

Title	音声明瞭度向上のための残響音声回復法に関する研究
Author(s)	戸井, 真智
Citation	
Issue Date	2005-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1918
Rights	
Description	Supervisor: 赤木 正人, 情報科学研究科, 修士

修 士 論 文

**音声明瞭度向上のための残響音声回復法
に関する研究**

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

戸井真智

2005 年 3 月

修 士 論 文

音声明瞭度向上のための残響音声回復法 に関する研究

指導教官 赤木正人 教授

審査委員主査 赤木 正人 教授

審査委員 党 健武 教授

審査委員 小谷 一孔 助教授

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

310072 戸井真智

提出年月: 2005 年 2 月

概要

本研究では、残響の影響を受けた音声から、ブラインド的に残響回復を行う手法を構築し、残響によって低下した音声明瞭度を向上させることを目標とする。

残響音声の回復は、拡声会議システムや補調システムのための音声強調や音声認識システムの前処理などの、様々な音声信号処理において重要な課題である。これまでの残響音声の回復法では、室内伝達特性の逆フィルタ処理を用いた手法が多く提案されている。しかし、これらの手法では、時変的な室内伝達特性をその都度測定する必要があるため、様々な音声アプリケーションに用いることは困難である。

酒田らは、MTFに基づいて、室内インパルス応答を測定すること無く、観測した音声情報から残響音声を回復する手法を提案した。しかし、この手法には解決すべき点がいくつか残されている。一つは、パワーエンベロープの回復部において、まだ改良の余地が考えられるということ。もう一つは、キャリア再生成部において、位相の制御を行っていないため、回復音声の品質が良くないということ。また、回復音声の明瞭度について、何も考察されていないという検討事項もある。

本研究では、この酒田らの手法の問題点を解決し、音声明瞭度を向上させるための残響回復法を提案する。音声信号をエンベロープとキャリアに分割し、それぞれに対して回復処理を行う。まず、パワーエンベロープ回復部では、酒田らの提案したパワーエンベロープ回復法に対して、適応的な時間-周波数分割処理を適用することにより、パワーエンベロープの回復精度をさらに向上できることがわかった。また、パワーエンベロープを抽出する際に利用していたカットオフ周波数についても検討を行った。MTF理論が最も成立しやすくなるようにカットオフ周波数の値を定めてやることによって、さらにパワーエンベロープの回復精度が向上した。次に、キャリア再生成部では、音源の位相情報の群遅延を制御することによって、回復音声の音質を大きく改善することができた。提案モデルがどれくらい残響回復できるかを検討するため、シミュレーションと聴取実験によって評価した。その結果、主観評価においても客観評価においても、提案モデルの高い有効性が示され、音声の明瞭度が向上する結果が得られた。

目次

第1章 序論	1
1.1 背景	1
1.2 目的	2
1.3 本論文の構成	3
第2章 残響音声回復モデルの概要	4
2.1 パワーエンベロープ回復部	4
2.2 キャリア再生成部	4
第3章 パワーエンベロープ回復法	6
3.1 パワーエンベロープ回復法の原理	6
3.1.1 変調伝達関数 (MTF)	6
3.1.2 パワーエンベロープ逆フィルタ法	7
3.1.3 パラメータの決定法	8
3.1.4 パワーエンベロープ逆フィルタ法の問題点	9
3.2 時間-周波数分割処理	10
3.2.1 時間分割処理	10
3.2.2 最適な閾値の決定	10
3.2.3 周波数分割処理	11
3.2.4 最適な相関値の決定	12
3.3 時間-周波数分割処理の評価	12
第4章 カットオフ周波数の検討	23
4.1 最適なエンベロープ抽出を行うカットオフ周波数	23
4.2 カットオフ周波数の導出過程	23
4.3 評価シミュレーション	26
第5章 キャリア再生成法	27
5.1 キャリア再生成法の原理	27
5.1.1 原理の説明	27
5.1.2 キャリア再生成法の問題点	28

5.2	キャリア再生成法の改良	28
第 6 章	提案モデルの主観的評価	31
6.1	実験条件	31
6.2	音声品質に対する評価	31
6.3	残響感に対する評価	32
6.4	音声明瞭度に対する評価	33
6.5	まとめ	34
第 7 章	結論	35
7.1	本論文で明らかにしたこと	35
7.2	今後の課題	35
	謝辞	38
	参考文献	39
	研究業績	42

目 次

2.1	残響音声回復モデル.	5
3.1	MTF に基づいたパワーエンベロープの関係：(a) 原信号 $x(t)$ と (b) パワーエンベロープ $e_x(t)^2$, (c) 室内インパルス応答 $h(t)$ と (d) パワーエンベロープ $e_h(t)$, (e) 残響信号 $y(t)$ と (f) パワーエンベロープ $e_y(t)$, (g) 残響回復されたパワーエンベロープ $\hat{e}_x(t)^2$	14
3.2	パワーエンベロープ逆フィルタ処理の流れ.	15
3.3	時間-周波数分割処理によるパワーエンベロープ.	15
3.4	逆フィルタの (a) 一括処理と (b) 時間分割処理, 太線：原音声パワーエンベロープ $e_x(t)^2$, 細線：回復パワーエンベロープ $\hat{e}_x(t)^2$, 破線：残響パワーエンベロープ $e_y(t)^2$	16
3.5	閾値 θ の決定：従来法からの (a) 相関と (b) SNR の平均改善量.	17
3.6	フィルタバンクにおける音声信号の SNR 回復精度: (a) 一括処理 (b) 時間分割処理.	18
3.7	帯域分割幅の決定：(a) 相関図と (b) 帯域分割幅.	19
3.8	相関図：相関値が (a) 0.9, (b) 0.85 の場合.	19
3.9	相関値の決定：(a) 相関と (b) SNR の平均改善度.	20
3.10	帯域分割法の比較：(a) 相関と (b) SNR.	21
3.11	帯域分割法の比較：(a) 相関と (b) SNR の平均改善度.	22
3.12	LSD 改善度.	22
4.1	カットオフ周波数の例 (mau, /aikawarazu/, $T_R = 2.0$ s).	24
4.2	カットオフ周波数の比較：カットオフ周波数 $f_c = 20$ Hz を用いた場合の (a) 相関と (c) SNR の改善度, 可変カットオフ周波数を用いた場合の (b) 相関と (d) SNR の改善度.	25
4.3	評価シミュレーション：(a) 相関と (b) SNR の平均改善度, Imp: 提案法 (可変カットオフ周波数), Prv: 従来法 (カットオフ周波数 $f_c = 20$ Hz).	26
5.1	PIFM モデル.	28
5.2	キャリア再生成処理の改良モデル.	30
6.1	音声品質, Prev: 従来法, Imp: 提案法, Org: 原音声	32

6.2	残響感, Rev:残響音声, Prev:従来法, Imp:提案法, Org:原音声	32
6.3	単語了解度試験.	33
7.1	mis,/tarukamu/のスペクトルグラム:(a) 原音声エンベロープ, (b) 残響音声エンベロープ, (c) 回復音声エンベロープ, (d) チューニングされたエンベロープを用いて作成された音声	37

第1章 序論

1.1 背景

実環境で観測される音声信号は、周囲の雑音や残響の影響により歪んでしまう。そのため、音声認識システムや補聴システムなどの音声信号処理を実環境で行う場合、これらの影響を取り除く必要がある。特に残響環境下では、壁や天井などによる反射音が直接音に歪みを与えるため、音声の明瞭度を著しく低下させるだけでなく、音声認識における認識精度低下の原因にもなっている。そのため、これらの問題点を改善するためには、残響の影響を除去し、明瞭度の高い音声を生成する必要がある。また、残響の影響は環境の様々な変化（室温や室内形状など）と共に変動するため、これを考慮して適応的に残響音声を回復する必要がある。

これまでに、室内音場で観測された残響音声から原音声を回復する逆フィルタ処理が数多く検討されている。例えば、Neely と Allen は、単一マイクロホンで受音された信号から室内伝達特性の最小位相成分のみを取り除いて回復させる方法を提案している [4]。しかしこの手法は、室内音場が最小位相特性を持つときのみしか有効でない。これに対し、三好と金田は、音源の数に対しマイクロホンを1つ以上多く配置することで、室内インパルス応答が最小位相特性を有していなくても音源そのものを正確に復元できる音場逆フィルタ理論を提唱している [5]。それぞれの音源からマイクロホンに至る伝達関数の零点が重複していなければ、伝達系が最小位相を有していなくても音源信号を正確に復元できる。また、王と板倉は、それぞれの帯域ごとに音源信号と回復信号の誤差が最小になる最適なインパルス応答の逆フィルタを求め、音源波形そのものを復元する帯域分割逆フィルタ理論を提唱している [6]。これらの伝達特性の逆特性を逆フィルタ処理で実現する方法では、逆フィルタを決定する上であらかじめ残響のインパルス応答を測定しておかなければならない。しかし、室内のインパルス応答は室内環境の様々な変動（室内の形状や温度など）に伴い時々刻々と変動するため、その都度伝達系を測定し直さなければいけない。これは様々な音声アプリケーションにこれらの手法を用いるためには重要な問題となる。

一方、室内のインパルス応答を測定することなく、観測した残響音声情報からエンベロープを回復させる、時間波形エンベロープ逆フィルタ法も提案されている。例えば、Langhaus と Strube [8] や Avendano と Hermansky [9] の手法である。これらは、変調伝達関数 (MTF) の概念 [1] に基づいて、短時間フーリエ変換 (STFT) によるパワースpekトルの高域通過処理を用いて、時間波形エンベロープの逆特性を畳み込むという手法である。また、Mourjopoulos と Hammond [10] や 広林ら [11] の手法もある。これらの手法も

MTF の概念に基づいており、STFTではなく、1つ以上のチャンネルのフィルタバンクを用いて、残響音声から時間波形エンベロープを回復する手法である。これらの手法により、室内インパルス応答を測定することなく残響信号のエンベロープを回復することができる。しかし、広林らの手法では、残響信号からのエンベロープ抽出法や、残響時間の推定法などが問題点として挙げられる。また古川らは、広林らの問題点に対する検討および、解決方法を提案した [13, 17]。これは、観測された残響音声の情報のみから残響時間を推定し、パワーエンベロープを回復させる手法である。しかしこれらのパワーエンベロープのみの回復法では、音声明瞭度はあまり向上しないとされている。これは、音声のキャリア部も残響の影響を受けているからである。

これに対して酒田らは、この MTF に基づいたパワーエンベロープ逆フィルタ法を用いて、残響音声から回復音声を波形レベルで再構築する手法を提案した [14, 16]。これは、広林らや古川らの提案したパワーエンベロープの回復に加えて、音声のキャリア情報も再構築することによって、回復音声を時間波形として出力する手法である。この手法により、残響感は解消できたが、回復音声の音質が悪く、また音声明瞭度も低いと予想される。これら以外にも、パワーエンベロープ回復処理にまだ改善の余地が考えられる点や、F0 も既知として処理を行っているといった課題も挙げられる。

1.2 目的

本研究の目的は、時変的な性質を持つ室内伝達系のインパルス応答を測定することなく、ブラインド的に残響音声を回復する手法を構築し、残響によって低下した音声明瞭度を向上することである。本研究では、室内インパルス応答を測定することなく回復処理を行うことができる酒田らの手法をベースに、残響回復モデルの構築を行う。さらにこの手法は、パワーエンベロープの回復だけでなく、残響の影響を受けたキャリアに対しても回復処理を行うので、音声明瞭度を向上できる可能性があると考えられる。

本研究のモデルでは、音声信号をパワーエンベロープとキャリアに分けてモデル化し、それぞれに対して回復処理を行う。パワーエンベロープ回復部では、酒田らの提案したパワーエンベロープ逆フィルタ法に対して、適応的な時間-周波数分割処理を適用することにより、エンベロープの回復精度の向上を目指す。また、音声信号からパワーエンベロープを抽出する際に利用していたカットオフ周波数に対して、より MTF 理論が成立するような値を求めてやることによって、パワーエンベロープの回復精度をさらに向上させることを試みる。次に、キャリア再生部では、音源の位相情報を制御することによって、回復音声の音質を改善する。最終的に、本研究で提案した手法を総合的に評価するため、シミュレーションによる客観的評価に加えて、聴取実験による主観的評価も行う。この主観評価実験によって、提案法の音声明瞭度を評価する。

この手法により残響音声回復が可能となれば、補聴システムや遠隔会議システムなどでの明瞭度の向上や音声認識における認識精度の向上などに貢献できる。

1.3 本論文の構成

本論文は全7章により構成されている。以下に各章の概要を述べる。第2章では、提案モデルの概要について説明する。第3章では提案モデルのパワーエンベロープ回復法の原理の説明とその問題点について述べ、改良法である時間-周波数分割処理の原理について説明する。第4章では、パワーエンベロープを音声から抽出する際に利用するカットオフ周波数の検討を行う。第5章では、キャリア再生処理の原理の説明とその問題点について述べ、その改良法について検討する。第6章では、提案モデルの有効性を示すため、主観評価実験を行い結果の考察を行う。第7章では、本研究のまとめと今後の課題について述べる。

第2章 残響音声回復モデルの概要

本研究で用いる残響音声回復モデル [20] を図 2.1 に示す. このモデルはパワーエンベロープ回復部とキャリア再生成部の 2 つの部分で構成される. まず, N チャンネルのフィルタバンクを通した残響音声信号をエンベロープとキャリアに分解する. その後, パワーエンベロープ回復部では, パワーエンベロープ逆フィルタ法を用いて, 残響音声エンベロープからパワーエンベロープを回復させる. キャリア再生成部では, 残響音声中から推定した基本周波数 (F_0) を基に, 有聲/無聲音ごとにキャリアの再生成処理を行う. 最後に, 回復されたエンベロープと再生成されたキャリアを掛け合わせ各チャンネル信号を作り, フィルタバンクによって残響信号から回復した音声を再構築する.

2.1 パワーエンベロープ回復部

ここでは, 帯域分割された残響音声からパワーエンベロープを抽出し, パワーエンベロープ逆フィルタ処理を行う. 本研究では, まず帯域分割の際に, 各帯域のエンベロープの相関を計算し, 共変調の性質に基づいて, 帯域幅を入力された音声のエンベロープに依存して適応的に計算する. その後, 各チャンネルのエンベロープに対して, 無音区間を検出し, その前後の区間でそれぞれ残響時間パラメータを推定する. その残響時間を基にパワーエンベロープ逆フィルタ処理を各区間ごとに行う.

2.2 キャリア再生成部

有聲音区間では, 推定された F_0 を基に, その高調波を足し合わせて有聲音源を生成する. 次に無聲音区間では, 白色雑音を生成する. そして, これらを足し合わせてフィルタバンクに通し, 各チャンネルのキャリアとなる. 本研究ではこの音源に対し, 位相の群遅延を操作するためのオールパスフィルタを畳み込み, キャリアの位相を操作する.

