical environments," J. Acoust. Soc. Japan (E), vol. 20, no. 3, pp. 225–231, 1999.

[83] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process. ASSP-27, no. 2, pp. 113–120, 1979.

[84] K. K. Paliwal, A. Basu, "A speech enhancement method based on kalman filtering," Proc. of ICASSP87, pp. 177–180, 1987.

[85] D. C. Popescu, I. Zelijković "Kalman filtering of colored noise for speech enhancement," Proc. of ICASSP98, pp. 997–1000, 1998.

[86] G. Shafer, *A mathematical theory of evidence,* Princeton University Press, 1976.

[87] 石塚 満, "Dempster&Shafer の確率理論," 電子通信学会誌, Vol. 66, No. 9, pp. 900–903, 1983.

[88] R. D. Patterson, J. Holdsworth, *A functional model of neural activity patterns and auditory images*, In: Ainsworth, W.A., Evans, E.F., Hackney, C.M. (Eds.), *Advances in speech*, Hearing and language processing. Vol.3, JAI Press, London, 1991.

[89] B. R. Glasberg, B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," Hearing Research 47, pp. 103–138, 1990.

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

[91] 柏野 邦夫, 田中 英彦, "二つの周波数成分の分離知覚に関する工学的モデル – 複数の要因の評価と統合–", 電子情報通信学会論文誌, vol. J77-A, no. 5, pp. 731–740, 1994.

[92] A. de Cheveigné, "Concurrent vowel identification III: A neural model of harmonic interference cancellation," J. Acoust. Soc. Am., vol. 101, no. 5, pp. 2857–2865, 1997.

本研究に関する発表論文

論文

1. Y. Ishimoto, K. Ishizuka, K. Aikawa, M. Akagi, “Fundamental frequency estimation for noisy speech using entropy-weighted periodic and harmonic features,” IEICE Trans. Inf. & Syst., Vol. E87-D, No. 1, pp. 205-214, 2004.
2. Y. Ishimoto, M. Unoki, M. Akagi, “Fundamental frequency estimation for noisy speech based on instantaneous amplitude and frequency,” Speech Communication. (条件付採録, 改訂版投稿中)

国際会議

1. Y. Ishimoto, M. Akagi, “A fundamental frequency estimation method for noisy speech,” Proc. WESTPRAC VII, Vol. 1, pp.161-164, 2000.
2. Y. Ishimoto, M. Unoki, M. Akagi, “A fundamental frequency estimation method for noisy speech based on periodicity and harmonicity,” ICASSP2001, SPEECH-SF3.1,2001.
3. Y. Ishimoto, M. Unoki, M. Akagi, “A fundamental frequency estimation method for noisy speech based on instantaneous amplitude and frequency,” A workshop on consistent and reliable acoustic cues for sound analysis (CRAC), 2001.
4. Y. Ishimoto, M. Unoki, M. Akagi, “A Fundamental Frequency Estimation Method for Noisy Speech Based on Instantaneous Amplitude and Frequency,” Proc. Eurospeech2001, Vol. 4, pp. 2439-2442, 2001.

口頭発表

1. 石本 祐一, 赤木 正人, “雑音が付加された音声の基本周波数推定と雑音抑圧,” 電子情報通信学会技術報告, SP99-169, 2000.
2. 石本 祐一, 赤木 正人, “雑音中の音声基本周波数推定法の提案,” 日本音響学会 平成 12 年度春季研究発表会 講演論文集, pp.253-254, 2000.
3. 石本 祐一, 鵜木 祐史, 赤木 正人, “周期性と調波性を考慮した雑音環境における基本周波数推定,” 平成 12 年度 電気関係学会北陸支部連合大会 講演論文集, p.452, 2000.
4. 石本 祐一, 鵜木 祐史, 赤木 正人, “周期性と調波性を考慮した雑音環境における基本周波数推定,” 日本音響学会 平成 12 年度秋季研究発表会 講演論文集, pp.243-244, 2000.
5. 石本 祐一, 鵜木 祐史, 赤木 正人, “周期性と調波性を考慮した雑音環境における基本周波数推定,” 日本音響学会 聴覚研究会資料, H-2000-81, 2000.
6. 石本 祐一, 鵜木 祐史, 赤木 正人, “瞬時振幅の周期性と調波性を考慮した雑音環境における基本周波数推定,” 話し言葉の科学と工学ワークショップ 講演予稿集, pp. 149-156, 2001.
7. 石本 祐一, 鵜木 祐史, 赤木 正人, “周期性と調波性を考慮した雑音環境における基本周波数推定法の改良,” 日本音響学会 平成 13 年度春季研究発表会 講演論文集, pp. 243-244, 2001.
8. 石本 祐一, 鵜木 祐史, 赤木 正人, “周期性雑音を含む音声に対する瞬時振幅を利用した基本周波数推定法,” 日本音響学会 平成 13 年度秋季研究発表会 講演論文集, pp. 253-254, 2001.
9. 石本 祐一, 石塚 健太郎, 相川 清明, 赤木 正人, “エントロピーによる重み付けを用いた雑音環境下での基本周波数推定,” 電子情報通信学会技術報告, SP2002-52, 2002.
10. 石本 祐一, 石塚 健太郎, 相川 清明, 赤木 正人, “エントロピーで重み付けした周期性・調波性特徴を用いた雑音下 F0 推定,” 日本音響学会 平成 14 年度秋季研究発表会 講演論文集, pp. 371-372, 2002.
11. 石本 祐一, 赤木 正人, “瞬時振幅の周期性・調波性を基にした相関係数統合に

よる基本周波数推定,” 日本音響学会 平成 15 年度春季研究発表会 講演論文集, pp. 337-338, 2003.