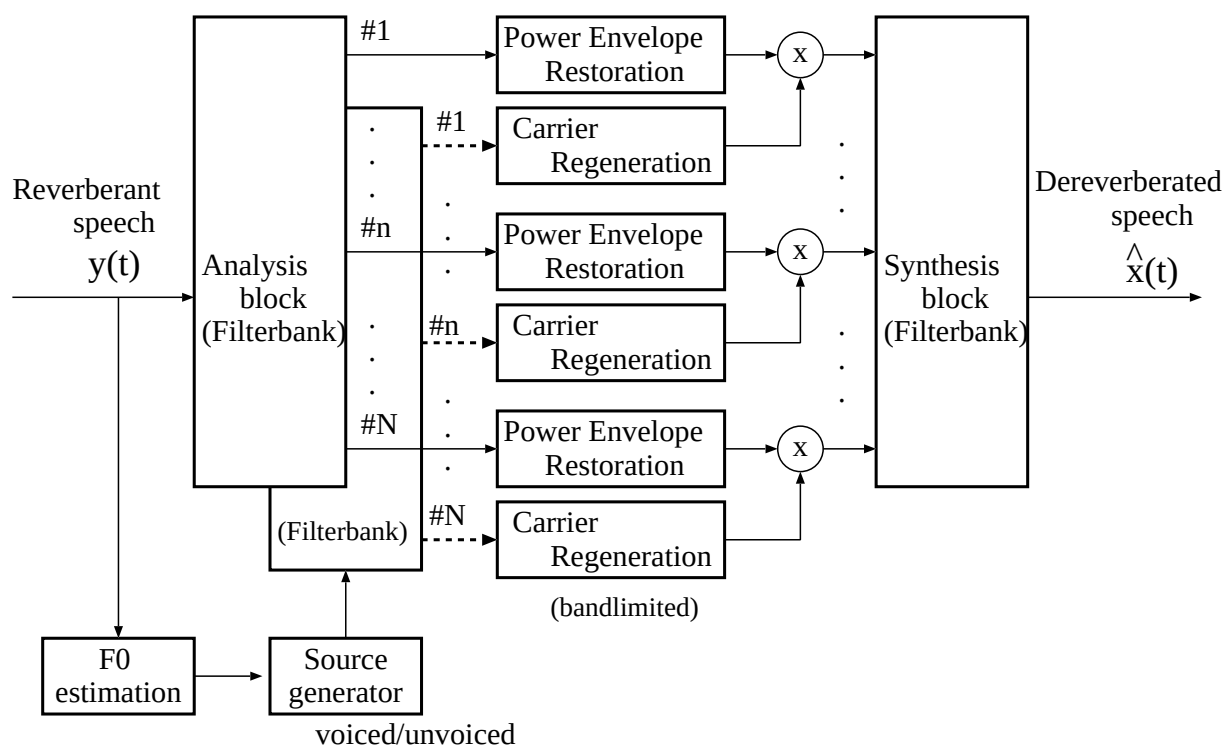


図 2.1: 残響音声回復モデル.

第3章 パワーエンベロープ回復法

3.1 パワーエンベロープ回復法の原理

3.1.1 変調伝達関数 (MTF)

Houtgast らが提案した変調伝達関数 (MTF) [1, 2] は, 室内伝達特性によって低下した音声明瞭度を測定する尺度として用いられている. その尺度は, 音声のパワーエンベロープ歪みを評価するものである. 変調伝達関数 $m(\omega)$ は以下の式で表される.

$$m(\omega) = \frac{\int_0^\infty h(t)^2 e^{-j\omega t} dt}{\int_0^\infty h(t)^2 dt}, \quad (3.1)$$

つまり, $m(\omega)$ は, 2 乗室内インパルス応答の複素フーリエ変換を全エネルギーで除算したものである. ここで, 室内インパルス応答 $h(t)$ を以下のように仮定する.

$$h(t) = e^{-t/\tau} n(t) = e^{-6.9t/T_R} n(t), \quad (3.2)$$

ただし, T_R は残響時間を示すパラメータであり, $h(t)$ のパワーが 60 dB 減衰するときの時間である. また, $n(t)$ は白色雑音を表す.

この式を MTF の式 3.1 に代入すると, 残響環境下における変調伝達関数を以下のよう求めることができる.

$$m(\omega) = \left[1 + \left(\omega \frac{T_R}{13.8} \right)^2 \right]^{-1/2}, \quad (3.3)$$

この式は, 残響の影響を受けることでパワーエンベロープの変調度が減少することを意味している. これより, パワーエンベロープの歪みが室内の残響音による影響である場合, MTF は室内のインパルス応答から求めることができる.

一例を図 3.1 に示す. 帯域雑音を変調周波数 $F_c = 10$ Hz の正弦波で 100 % (変調度 $m = 1$) 振幅変調した信号を考える. この時間波形とパワーエンベロープを図 3.1(a)(b) に示す. これに残響時間 $T_R = 0.5$ s の室内インパルス応答 (図 3.1(c)) を畳み込んだ残響信号を図 3.1(e) に示す. このとき, 観測された残響信号のパワーエンベロープは図 3.1(f) のようになり, 変調度が 1 から 0.29 に減少している. 式 3.3 に代入して得られた値 $m(F_c) = 0.31$ とほぼ等しい. このように, 室内インパルス応答から得られた MTF によって, 残響によってどれだけ影響を受けたかを知ることができる.

3.1.2 パワーエンベロープ逆フィルタ法

パワーエンベロープ逆フィルタ法 [8,9] において, MTF の概念 [11] に基づいて, 観測された残響信号を $y(t)$, 原信号を $x(t)$, そして室内インパルス応答を $h(t)$ とし, 以下のようにモデル化する.

$$\mathbf{y}(t) = \mathbf{x}(t) * \mathbf{h}(t), \quad (3.4)$$

$$\mathbf{x}(t) = e_x(t)\mathbf{n}_1(t), \quad (3.5)$$

$$\mathbf{h}(t) = e_h(t)\mathbf{n}_2(t) = a \exp(-6.9t/T_R)\mathbf{n}_2(t), \quad (3.6)$$

$$\langle \mathbf{n}_k(t), \mathbf{n}_k(t-\tau) \rangle = \delta(\tau) \quad (3.7)$$

ここで, $*$ は畳み込み演算, $e_x(t)$ と $e_h(t)$ は $\mathbf{x}(t)$ と $\mathbf{h}(t)$ のエンベロープ, $\mathbf{n}_1(t)$ と $\mathbf{n}_2(t)$ は白色雑音である. インパルス応答のパラメータ, a と T_R はそれぞれ振幅項と残響時間である [4]. 図 3.1 に, MTF に基づいたパワーエンベロープの関係を示す. モデル化された時間波形を左図 ((a) 原信号 $x(t)$, (c) 残響時間 $T_R = 0.5$ s のインパルス応答 $h(t)$, (e) 残響信号 $y(t)$) に示し, それぞれのパワーエンベロープを右図 ((b) $e_x(t)^2$, (d) $e_h(t)^2$, (f) $e_y(t)^2$) に示す.

ここで $\mathbf{y}(t)$ の集合平均を求める.

$$\begin{aligned} \langle \mathbf{y}(t)^2 \rangle &= \left\langle \left(\int_{-\infty}^{\infty} \mathbf{x}(t)\mathbf{h}(t-\tau) \right)^2 \right\rangle, \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e_x(\tau_1)e_x(\tau_2)e_h(t-\tau_1)e_h(t-\tau_2), \\ &\quad \langle \mathbf{n}_1(\tau_1)\mathbf{n}_1(\tau_2) \rangle \langle \mathbf{n}_2(t-\tau_1)\mathbf{n}_2(t-\tau_2) \rangle d\tau_1 d\tau_2 \\ &= \int_{-\infty}^{\infty} e_x(\tau)^2 e_h(t-\tau)^2 d\tau, \\ &= e_x(t)^2 * e_h(t)^2 \end{aligned} \quad (3.8)$$

また, $\langle \mathbf{y}(t)^2 \rangle$ は $\langle \mathbf{y}(t)^2 \rangle = \langle e_y(t)^2 n(t)^2 \rangle = e_y(t)^2$ となることから, 以下の式が得られる.

$$\langle \mathbf{y}(t)^2 \rangle = e_x(t)^2 * e_h(t)^2 = e_y(t)^2, \quad (3.9)$$

ここで $\langle \cdot \rangle$ は集合平均である. この式は原信号のエンベロープと室内インパルス応答のエンベロープの関係を示している. この結果に基づけば, $e_y(t)^2$ に $e_h(t)^2$ の逆特性を畳み込むことで, $e_x(t)^2$ を回復することができる. ここで, パワーエンベロープ $e_x(t)^2$, $e_h(t)^2$, $e_y(t)^2$ の z 変換をそれぞれ $E_x(z)$, $E_h(z)$, $E_y(z)$ とおく. これより, 室内インパルス応答の伝達関数 $E_h(z)$ は以下のように計算される.

$$E_h(z) = \frac{a^2}{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}}, \quad (3.10)$$

ここで、 f_s はサンプリング周波数 (20 kHz) である。したがって、原信号のパワーエンベロープの伝達関数 $E_x(z)$ は、次のように決定できる。

$$E_x(z) = \frac{E_y(z)}{a^2} \left\{ 1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1} \right\}, \quad (3.11)$$

この $E_x(z)$ の逆 z 変換を求めることにより、パワーエンベロープ $e_x(t)^2$ を求めることができる。

以上より、観測した残響信号 $y(t)$ のパワーエンベロープを抽出し、各パラメータ a, T_R を決定できれば、パワーエンベロープを正確に回復することができる。

例として、回復したパワーエンベロープ $\hat{e}_x(t)^2$ を図 3.1(g) に示す。適切な残響時間 (ここでは、 $T_R = 0.5$ s) を利用して逆フィルタ処理を行えば、ほぼ完全にパワーエンベロープ (実線) を回復できることがわかる。また、異なる T_R を利用して逆フィルタ処理を行えば、不十分な T_R であれば、不十分な回復を、過剰な T_R であれば、過剰な回復をすることがわかる。

3.1.3 パラメータの決定法

本研究では、以下のアルゴリズムを用いて、各パラメータを決定している。

(a) パワーエンベロープ $e_y(t)^2$ の抽出:

パワーエンベロープ逆フィルタ処理を行うには、観測信号 $y(t)$ からパワーエンベロープ $e_y(t)^2$ を抽出する必要がある。本研究では、Hilbert 変換を用いたパワーエンベロープ抽出法を用いる [17]。

$$\hat{e}_y(t)^2 = \text{LPF} \left[|y(t) + j \cdot \text{Hilbert}(y(t))|^2 \right]. \quad (3.12)$$

ここで、 $\text{Hilbert}()$ はヒルベルト変換である。本研究では、低域通過フィルタ $\text{LPF}[\cdot]$ のカットオフ周波数は 20 Hz とした。

(b) 残響時間 T_R の推定:

図 3.1(g) からわかるように逆フィルタ処理は、残響時間パラメータ T_R の値が大きくなるほど、パワーエンベロープの山と谷を強調させる働きがあり、ある程度大きくすると、パワーエンベロープが負の値を持つようになる。音源信号には必ず無音区間が存在する、すなわち $e_x(t)^2$ の変調度が 1 である仮定の基で、逆フィルタ処理のこの特徴を利用して、以下のようにパワーエンベロープの負の値を持つ直前の残響時間パラメータを調べること、回復処理に適した残響時間パラメータを決定できる。

$$T_P(T_R) = \min \left(\arg \min_{t_{\min} \leq t \leq t_{\max}} \left| \hat{e}_{x,T_R}(t)^2 - \theta \right| \right), \quad (3.13)$$

$$\hat{T}_R = \arg \min_{0 \leq T_R \leq T_{R,\max}} \left\{ \frac{dT_P(T_R)}{dT_R} \right\}. \quad (3.14)$$

ここで、 θ は $e_y(t)^2$ の最大値から決定した閾値である．本研究では、 θ を $e_y(t)^2$ の最大値から -20 dB の位置とした．また、 $t_{\min}$ と $t_{\max}$ はそれぞれ、 t の下限と上限とする． T は $y(t)$ の区間、 $\hat{e}_{x,T_R}(t)^2$ は T_R の関数として残響回復されたパワーエンベロープの候補の集合、 $T_{R,\max}$ は T_R の範囲の上限である．

(c) 振幅項 a の決定:

残響の特性が信号を増加させるのではなく、主に信号成分の位相遅れに影響を与えるものと考え、室内インパルス応答のパワーエンベロープの面積が 1 となるように a を決定した．

$$a = \sqrt{1 / \int_0^T \exp(-13.8t/T_R) dt}. \quad (3.15)$$

図 3.2 に、これらのパラメータの決定をふまえたパワーエンベロープ逆フィルタ処理の流れを示す．

3.1.4 パワーエンベロープ逆フィルタ法の問題点

従来法 [18] では、 N チャネル ($N = 100$, 帯域幅 100 Hz) の定帯域フィルタバンクで帯域分割を行った残響音声に対して、パワーエンベロープ逆フィルタ処理を行っている．この帯域幅は、音声のパワーエンベロープの共変調の性質と、MTF 理論の成立を考慮して決定された．帯域幅を狭くするにつれ、帯域幅内でのパワーエンベロープの共変調とみなすことができる．しかしその一方で、キャリア間の無相関性である仮定が崩れてしまい、式 3.9 に示す MTF のパワーエンベロープの関係が成り立たなくなってしまう．従って、この共変調の性質と MTF の関係はトレードオフの関係にあり、両者を満たす最適な帯域幅を決定する必要があった．そのため従来法では、全帯域で平均 100 Hz が最も良い帯域幅であるとした．

しかし、音声のパワーエンベロープの共変調の性質は、各音声の帯域毎に異なる．そのため、共変調の性質と MTF 成立のトレードオフの関係も各帯域毎に異なる．従って、全帯域で一定の帯域分割を行うよりも、各帯域毎に適切に帯域幅を変えてやることによって、さらに回復精度が向上する可能性が考えられる．

また、従来法では、全時間区間一括で逆フィルタ処理を行っている．しかし、無音区間やエンベロープの谷が多く存在した場合や、山谷の起伏が多い場合などは、回復精度が低

下してしまうという問題点がある。

3.2 時間–周波数分割処理

3.1.4 節で挙げた問題点を解決するために、本研究では、対象となる音声信号に依存して適応的に時間–周波数分割を考慮したパワーエンベロープ回復法を提案する。時間領域では、無音区間の前後でそれぞれ残響時間パラメータを決定し、各区分毎でそれぞれ逆フィルタ処理を適応する。また、周波数領域では、各帯域毎にパワーエンベロープの相関を計算し、可変的に帯域分割を行う。各音声毎に、適切な帯域幅や時間区間を定めることでパワーエンベロープの回復精度をさらに高められる可能性がある。

時間–周波数分割処理による、パワーエンベロープの概念を図 3.3 に示す。図に示すように周波数領域では、帯域毎に帯域幅が異なる。また、時間領域では、それぞれ時間区間ごとに分割され、各区分ごとに逆フィルタ処理を行う。この分割幅は各音声ごとに異なる。

3.2.1 時間分割処理

時間分割処理では、閾値処理を利用して各チャンネルごとに時間分割を行う。本稿では、閾値処理を利用して無音区間の中心点を検出し、その中心点の前後で区間を分割して回復処理を行う。例として図 3.4 に、ATR データベース中の単語 (mau, /aikawarazu/) の 76 チャンネル目のパワーエンベロープを示す。太線は原音声 $e_x(t)^2$ を示す。細線と破線はそれぞれ回復音声のパワーエンベロープ $\hat{e}_x(t)^2$ と残響音声のパワーエンベロープ $e_y(t)^2$ を示す。図 3.4(a) は全区間一括で回復処理を行った結果を示し、図 3.4(b) では音声区間を $\theta = -10$ dB とした閾値処理によって 3 つのセグメントに分割して、それぞれの区間に対し回復処理を行った結果を示す。推定残響時間 $\hat{T}_R(t)$ は、3 つのセグメントに分けた場合は 0.56, 1.50, 0.41 s であるのに対し、一括処理の場合は 0.29 s である。図 3.4(a)(b) の 2 つの図で $e_x(t)^2$ と $\hat{e}_x(t)^2$ の近似具合を比較すると、時間分割処理を行ったほうが精度良く $\hat{e}_x(t)^2$ を回復できている。

3.2.2 最適な閾値の決定

ここでは、時間分割処理において、パワーエンベロープ $e_y(t)^2$ の最大振幅から何 dB 下がった位置を無音区間検出の閾値 θ にするかを検討する。

シミュレーション条件は、F0 が事前に正しく推定されているものとし、サンプリング周波数 f_s は 20 kHz とした。音声信号として、ATR 音声データベース中の 10 人の話者（男性 5 名、女性 5 名）が発話した 3 文章（“相変わらず”，“新聞”，“冗談”）を利用し、これに 500 種類の室内インパルス応答（式 (3), 100 種類の $n(t)$, $T_R = 0.1, 0.3, 0.5, 1.0, 2.0$ s）を畳み込んだ残響音声 $y(t)$ に対して回復処理を行った。帯域幅 100 Hz の定帯域フィル

タバンクを用い、最大振幅からの位置を -8 , -10 , -15 , -20 dB と変化させ、時間分割処理を行った。評価は、従来法（一括処理）からの相関と SNR の各チャネルの平均改善量で比較した。相関と SNR は以下のように計算する。

$$\text{Corr}(e_x^2, \hat{e}_x^2) = \frac{\int_0^T (e_x(t)^2 - \overline{e_x(t)^2}) (\hat{e}_x(t)^2 - \overline{\hat{e}_x(t)^2}) dt}{\sqrt{\left\{ \int_0^T (e_x(t)^2 - \overline{e_x(t)^2})^2 dt \right\} \left\{ \int_0^T (\hat{e}_x(t)^2 - \overline{\hat{e}_x(t)^2})^2 dt \right\}}}, \quad (3.16)$$

$$\text{SNR}(e_x^2, \hat{e}_x^2) = 20 \log_{10} \frac{\int_0^T e_x(t)^2 dt}{\int_0^T (e_x(t)^2 - \hat{e}_x(t)^2) dt}, \quad (\text{dB}) \quad (3.17)$$

ここで、 $\overline{e_x(t)^2}$ は $e_x(t)^2$ の平均である。相関の改善度を $\text{Corr}(e_x^2, \hat{e}_x^2) - \text{Corr}(e_x^2, e_y^2)$ 、SNR の改善度を $\text{SNR}(e_x^2, \hat{e}_x^2) - \text{SNR}(e_x^2, e_y^2)$ で計算する。

結果を図 3.5 に示す。4 つの閾値を比較すると、 $\theta = -10$ dB 時に SNR、相関ともに最も高い改善度が得られた。

次に、例として、この閾値 $\theta = -10$ dB としたときの各チャネル毎の SNR の改善度を図 3.6 に示す。残響時間が 1.0 s の場合の (a) 従来法の一括処理、(b) 時間分割処理の改善度を示す。2 つの図を比較すると、20, 40, 58 チャネル付近において、時間分割処理による改善効果が見られる。他の残響時間 ($T_R = 0.1, 0.3, 0.5, 2.0$ s) においても時間分割処理による改善効果が見られた。

3.2.3 周波数分割処理

周波数分割処理では、対象となる音声の共変調に基づいてフィルタバンクの適切な帯域幅の決定を行う。本稿では、狭帯域 (40 Hz) フィルタバンクで分割されたパワーエンベロップ間の相関を求め、その相関に基づいて適応的に帯域幅を決定する。例として図 3.7 では、ATR データベース中の単語 (mau/aikawarazu/) に対して、相関値 0.98 と見なせるエンベロップの幅を等高線で示し、右図にはその帯域分割幅を示す。横軸はチャネルの中心周波数を示し、縦軸は隣接したチャネルの周波数幅を示す。つまり、ある中心周波数に対する等高線の幅が、その帯域の分割幅を示している。この図 3.7 の等高線の場合、平均 90 Hz が全周波数帯域で最適な帯域分割幅ということになる。図 3.7(右) に適応的に計算した帯域分割幅を示す。これらの帯域幅は対象となる音声や条件によって変化する。

もし、対象となる音声ごとに、この等高線に基づいたフィルタバンクの帯域幅を適応的に決定できれば、パワーエンベロップ逆フィルタ処理は各チャネルごとにさらに正確なパワーエンベロップを回復できると考えられる。

また、相関値を変えた場合の等高線を図 3.8 に示す。(左) が相関値 0.9 の場合、(右) が相関値 0.85 の場合の等高線である。この図から、相関値を小さくすることにより、帯域幅は広く計算されることがわかる。3.1.4 節でも述べたように、MTF の関係と共変調の性質はトレードオフの関係にあるため、両者を満たす最適な相関値を決定する必要がある。

3.2.4 最適な相関値の決定

ここでは、周波数分割処理における最適な相関値の決定を行う。3.2.2 節と同じシミュレーション条件で回復処理を行った。帯域幅を決定する相関値を 0.70, 0.80, 0.95, 0.98, 0.99 と変化させて、周波数分割を行った結果を図 3.9 に示す。評価は、周波数 1 Hz あたりの SNR と相関の平均改善度で行った。結果を比較すると、相関値が 0.98 のとき、SNR, 相関ともに最も改善度が高いことがわかった。これは相関値が高いほど、音声の特徴を良くとらえることができるからと考えられる。また、相関値がこれより高くなると、帯域幅は一定の狭帯域分割に近づいてしまうため、改善度は低下する。

3.3 時間–周波数分割処理の評価

改良法が最適な帯域分割法であるかを調査するため、他の帯域分割法との比較を行う。3.2.2 節と同じ条件で回復処理を行った。比較として、定帯域分割（従来法、帯域幅 100 Hz）、定 Q 分割（チャンネル数 32 ch, 中心周波数 600 Hz）、時間分割（閾値 −10 dB）、周波数分割（相関値 0.98）、そして適応的な時間–周波数分割（提案法、閾値 −10 dB, 相関値 0.98）の 5 つの回復法を評価した。

まず、図 3.10 に各音声の絶対的な SNR と相関の値を示す。残響音声に比べて、どの回復法を用いた場合も非常に大きな改善が得られている。次に、図 3.11 に、SNR と相関の平均改善度を示す。周波数分割、定帯域分割、定 Q 分割を比較すると、周波数分割の改善度が最も高いことがわかる。つまり、周波数領域における帯域分割法では、本研究で提案した適応的な周波数分割が最も適切な分割法であることがわかった。これは、対象となる音声とは無関係に一定で帯域を分割するより、音声に依存して適応的に分割を行う方が、改善度が向上するというを示している。また、同じ図 3.11 に、時間分割と周波数分割を組み合わせた提案法の結果も示す。2 つの手法を組み合わせることにより、従来法に比べ、 $T_R = 0.5$ s 時の SNR で約 1.2 dB の改善が得られた。特に、周波数分割処理が結果の向上に大きく貢献していることがわかる。これは、MTF が成立すると考えられる帯域内で、より音声の特徴をとらえているためと考えられる。

次に、以下のように計算される対数スペクトル歪み（LSD）による残響音声回復の評価を行った。

$$\text{LSD} = \sqrt{\frac{1}{W} \sum_{\omega} \left(20 \log_{10} \frac{|S(\omega)|}{|\hat{S}(\omega)|} \right)^2}, \quad (\text{dB}) \quad (3.18)$$

ここで、 W は周波数の上限（ここでは 10 kHz とした）、 $|S(\omega)|$ と $|\hat{S}(\omega)|$ はそれぞれ $\tilde{x}(t)$ と $\tilde{y}(t)$ もしくは $\hat{x}(t)$ の振幅スペクトルである。ここで、 $\tilde{x}(t)$ は、原音声にキャリア再生成処理を行った音声であり、 $\tilde{y}(t)$ は $\tilde{x}(t)$ に残響を付加した音声 ($\tilde{x}(t) * h(t)$) である。

これらの評価において、4 つの手法（定帯域分割処理（従来法）、時間分割処理、周波数分割処理、時間–周波数分割処理（提案法））の比較を行い、時間分割処理と周波数分

割処理を用いた場合の有効性を示す．図 3.12 に，4 つの手法を用いたときの LSD の改善度を示す．全てのモデルにおいて，対数スペクトル歪みが減少しているのがわかる．そして，従来法を用いた場合に比べ，時間分割処理や周波数分割処理を用いた場合の方がさらに効果的に改善されているのがわかる．提案法（時間+周波数分割処理）では，全ての条件において最も改善精度が良く，従来法に比べ LSD の改善度は約 1 dB 向上した．

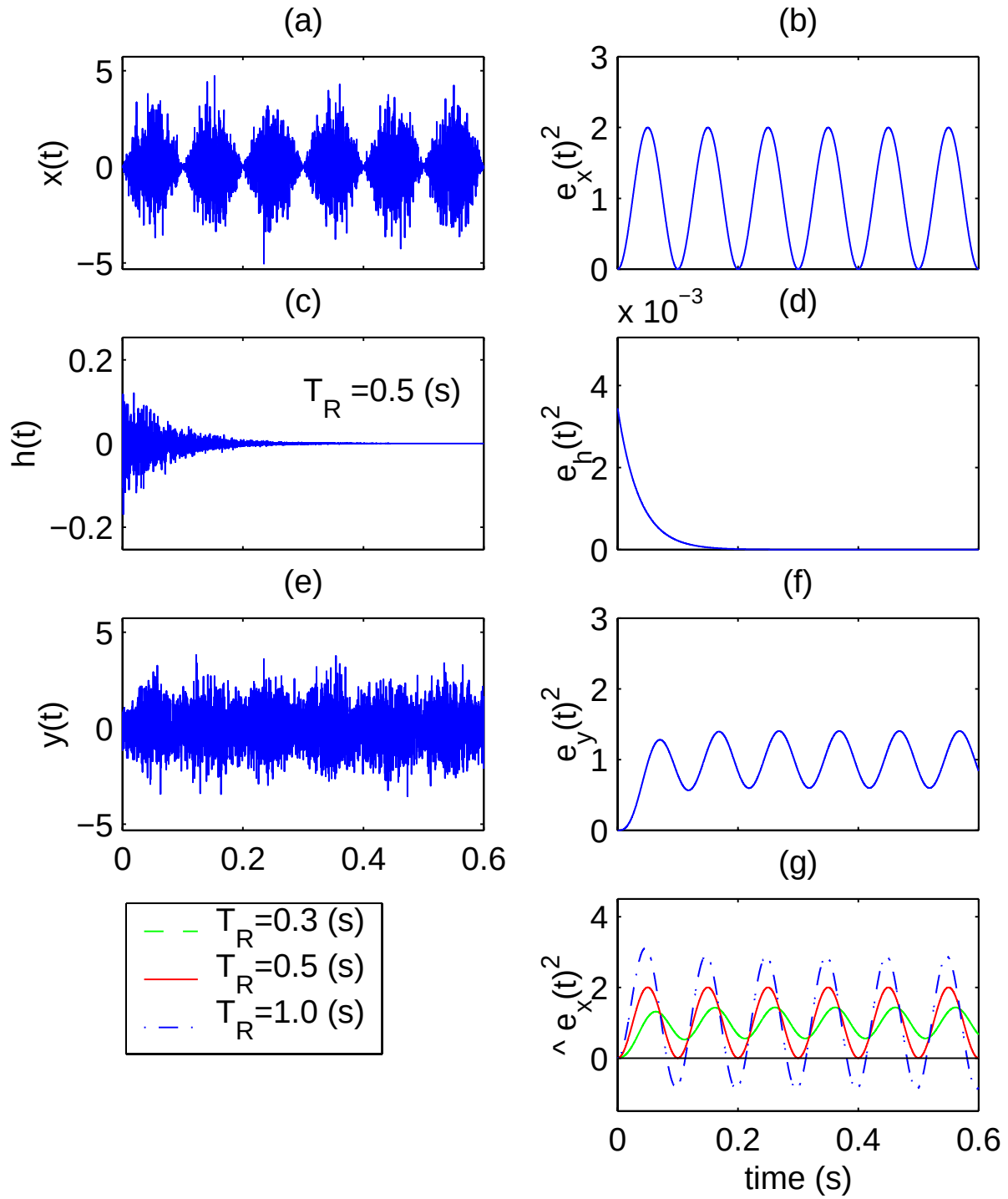


図 3.1: MTF に基づいたパワーエンベロープの関係：(a) 原信号 $x(t)$ と (b) パワーエンベロープ $e_x(t)^2$, (c) 室内インパルス応答 $h(t)$ と (d) パワーエンベロープ $e_h(t)$, (e) 残響信号 $y(t)$ と (f) パワーエンベロープ $e_y(t)$, (g) 残響回復されたパワーエンベロープ $\hat{e}_x(t)^2$

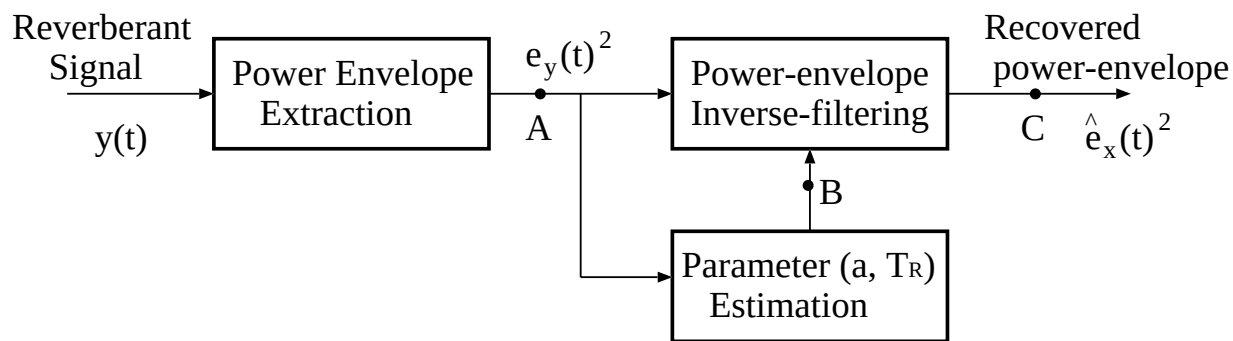


図 3.2: パワーエンベロープ逆フィルタ処理の流れ.

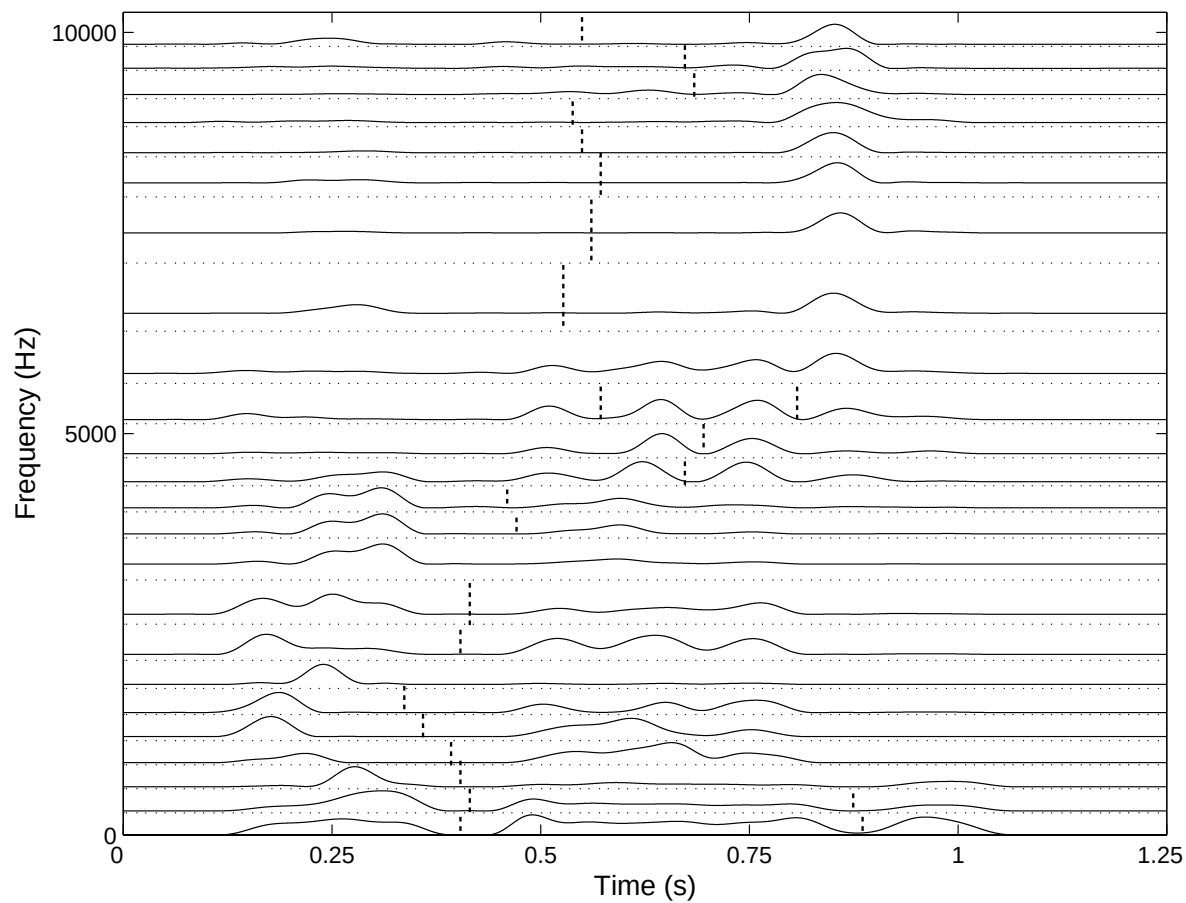


図 3.3: 時間-周波数分割処理によるパワーエンベロープ.

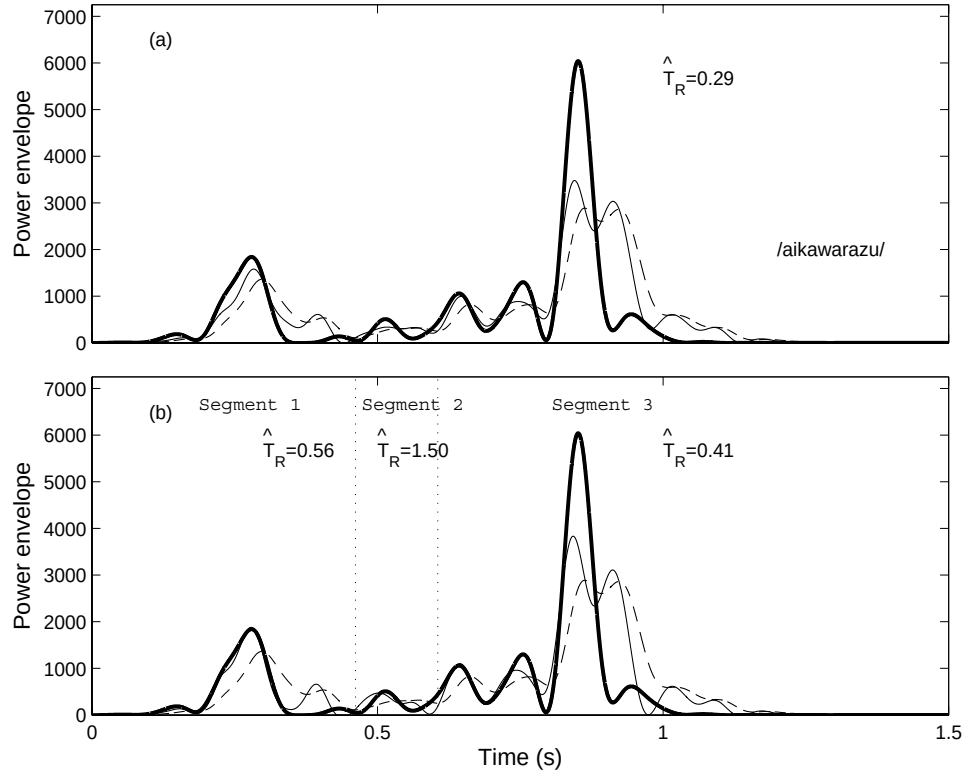


図 3.4: 逆フィルタの (a) 一括処理と (b) 時間分割処理, 太線: 原音声パワーエンベロープ $e_x(t)^2$, 細線: 回復パワーエンベロープ $\hat{e}_x(t)^2$, 破線: 残響パワーエンベロープ $e_y(t)^2$.

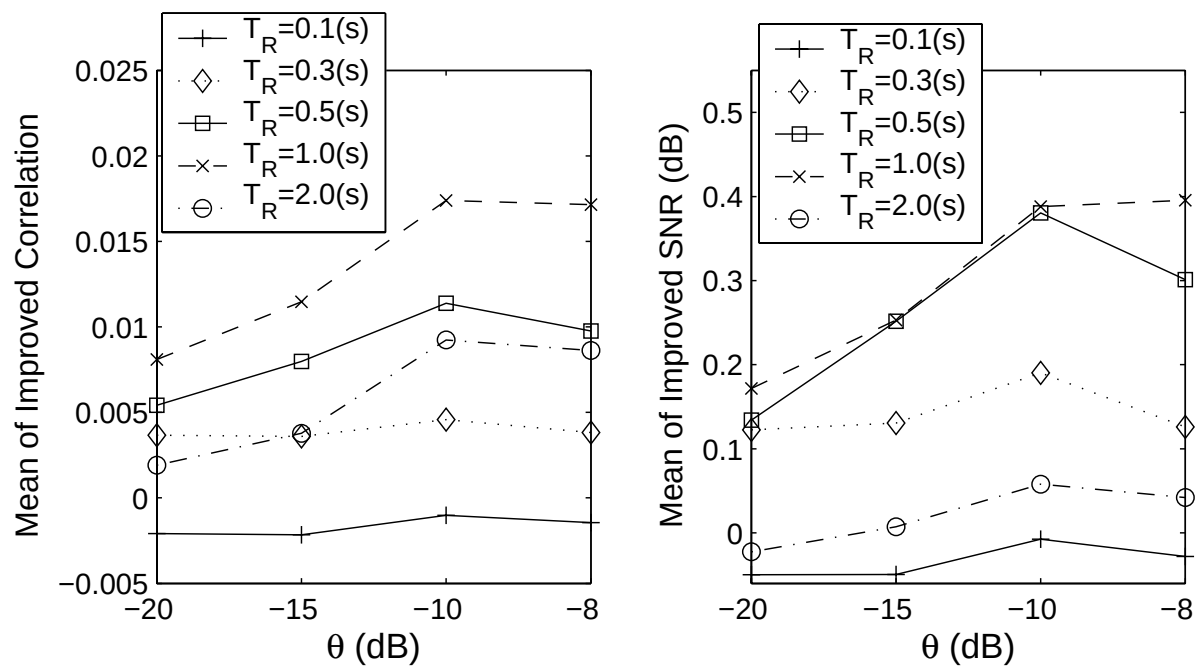


図 3.5: 閾値 θ の決定：従来法からの (a) 相関と (b) SNR の平均改善量.

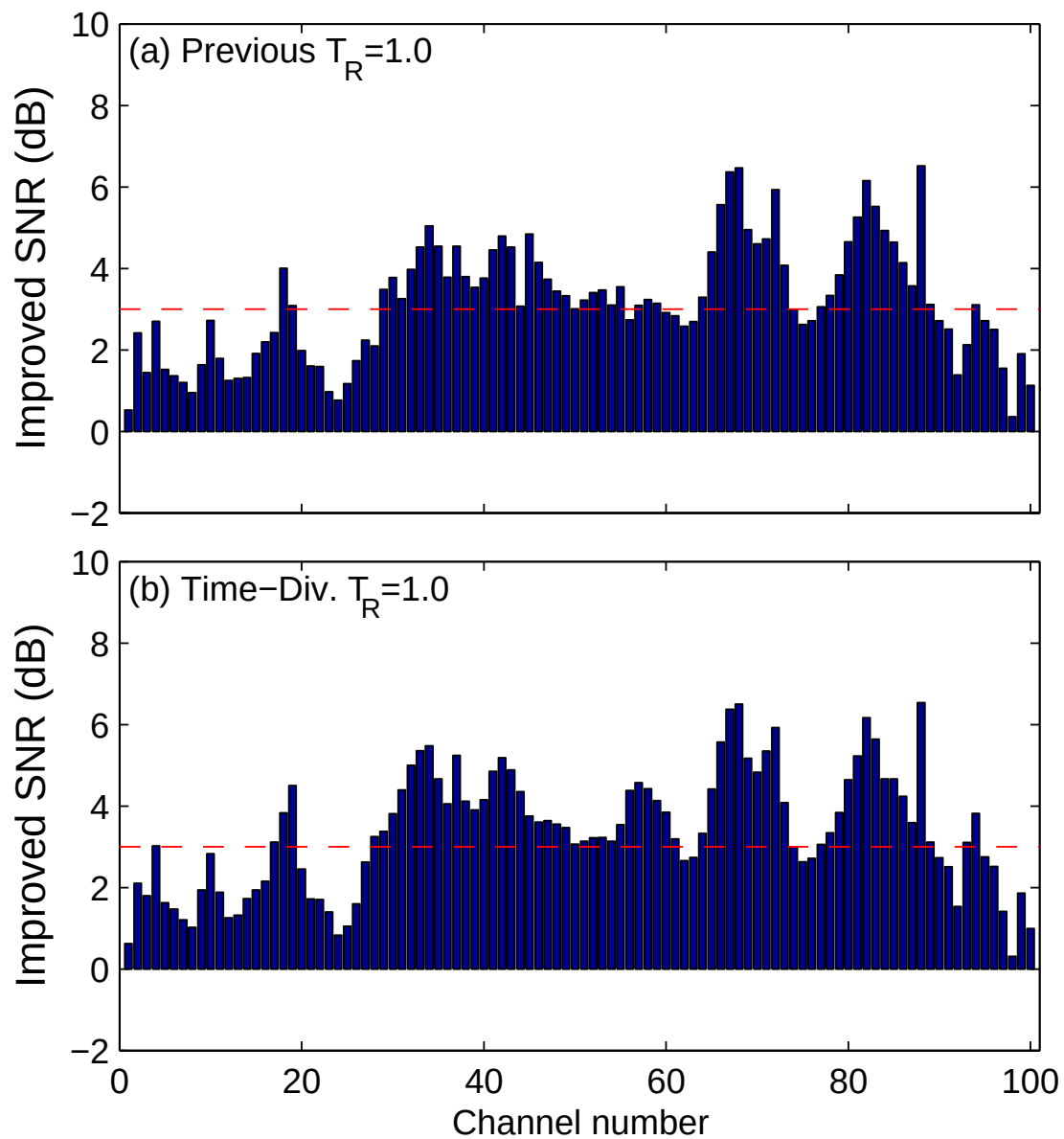


図 3.6: フィルタバンクにおける音声信号の SNR 回復精度: (a) 一括処理 (b) 時間分割処理.

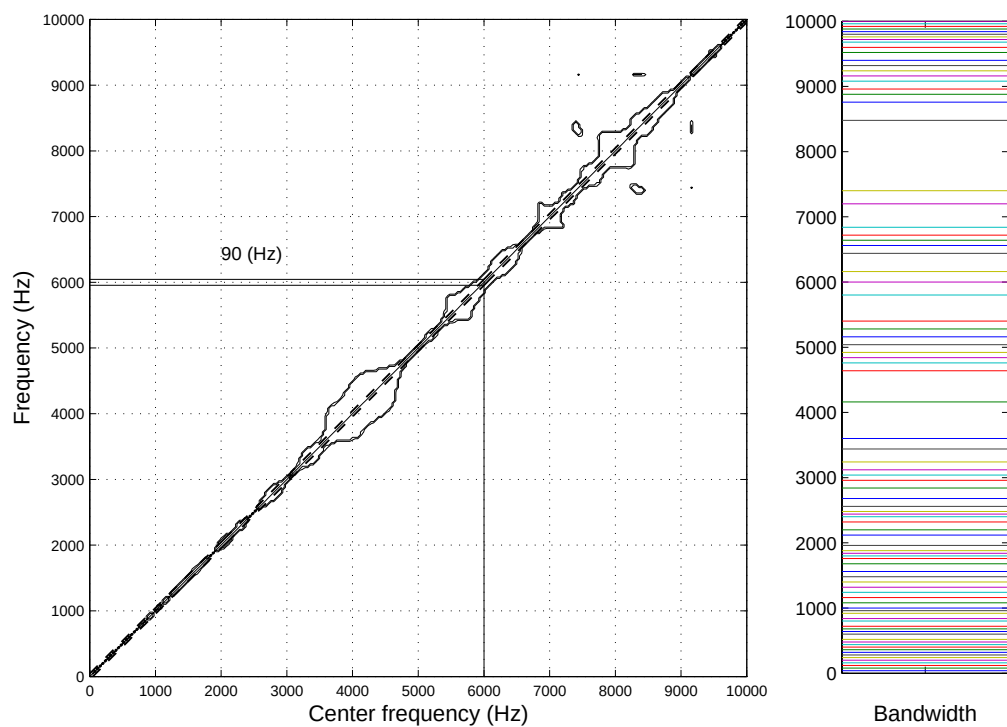


図 3.7: 帯域分割幅の決定：(a) 相関図と (b) 帯域分割幅.

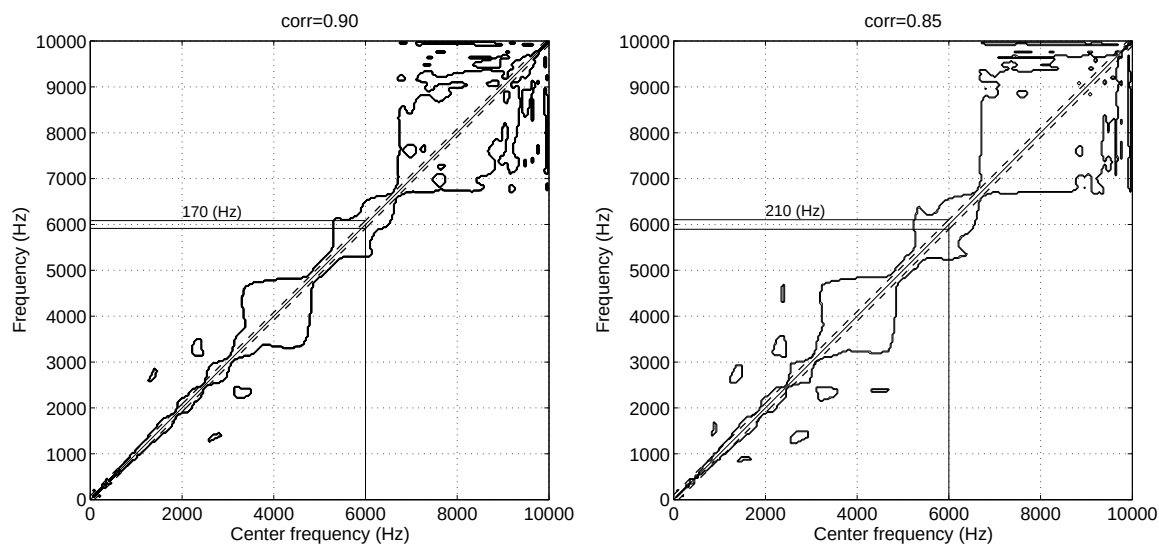


図 3.8: 相関図：相関値が (a)0.9, (b)0.85 の場合.

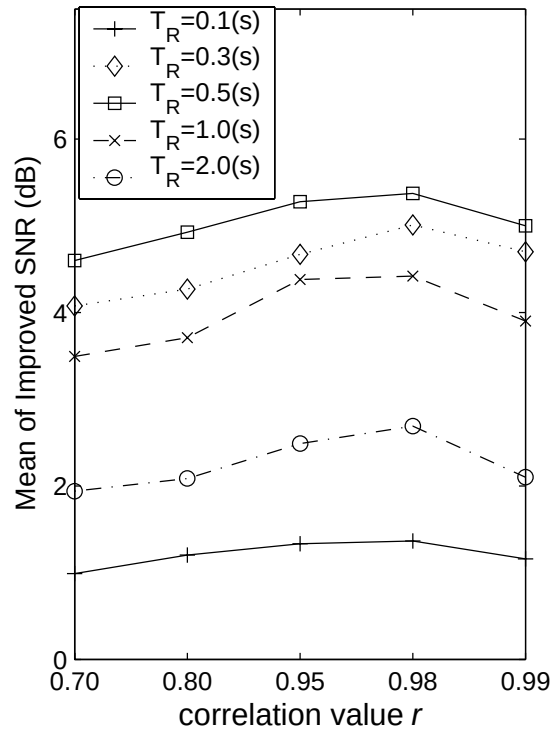
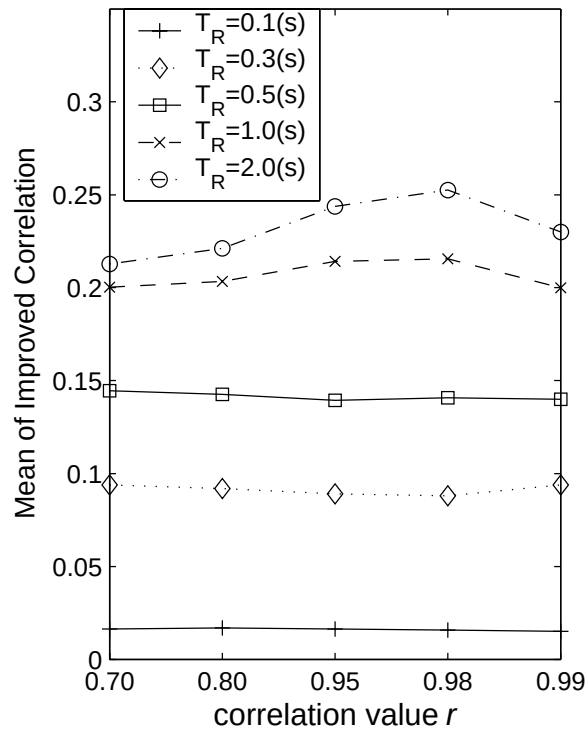


図 3.9: 相関値の決定：(a) 相関と (b) SNR の平均改善度.

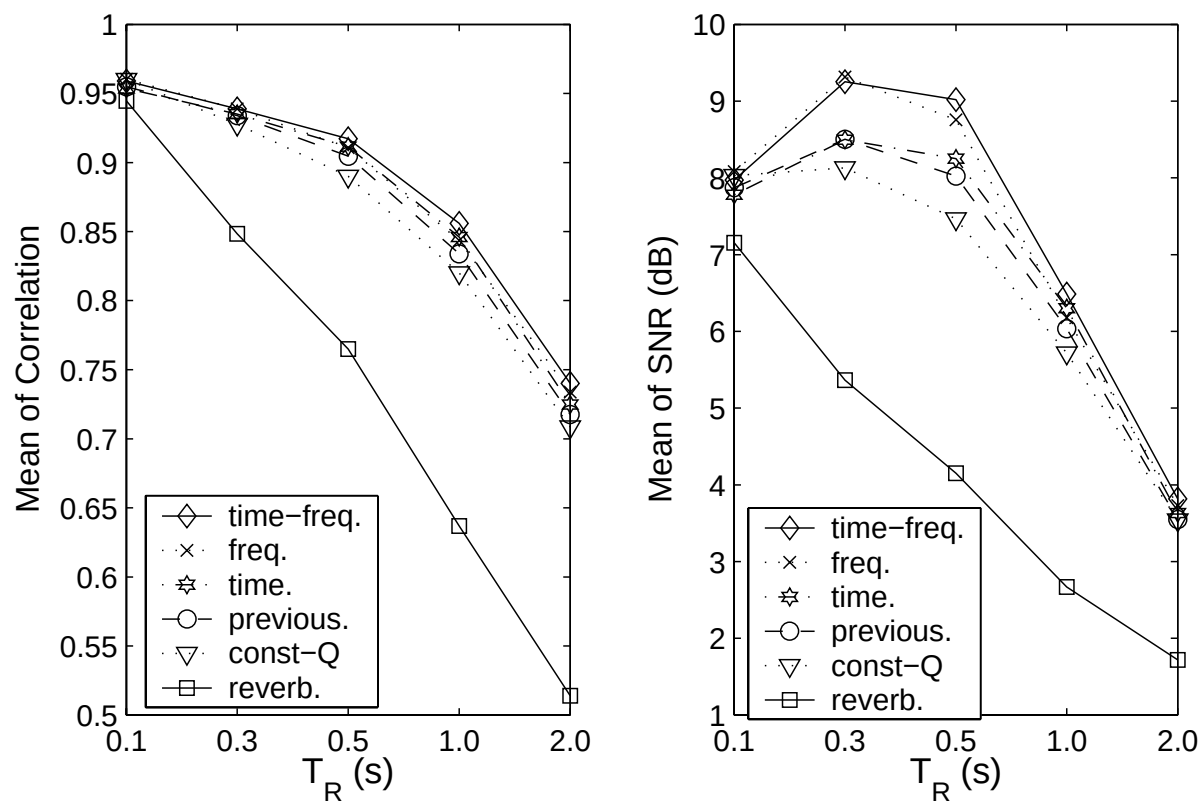


図 3.10: 帯域分割法の比較：(a) 相関と (b) SNR.

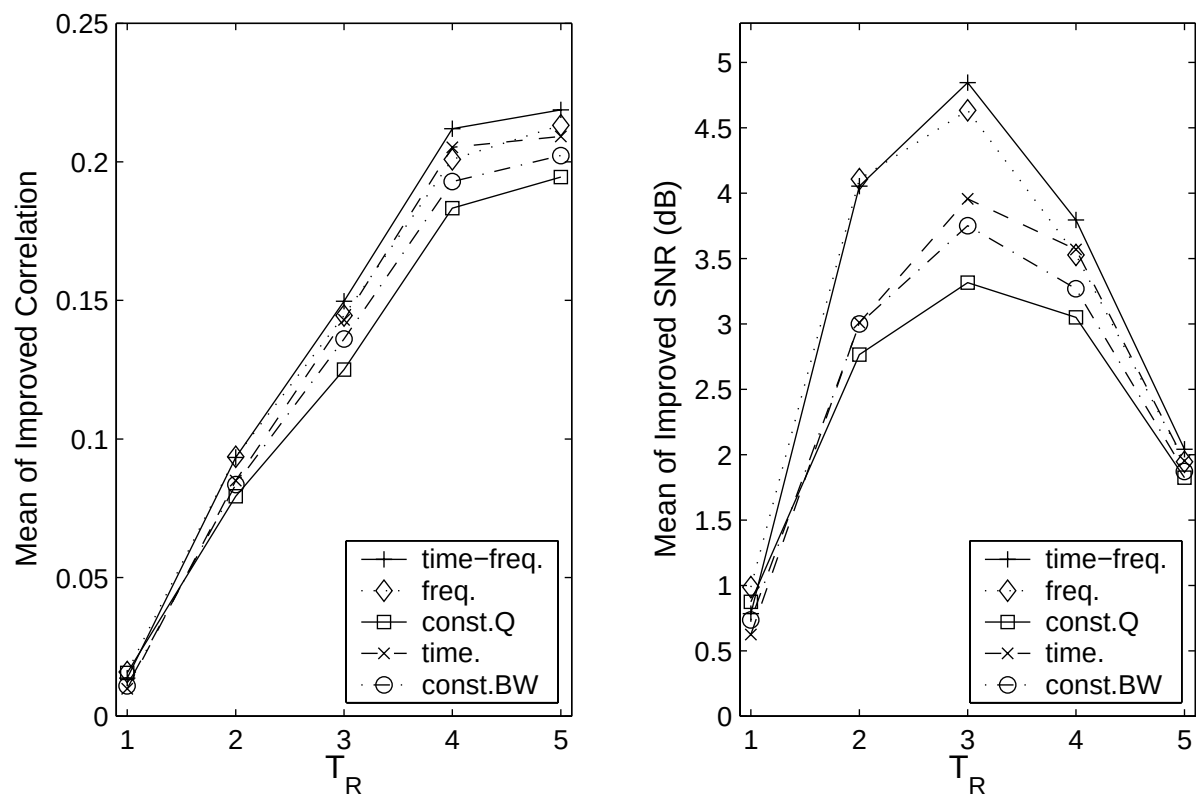


図 3.11: 帯域分割法の比較：(a) 相関と (b) SNR の平均改善度.

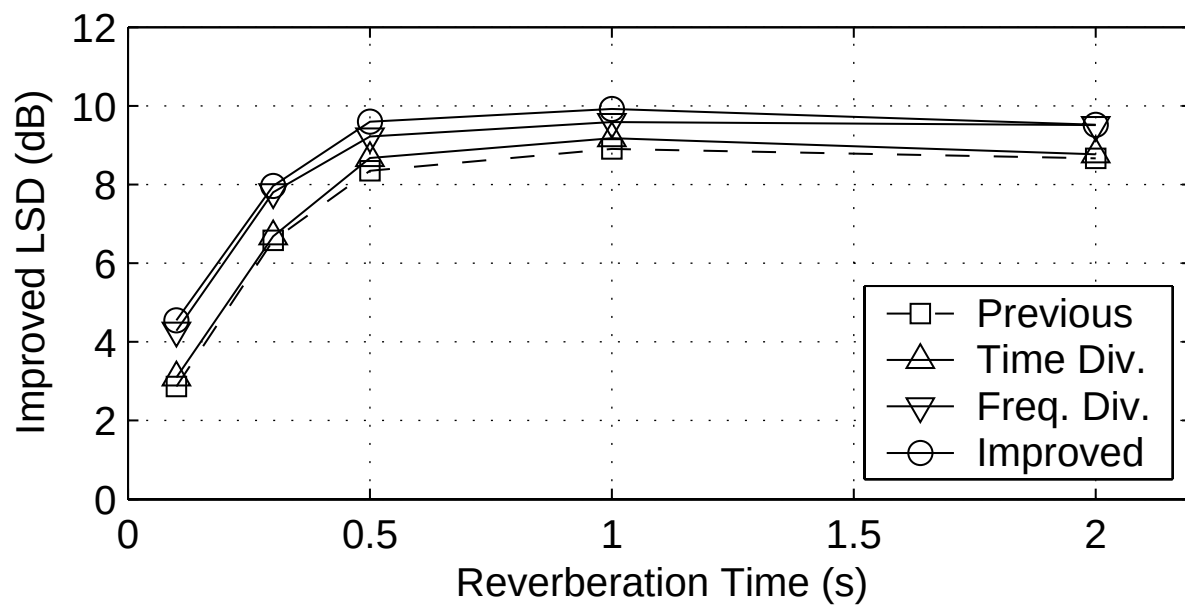


図 3.12: LSD 改善度.

第4章 カットオフ周波数の検討

4.1 最適なエンベロープ抽出を行うカットオフ周波数

これまで、式 3.12 に示すパワーエンベロープ抽出のためのカットオフ周波数を $f_c = 20$ Hz として処理を行っていた。しかし、この値はアドホックに決定した値であり、最適なエンベロープ抽出を行うカットオフ周波数であるという考察はない。ここでは、パワーエンベロープ抽出におけるカットオフ周波数についての検討を行う。

まず、式 3.9 に示す MTF に基づいたパワーエンベロープの関係を考える。このとき、残響音声から抽出したパワーエンベロープ $e_y(t)^2$ と、 $e_x(t)^2 * e_h(t)^2$ から求められたパワーエンベロープ $\tilde{e}_y(t)^2$ が最も近づく場合の LPF のカットオフ周波数を求める。例として、mau, /aikawarazu/ の音声に、残響時間 $T_R = 2.0$ s のインパルス応答を畳み込んだ残響音声に対して、最も $e_y(t)^2$ と $\tilde{e}_y(t)^2$ が近づくときのカットオフ周波数を図 4.1 に示す。この結果を見ると、従来用いていた $f_c = 20$ Hz よりもカットオフ周波数は小さな値になっていることがわかる。また、各チャンネルごとにカットオフ周波数は変動する。

このカットオフ周波数を用いて、mau, /aikawarazu/ の音声に、残響時間 $T_R = 2.0$ s のインパルス応答を畳み込んだ残響音声に対して回復処理を行った結果の例を図 4.2 に示す。評価は相関と SNR の改善度で示す。図の上が従来のカットオフ周波数 $f_c = 20$ Hz を用いてパワーエンベロープを抽出する場合の相関と SNR の改善度、図の下が提案した式 4.4 で求めた可変カットオフ周波数でパワーエンベロープを抽出する場合の相関と SNR の改善度である。この図より、カットオフ周波数を各チャンネルごとに可変的に求めることにより、相関も SNR も従来の $f_c = 20$ Hz を用いてパワーエンベロープを抽出する場合に比べて、回復精度は大きく向上する。これは、MTF の関係が最も成立するようにカットオフ周波数を求めているからである。

4.2 カットオフ周波数の導出過程

上記で利用したカットオフ周波数は、原音声のパワーエンベロープ情報 $e_x(t)^2$ を用いて計算しているため、実際のシステムでは利用できない。そこで、このカットオフ周波数を $e_x(t)^2$ を用いずに理論的に導出することを考える。

まず、室内インパルス応答 $h(t)$ のパワーエンベロープ $e_h(t)^2$ について、MTF の式 3.9 を考える。

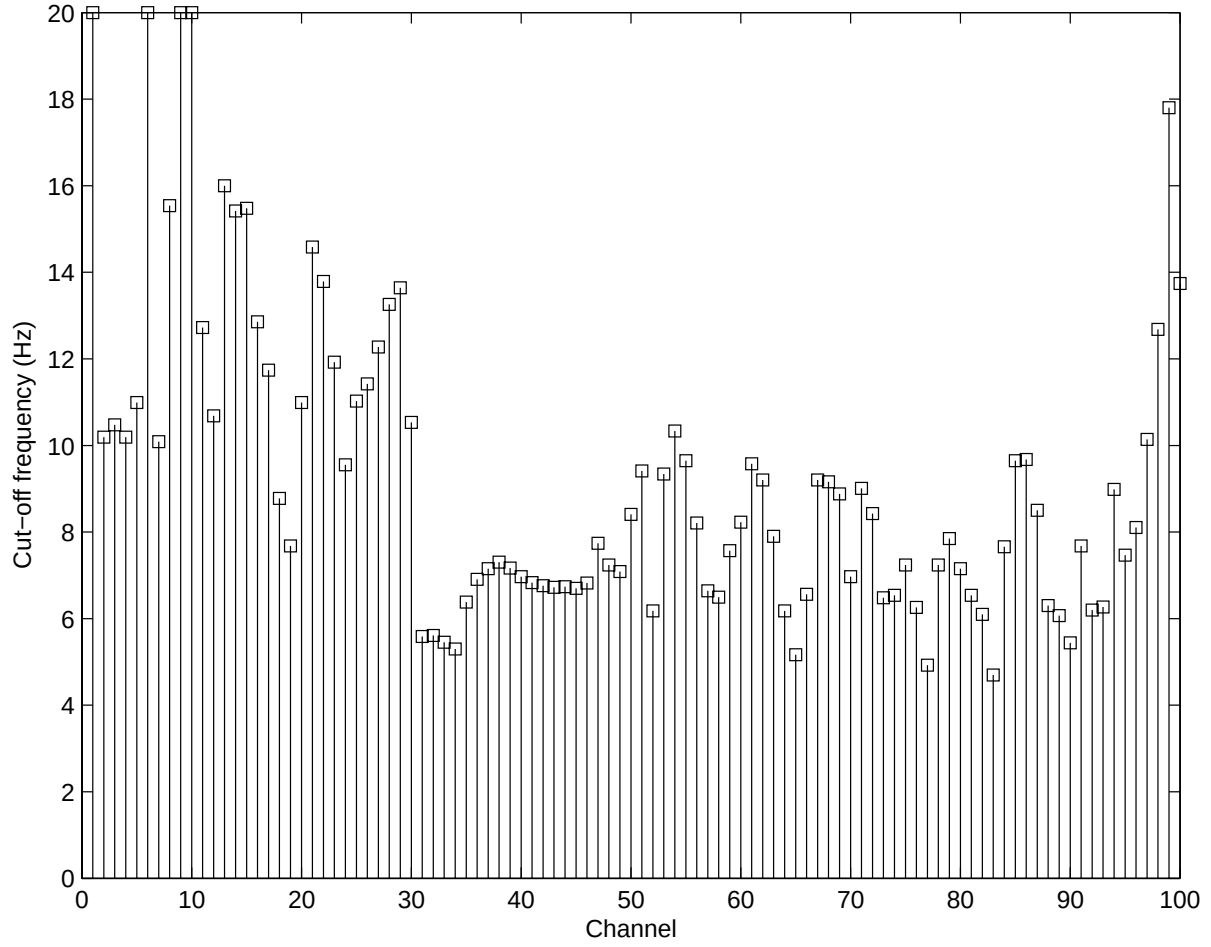


図 4.1: カットオフ周波数の例 (mau, /aikawarazu/, $T_R = 2.0$ s) .

$$H(\omega) = \frac{\int_0^\infty e_h(t)^2 e^{-j\omega t} dt}{\int_0^\infty e_h(t)^2 dt}, \quad (4.1)$$

室内インパルス応答のパワーエンベロープは式 3.6 より, $e_h(t)^2 = a^2 \exp(-13.8t/T_R)$ であるので,

$$\begin{aligned} H(\omega) &= \frac{\int_0^\infty a^2 \exp(-13.8t/T_R) e^{-j\omega t} dt}{\int_0^\infty a^2 \exp(-13.8t/T_R) dt}, \\ &= \frac{a^2 / (13.8/T_R + j\omega)}{a^2 T_R / 13.8}, \\ &= \frac{1}{1 + j\omega T_R / 13.8}, \\ |H(\omega)| &= \frac{1}{\sqrt{1 + (\omega T_R / 13.8)^2}}, \end{aligned} \quad (4.2)$$

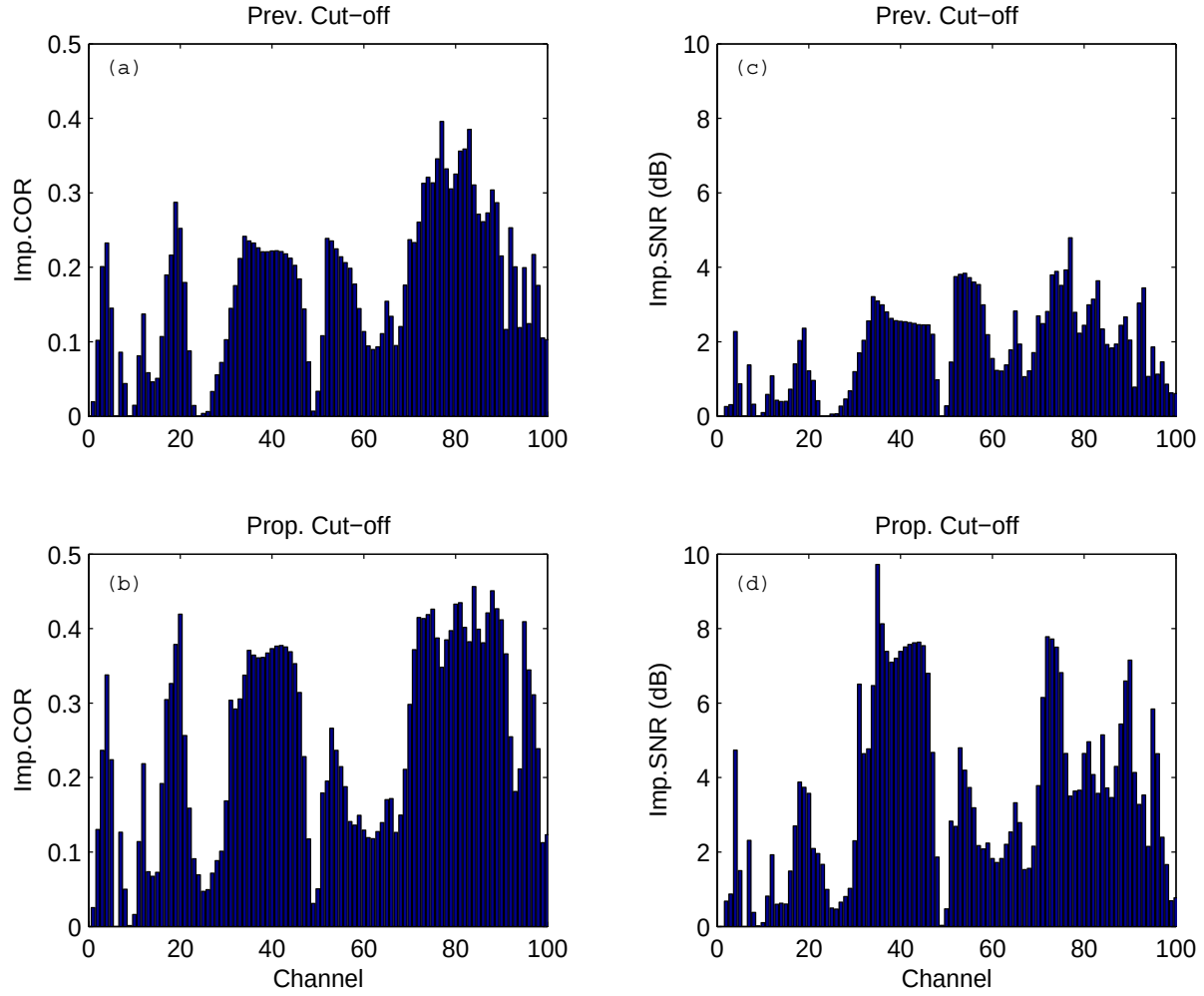


図 4.2: カットオフ周波数の比較：カットオフ周波数 $f_c = 20$ Hz を用いた場合の (a) 相関と (c) SNR の改善度，可変カットオフ周波数を用いた場合の (b) 相関と (d) SNR の改善度．

ここで， $10 \log_{10} |H(\omega)| = -3$ dB のときのカットオフ周波数 F_c を求めると，

$$\begin{aligned}
 10 \log_{10} \frac{1}{\sqrt{1 + (\omega T_R / 13.8)^2}} &= -3, \\
 \log_{10} \{1 + (\omega T_R / 13.8)^2\} &= \frac{3}{5}, \\
 1 + (\omega T_R / 13.8)^2 &= 10^{\frac{3}{5}}, \\
 \omega = 2\pi F_c &= \frac{\sqrt{10^{\frac{3}{5}} - 1}}{T_R / 13.8} \quad (4.3)
 \end{aligned}$$

これより，パワーエンベロープの MTF の関係（式 3.9）が最も成立するためのカットオフ周波数 F_c は以下のように求められる．

$$F_c = \frac{13.8(10^{\frac{3}{5}} - 1)^{\frac{1}{2}}}{2\pi T_R} \quad (\text{Hz}) \quad (4.4)$$

4.3 評価シミュレーション

この可変的なカットオフ周波数の効果を調べるために、評価シミュレーションを行った。シミュレーションに用いた音声信号は、NTT 親密度別単語理解度試験用音声データベース中の話者 (mis) が発話した 12 単語を利用し、これに 50 種類の室内インパルス応答 (式 3.6 の 10 種類の $n(t)$, $T_R = 0.1, 0.3, 0.5, 1.0, 2.0$ s) を畳み込んだ残響音声 $y(t)$ に対して回復処理を行った。このデータベースは、第 7 章の主観評価実験でも用いているため、ここではこのデータベースによる客観的な回復効果を調べた。F0 が事前に正しく推定されているものとし、サンプリング周波数 f_s は 20 kHz とした。従来法と提案法を相関と SNR の平均改善量で比較した。

図 4.3 に結果を示す。従来法に比べて、SNR の改善度も相関の改善度も向上している。特に、従来法では残響時間 T_R が大きくなるにつれ改善量が減少しているのに対して、提案法ではこの改善量の低下が少ない。この結果より、エンベロープ抽出を行う際に用いるカットオフ周波数を、対象となる音声ごとに適した値を定めてやることにより、MTF 理論が成立し、回復精度がさらに向上することがわかった。

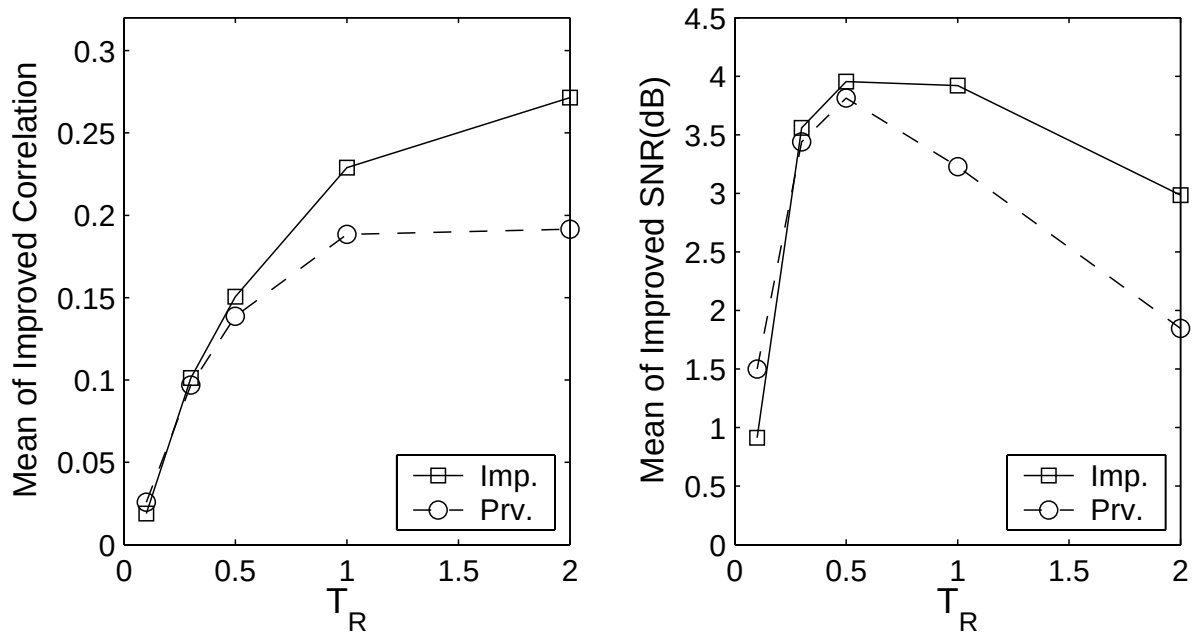


図 4.3: 評価シミュレーション：(a) 相関と (b) SNR の平均改善度，Imp：提案法（可変カットオフ周波数），Prv：従来法（カットオフ周波数 $f_c = 20$ Hz）。

第5章 キャリア再生成法

5.1 キャリア再生成法の原理

キャリア再生成部では、濱上が提案した PIFM (Pulse source Interpolated Frequency Modulation) モデル [22] に基づいて、音源信号を生成する。PIFM モデルを図 5.1 に示す。

5.1.1 原理の説明

ここでは、推定された基本周波数 (F_0) が存在する有声音区間と、 F_0 が存在しない無声音区間に分けてそれぞれで処理を行う。本稿では、 F_0 は事前に正しく推定されているものとして処理を行う。

はじめに、有声音区間のキャリア再生成法の概要を示す。周期的な構造を持つ、すなわち F_0 の情報を持つ有声音は、 F_0 と同値の周波数である基本波と F_0 の倍数からなる複数の調波を足し合わせた波形構造である。また、基本波は時間的に緩やかに変化し、調波もそれに対応しながら緩やかに変化する。そこで、残響音声から推定された F_0 から基本波およびその倍音構造である調波を作成し、それらの波形を足し合わせてできた調波複号音を有声音源信号とする。ここでは、生成された有声音区間の音源信号 $\hat{c}_v(t)$ を次のように表す。

$$\hat{c}_v(t) = \frac{1}{K(t)} \sum_{k=1}^{K(t)} \sin \left(k \int_0^t 2\pi F_0(\tau) d\tau + \phi_k(t) \right), \quad (5.1)$$

ここで、 $F_0(t)$ は基本周波数、 $\phi_k(t)$ は初期位相、 k は調波の次数 ($1 \leq k \leq K(t)$)、そして $K(t)$ は調波の最大次数である。

次に、無声音区間のキャリア再生成法の概要を示す。無声音区間では、 F_0 の情報を持っておらず、非周期的で不規則な波形構造である。そこで、以下のように白色雑音を用いて無声音源 $\hat{c}_u(t)$ を生成する。

$$\hat{c}_u(t) = n(t), \quad (5.2)$$

ここで、 $n(t)$ は白色雑音である。これより、全区間の音源信号 $\hat{c}_x(t)$ は、

$$\hat{c}_x(t) = \hat{c}_v(t) + \hat{c}_u(t) \quad (5.3)$$

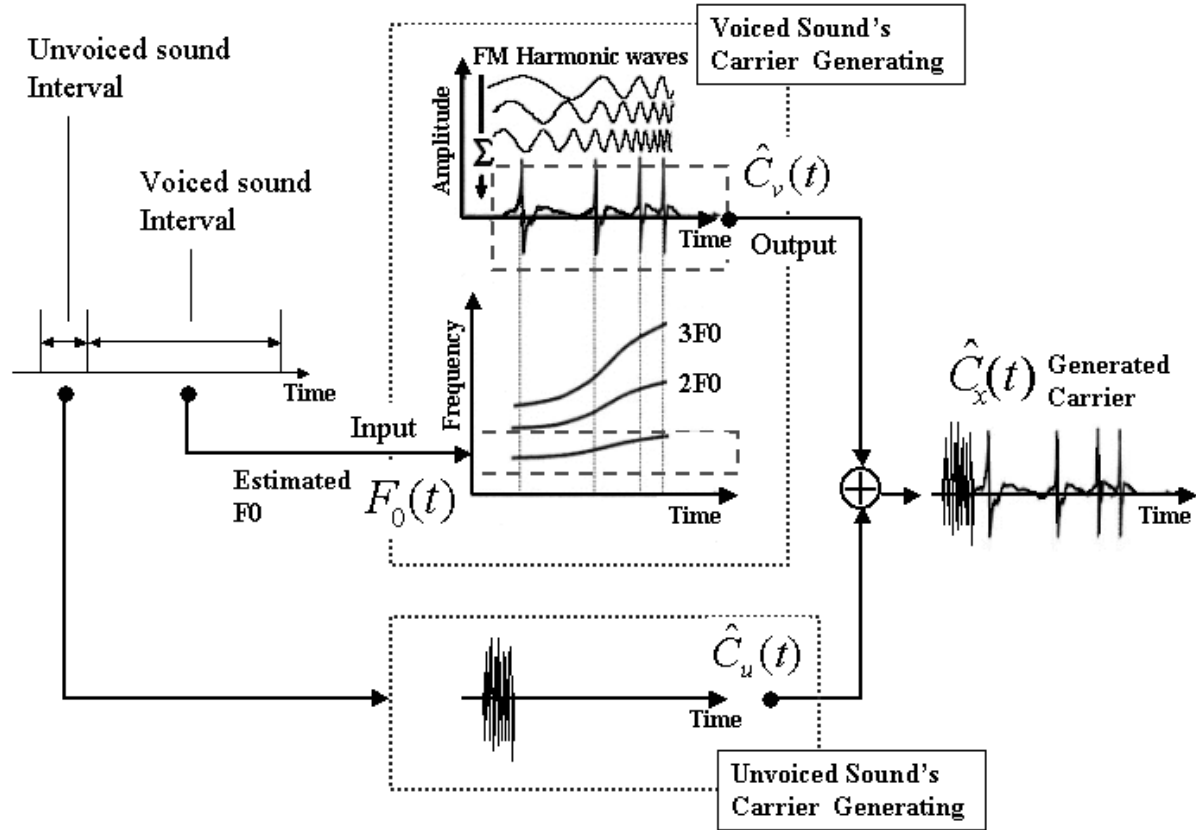


図 5.1: PIFM モデル.

のように、有声区間と無声区間の音源信号の和で求められる。

最終的に、この再生成された音源信号はフィルタバンクによって分解され、各チャネルのキャリア信号となり、回復されたパワーエンベロープと掛け合わせられる。

5.1.2 キャリア再生成法の問題点

従来法 [14, 16] では、式 5.1 における位相項 $\phi_k(t)$ が制御されていなかったため、回復された音声の品質が低下してしまうという問題点があった。そこで、この音質低下を解消するために、キャリア再生成処理の改良を試みる。

5.2 キャリア再生成法の改良

ここでは、音質を向上させるために、キャリアの位相を制御する手法を検討する。本研究では、位相を直接制御せず、位相の周波数微分である群遅延を用いてキャリアを制御した。これは、群遅延が時間領域で直接遅延量に相当しており、直感的に理解しやすく、制

御しやすいからである.

本研究では, 河原らが提案した STRAIGHT[19, 20] 合成系で利用された, 乱数による群遅延操作を行うオールパスフィルタを用いた. これは乱数を用いることによって, 残響音声から正確な位相を抽出することなく位相 (群遅延) 制御を行えるため, 本研究において利用した.

このフィルタは, 振幅特性が 1 で, 位相特性が次式で表される群遅延 $\tau_g(\omega)$ を持つように構成される.

$$\tau_g(\omega) = \rho(\omega) \frac{d_g v(\omega)}{\sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} |v(\omega)|^2 d\omega}} \quad (5.4)$$

$$\rho(\omega) = \frac{1}{1 + \exp(-(|\omega| - \omega_0)/b_w)} \quad (5.5)$$

$$v(\omega) = F^{-1}(W_s(\tau)N(\tau)) \quad (5.6)$$

$$W_s(\tau) = |\tau| \exp(-\pi(\tau/\tau_{bw})^2) \quad (5.7)$$

ここで, $n(\omega)$, $N(\tau)$ はガウス雑音とそのフーリエ変換を, $W_s(\omega)$ と $\rho(\omega)$ はそれぞれ重み関数と周波数加重を示す. 本研究では, 3 kHz 以降の群遅延が大きく変動するという音声の性質から, 境界値 $b_w = 3$ kHz と設定した. また, 標準偏差 d_g は, 音声に歪みが生じず, かつ 最もクリアな音が生成された $d_g = 4$ ms を用いることにした.

求められた群遅延 $\tau_g(\omega)$ を, 周波数上で積分することによって位相特性が求められる. 最終的に, 作成したオールパスフィルタを式 5.1 の有声音源信号 $\hat{c}_v(t)$ に畳み込むことで, キャリア音源の位相制御を行う.

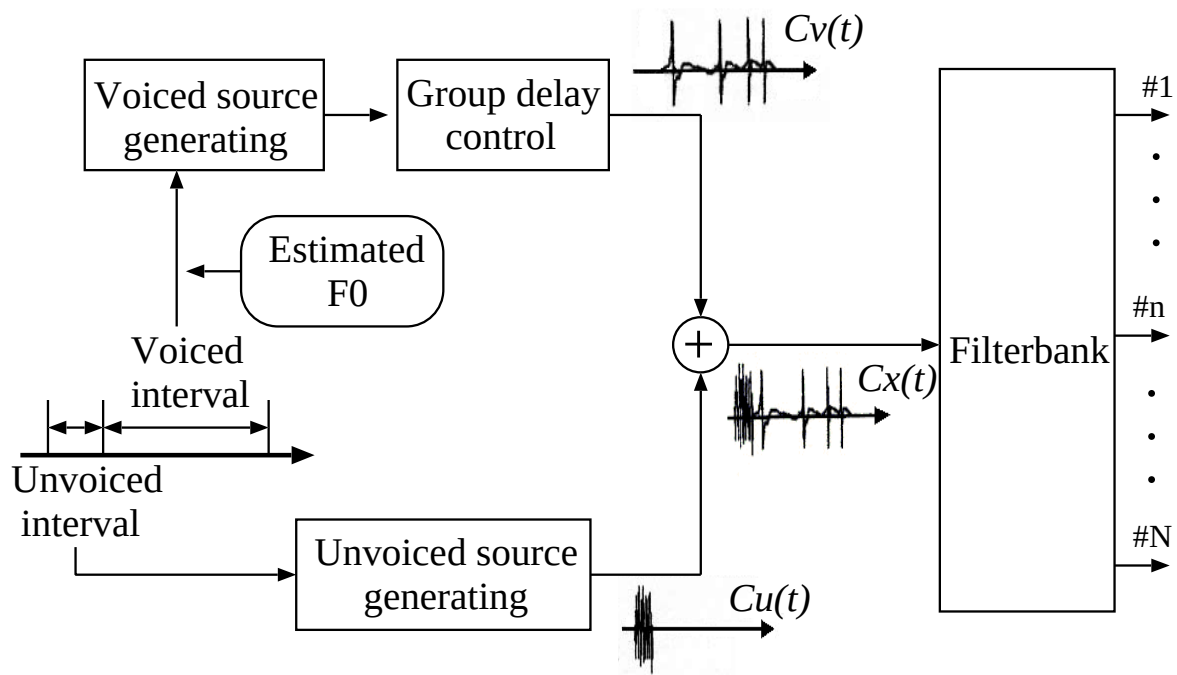


図 5.2: キャリア再生成処理の改良モデル.

第6章 提案モデルの主観的評価

この章では、本研究で提案した残響回復法の主観評価実験を行う。主観評価実験では、まず、音声品質（6.2 節）、残響感（6.3 節）について評価する。そして最後に、提案法によってどれだけ音声明瞭度（6.4 節）を向上できるのかを測定するため、単語了解度試験を行う。

6.1 実験条件

ここでは、NTT 親密度別単語了解度試験用音声データベース中の単語を利用し、人工的なインパルス応答を畳み込んで残響音声を作成した。ここでの残響音声の残響時間は $T_R = 2.0$ s とした。主観評価実験は、正常聴力を有する大学院生 5 名を被験者とし、防音室内で実施した。

6.2 音声品質に対する評価

まず、提案法の音声品質がどれだけ向上しているかを調査するために、Scheffe の一対比較法 [24] により回復音声の品質を評価した。音声信号はデータベースの中の 8 単語を利用し、従来法、提案法、原音声で比較実験を行った。

結果を図 6.1 に示す。この図は、数直線で右に行くほど高品質であることを表している。従来法に比べて、提案法の音声品質が良くなっていることがわかる。従来法による回復音声の劣化が解消されたことによって、音声品質が向上する結果が得られた。これはキャリアの改良が大きく効いていると考えられる。従来法では、キャリアの位相が制御されていなかったため、回復音声の品質が低く、バズ音が生じていた。しかし、キャリアの位相（群遅延）を操作した提案法では、このバズ音が解消され、自然な音声が生産されていることが聴感上判断できる。しかし、従来法に比べると音声品質は向上したものの、原音声と比べると品質は劣っている。また、原音声と比べると合成音特有の音色が感じられる。提案法では、乱数を用いて位相（群遅延）を制御しているため、音質の低下は軽減されたが、原音声と比較すると音色の違いが明確に出ている。これは、STRAIGHT [19, 20] が元来声質変換を目的としているため、原音にそれほど忠実ではないからである。この音色の差異が、原音声に比べて音声品質が劣ってしまった原因の一つと考えられる。

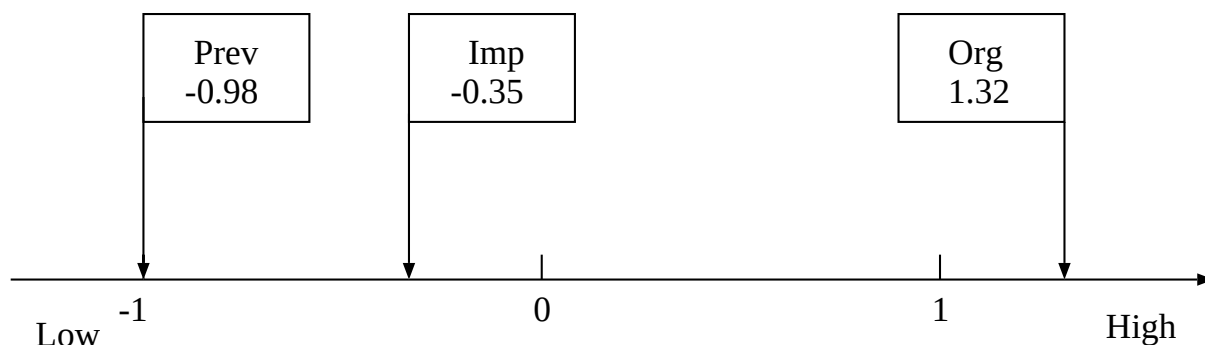


図 6.1: 音声品質, Prev:従来法, Imp:提案法, Org:原音声

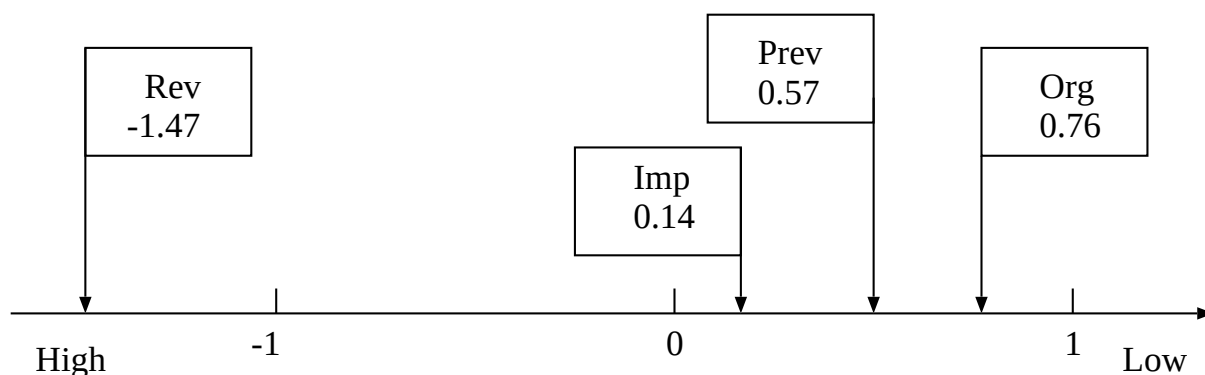


図 6.2: 残響感, Rev:残響音声, Prev:従来法, Imp:提案法, Org:原音声

6.3 残響感に対する評価

次に、残響感がどれだけ解消されているかを測定するために、Scheffe の一対比較法を用いて 6.2 節と同様の音声データで、回復音声の残響感を評価した。

結果を図 6.2 に示す。この図は、数直線で右に行くほど残響感が無いことを表している。残響音声に比べて従来法、改良法ともに残響感が大きく解消されていることがわかる。しかし、従来法に比べると、提案法の方が若干残響感が感じられる結果となった。エンベロープが従来法よりも正確に回復されており、音声の無音区間などがより顕著に回復されているため、提案法の方が残響感も解消されていると予想されたが、従来法よりも残響感が増える結果となった。この理由としては、キャリア部の群遅延を操作したことにより、音声の自然性を高めた分、反対に音声に自然な響き加わったためと考えられる。また、従来法特有のバズ音も被験者に対して、残響感を感じさせない要因の一つになったと考えられる。

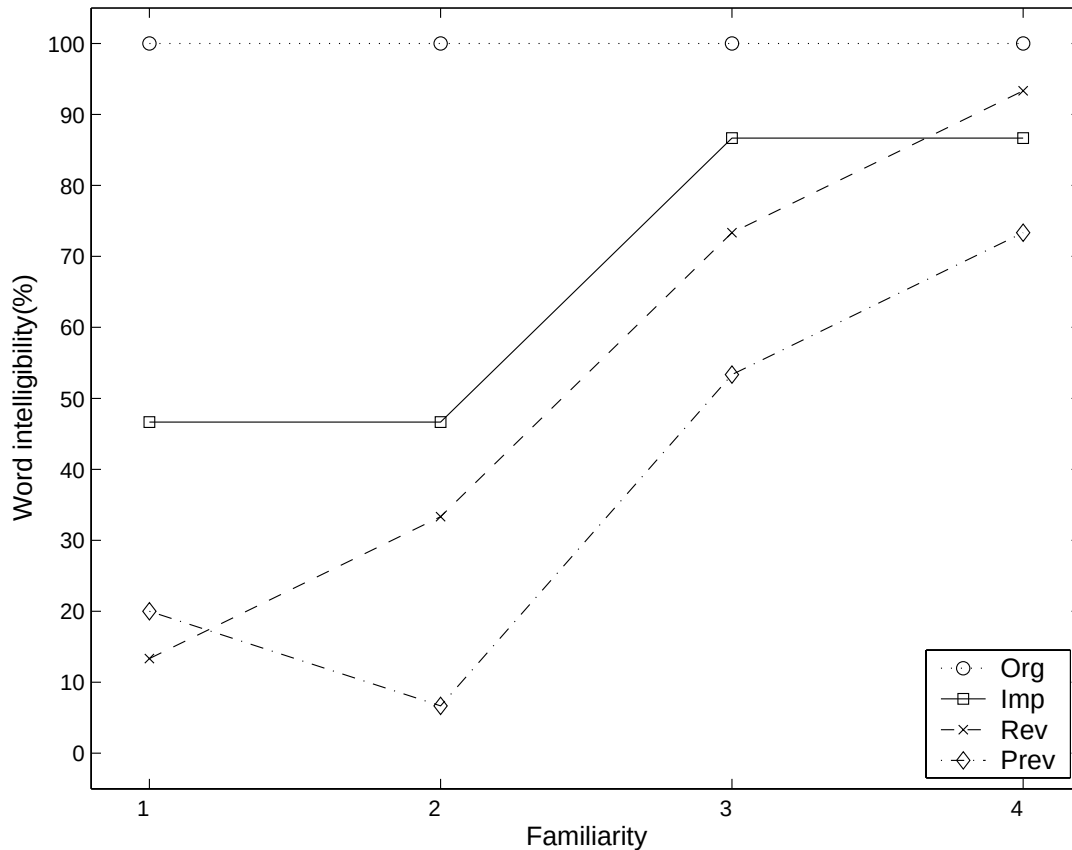


図 6.3: 単語了解度試験.

6.4 音声明瞭度に対する評価

最後に、音声明瞭度について評価を行うため、単語了解度試験を実施した。音声信号はデータベースの中の 12 単語（親密度 4 段階 × 3 単語）を利用し、残響音声と、従来法、提案法、原音声に対して試験を実施した。

図 6.3 に単語了解度試験の結果を示す。図の横軸は親密度を示し、数値が大きくなるほど親密度は高くなる。縦軸に正答率（単語了解度）を示す。結果を見ると、正答率は親密度に依存し、どの音声の場合でも、親密度が低くなるにつれ、正答率も低下する傾向にあるということがわかる。また、従来法や残響音声と比較して、提案法の正答率が高いことがわかる。特に低い親密度で正答率の差がより大きく出ており、親密度が 1 の場合では、20 % 以上の差が出ている。逆に、従来法では、明瞭度が残響音声よりも低い結果となった。

6.5 まとめ

提案法の回復精度を主観評価により総合的に評価した。その結果、本研究の目標である音声明瞭度が向上した。特に、親密度が低い音声でも従来法や残響音声に比べて高い明瞭度が得られた。また、音声品質に関しても、従来法より高い改善が得られた。これは、キャリアの位相を制御することによって音質が向上した結果である。その反対に、残響感に関しては、従来法よりもやや劣る結果となった。

これらの結果から考察すると、音声品質の改善が音声明瞭度に大きく貢献していると考えられる。もちろん残響感が大きいと明瞭度は低下してしまうが、残響感が改善した従来法の結果を見ると、やはり音声品質が悪いと明瞭度は向上しないことがわかった。そのため、残響感のみ改善した従来法では、残響音声よりも明瞭度が低下したのに対し、残響感と音声品質共に改善した提案法では、明瞭度が向上する結果となった。

第7章 結論

7.1 本論文で明らかにしたこと

本研究では、残響によって低下した音声明瞭度を向上させるための残響回復法を提案した。提案モデルでは、音声をパワーエンベロープとキャリアに分割し、それぞれに対して回復処理を行った。パワーエンベロープ回復部では、酒田らの提案したパワーエンベロープ回復法に対して、適応的な時間-周波数分割処理を適用することにより、パワーエンベロープの回復精度をさらに向上できることがわかった。また、パワーエンベロープ抽出の際に利用していたカットオフ周波数を、各音声ごとに可変的に設定してやることにより、残響時間が長い音声に対しても効果的な回復が得られた。キャリア再生部では、群遅延を操作することによって音源信号の位相を制御する手法を検討した。その結果、回復音声の音質を大きく改善することができた。

提案モデルがどれだけ残響回復できるかを総合的に評価するため、音声品質、残響感、そして単語了解度について主観評価実験を行った。その結果、残響音声に比べて大きく残響感が解消され、また従来法よりも音声品質が向上する結果が得られた。それに伴い、提案モデルが、残響によって低下した音声明瞭度を回復できることも了解度試験からわかった。これより、主観的评价においても客観的评价においても、本研究の提案モデルの高い有効性が示された。

7.2 今後の課題

本研究で残った課題について説明する。

● 実環境下での残響回復精度の調査

今回使用した残響音声は、人工的に作成した室内インパルス応答を原音声に畳み込んで作成したため、その残響音声自体の明瞭度があまり低下しなかった。また、残響時間が 2.0 s より短い残響音声の場合はほとんど聞き取れてしまう。今後は、実環境で収録した残響音声に対して本手法での回復精度を検討する予定である。実環境で収録された残響音声を用いれば、残響音声の明瞭度は低下し、さらに提案法との回復効果の差が顕著に現れると予想される。

● パワーエンベロープ回復精度の向上

残響回復処理を行った際に、回復音声の音韻性が崩れ、異調を起こしてしまう場合がある。特に、残響時間の長い残響音声を回復させた場合によく生じてしまう。試験的にエンベロープを原音声のものをを用いて、回復音声を作成した場合、このような異調は生じない。つまり、エンベロープの回復にまだ問題が生じていると考えられる。

例として、mis, /tarukamu/という音声について検討してみる。そのスペクトルグラムを図 7.1 に示す。これは、キャリアは全て同じものを用いて、エンベロープのみを上から、原音声エンベロープ、残響エンベロープ（残響時間 2.0 s）、回復エンベロープに変えて再構築した音声のスペクトルグラムである。この回復させた音声を聞くと、/tarutau/のように聞こえてしまう。この図を見ると、0.35 s 付近に存在する子音部/k/において、原音声では 1 kHz 以上のパワーは見られないのに対し、回復音声では 1 kHz から 4 kHz までのパワーが大きく存在している。これによって、子音/k/が/t/に異調されてしまう原因となる。

試験的に、低域（1-3 kHz）のみ原音声のエンベロープを用い、それ以降の高域を回復エンベロープを用いて音声を作成した。そうすると、正確に/tarukamu/と聞こえるようになった。この試験的に作成した音声のスペクトルグラムを図 7.1 の一番下に示す。このように、エンベロープの回復処理によって、定量的には原音声のエンベロープに近づくが、回復が充分で無いために、知覚的に音韻性が壊れてしまうことが生じている。今後さらに明瞭度を向上させ、原音声に近づけるためには、エンベロープの回復をさらに精度の高いものとし、異調を起こさないようにする必要がある。

● F0 情報を用いない残響回復モデルの検討

現在の残響回復モデルでは、事前に F0 を正確に推定できているという仮定で、キャリア生成処理を行っている。残響音声から F0 を抽出する手法は、今だ提案されておらず、非常に難しいとされている。そのため、この F0 を用いずにキャリアを生成する手法を検討する必要がある。また、これに関連して、音声の有声/無声区間の検出法も検討する必要がある。

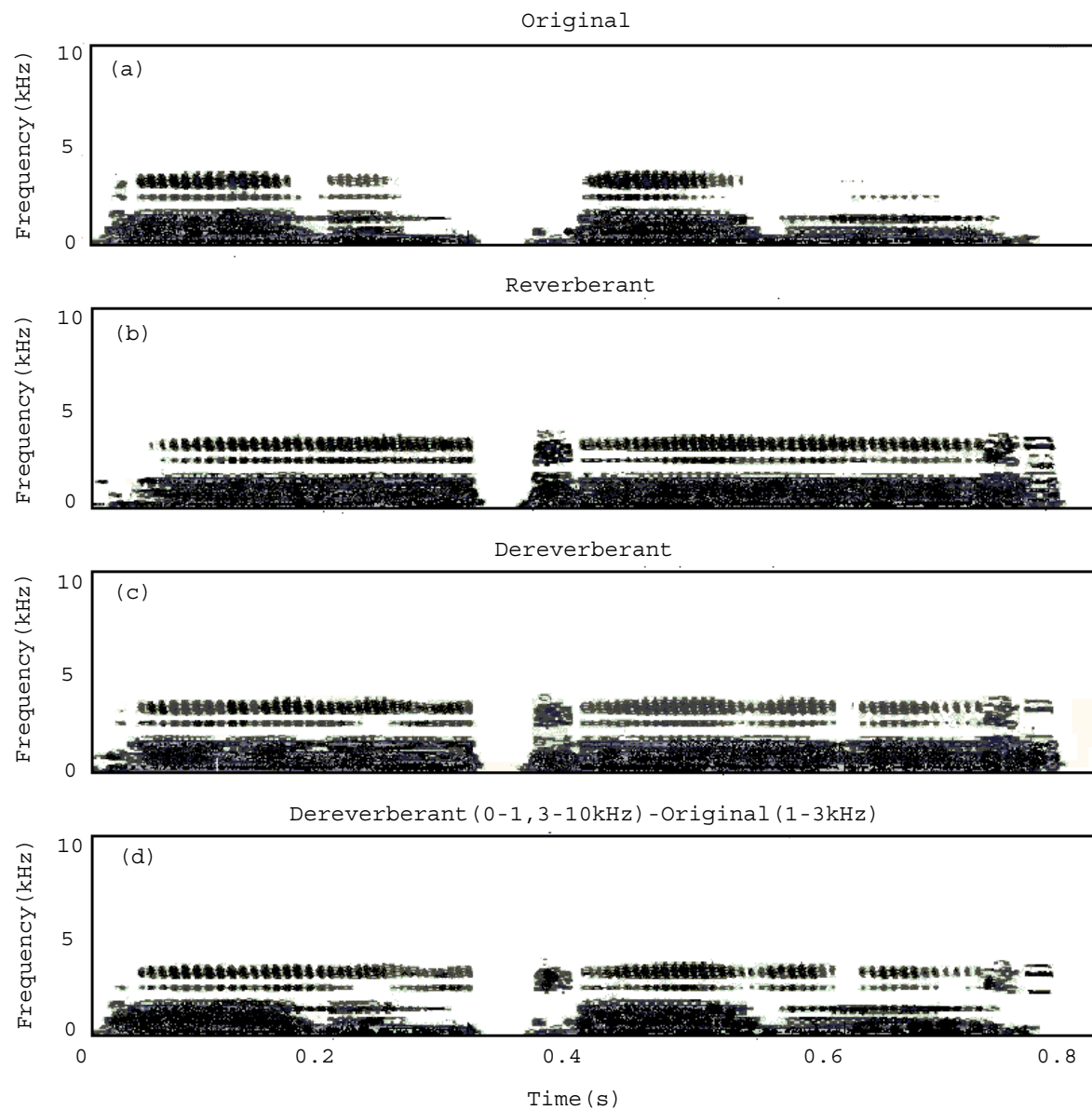


図 7.1: mis,/tarukamu/のスペクトルグラム：(a) 原音声エンベロープ，(b) 残響音声エンベロープ，(c) 回復音声エンベロープ，(d) チューニングされたエンベロープを用いて作成された音声

謝辞

本研究を進めるにあたり，多大な御助言を頂き熱心な御指導をして頂きました赤木正人教授に心から感謝致します．本研究を進めるにあたり，多大な御助言を頂き熱心な御指導をして頂きました党建武教授に心から感謝致します．本研究に関して，多大な御助言，御討論をして頂いた鶴木祐史助手に心より感謝致します．本研究に関して，多大な御助言をして頂いた研究員の伊藤一仁氏，石本祐一氏，博士後期課程の西本博則氏，斎藤毅氏，李軍鋒氏，羽二生篤氏，黄純芳氏に心より感謝致します．また，本研究を進めるにあたり有意義な討論並びに，有益な助言を賜った赤木・党研究室の皆様方に感謝致します．

参考文献

- [1] Houtgast, T. and Steenken, H. J. M., “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” J. Acoust. Soc. Am. vol.77, No.3, pp. 1069-1077, 1985.
- [2] Houtgast, T. and Steenken, H. J. M., “The Modulation Transfer Function in room acoustics as a predictor of speech intelligibility,” Acustica. vol.28, pp. 66-73, 1973.
- [3] Drullman, R., “Temporal envelope and fine structure cues for speech intelligibility,” J. Acoust. Soc. Am. vol.97, No.1, pp. 585-591, 1995.
- [4] Neely S. T. and Allen J. B., “Invertibility of a room impulse response,” J. Acoust. Soc. Am. Vol.66, No.1, July 1979.
- [5] Miyoshi, M. and kaneda, Y., “Inverse filtering of room acoustics,” IEEE Trans. ASSP, Vol.36, No.2, pp.145-152, Feb.1988.
- [6] Wang, H. and Itakura, F., “Realization of acoustic inverse filtering through multi-microphone sub-band processing,” IECE Trans. Fundamentals, vol.E75-A, no. 11, pp. 1474-1483, Nov. 1992.
- [7] Nakatani, T. and Miyoshi, M., “Blind dereverberation of single channel speech signal based on harmonic structure,” Proc. ICASSP2003, vol. I, pp. 92-95, 2003.
- [8] Langhans, T. and Strube, H. W., “Speech enhancement by nonlinear multiband envelope filtering,” Proc. ICASSP82, pp. 156-159, 1982.
- [9] Avendano, C. and Hermansky, H., “Study on the dereverberation of speech based on temporal envelope filtering,” Proc. ICSLP, pp.889-892,1996.
- [10] Mourjopoulos, J. and Hammond, J. K., “Modelling and enhancement of reverberant speech using an envelope convolution method” Proc. ICASSP83, pp. 1144-1147, 1983.
- [11] 広林 茂樹, 野村 博昭, 小池 恒彦, 東山 三樹夫, “パワーエンベロープ伝達関数の逆フィルタ処理による残響音声の回復,” 電子情報通信学会論文誌 A, Vol. J81-A, No.10, pp.1323-1330, 1998.

- [12] 広林 茂樹, 山淵 龍夫, “帯域分割を用いたパワーエンベロープ逆フィルタ処理の残響抑圧効果,” 電子情報通信学会論文誌 A, Vol. J83-A, No.8, pp. 1029–1033, 2000.
- [13] 古川 正和, 鶴木 祐史, 赤木 正人, “MTF に基づいた残響音声パワーエンベロープの回復方法,” 電子情報通信学会技術研究報告, SP2002-15, pp.49-54, 2002.
- [14] 酒田 恵吾, 鶴木 祐史, 赤木 正人, “MTF に基づいた残響音声の回復法の検討,” 電子情報通信学会技術研究報告, SP2003-03, pp.13-18, 2003.
- [15] Unoki, M., Sakata, K., Furukawa, M., and Akagi, M., “A method based on the MTF concept for dereverberating the power envelope from the reverberant signal,” Proc. of ICASSP2003, vol. I, pp. 840-843, 2003.
- [16] Unoki, M., Sakata, K., and Akagi, M., “A speech dereverberation method based on the MTF concept,” Proc. Eurospeech2003, pp. 1417-1420, 2003.
- [17] Unoki, M., Furukawa, M., Sakata, K., and Akagi, M., “An improved method based on the MTF concept for restoring the power envelope from a reverberant signal,” Acoust. Sci.&Tech. vol. 25, no.4, 2003.
- [18] Unoki, M., Sakata, K., Furukawa, M., and Akagi, M., “A speech dereverberation method based on the MTF concept in power envelope restoration,” Acoust. Sci.&Tech. vol. 25, no.4, 2003.
- [19] 河原 英紀, “高品質音声分析変換合成法 STAIGHT”, ATR 人間情報通信研究所, 和歌山大学, 平成 13 年 1 月 14 日
- [20] 河原 英紀, 津崎 実, Roy, D, Patterson., “オールパスフィルタの位相操作による時間構造制御とその知覚への影響について,” 日本音響学会聴覚研究会資料, H-96-79, 1996.
- [21] 坂野 秀樹, 陸 金林, 中村 哲, 鹿野 清宏, 河原 英紀, “時間領域平滑化群遅延を用いた短時間位相の効率的表現方法,” 電子情報通信学会論文誌, Vol. J84-D-2, pp. 621-628, 2001.
- [22] 濱上 知樹, “音源波形形状を高調波位相により制御する音声合成方式,” 日本音響学会誌, Vol. 54, No. 9, pp.623-631, 1998.
- [23] 加藤 和美, 天野 成昭, 近藤 公久, “雑音を付加した音声の単語了解度に対する親密度の影響,” 日本音響学会聴覚研究会資料, H-99-8, 1999.
- [24] 難波 精一郎, 桑野 園子, “音の評価のための心理学的測定法,” コロナ社, 1998.

- [25] Takeda, T., Sagisaka, Y., Katagiri, S., Abe, M., and Kuwabara, H., Speech Database User's Manual, ATR Technical Report, TR-I-0028, 1988.
- [26] Database for speech intelligibility testing using Japanese word lists. NTT-AT, March 2003.

研究業績

- [1] 鵜木 祐史, 戸井 真智, 赤木 正人, “残響音声のパワーエンベロープ回復処理の改良,” 日本音響学会講演論文集, vol. I, pp. 609-610, 2004 年 3 月.
- [2] Unoki, M., Sakata, K., Toi, M., and Akagi, M., “Speech deereverberation based on the concept of the Modulation Transfer Function,” NCSP2004, pp. 423-426, 2004.
- [3] Unoki, M., Toi, M., and Akagi, M., “A speech deereverberation method based on the MTF concept using adaptive time-frequency divisions,” Proc. EUSIPCO2004, pp. 423-426, 2004.
- [4] 戸井 真智, 鵜木 祐史, 赤木 正人, “適応的な時間-周波数分割を考慮した残響音声回復法の改良,” 日本音響学会聴覚研究会資料, H-2004-66, 2004.
- [5] 戸井 真智, 鵜木 祐史, 赤木 正人, “残響音声パワーエンベロープ回復法における最適な時間-周波数分割の検討,” 日本音響学会講演論文集, vol. I, pp. 643-645, 2004 年 9 月.
- [6] Toi, M., Unoki, M., and Akagi, M., “Development of adaptive time-frequency divisions and a carrier reconstruction in the MTF-based speech dereverberation method,” NCSP2005, 2005 (accepted).
- [7] 戸井 真智, 鵜木 祐史, 赤木 正人, “残響音声回復法におけるキャリア再生成処理の改良とその評価,” 日本音響学会講演論文集, 2005 (to be appear).