修 士 論 文

残響音声からの音声特徴量抽出法と音源波形再合成に関する研究

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

酒田 恵吾

2003年3月

修 士 論 文

残響音声からの音声特徴量抽出法と音源波形再合成に関する研究

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　小谷一孔 助教授
審査委員　下平博 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

110049 酒田 恵吾

提出年月: 2003 年 2 月

概 要

残響は音声に歪みを与える原因となる。遠隔会議システムや音声認識において残響の影響を抑圧することは大きな課題である。また残響の影響を抑圧する手法は、時間変動する残響特性に適用的できることが望まれる。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を手法を用いた手法が多く提案されている [3],[4]。これらの手法は、時間変動する室内伝達特性をその都度正確に計測する必要があり、実用化は困難である。

古川らは MTF の理論に基づき [5],[6],[7]、室内伝達特性を測定せず、観測信号の情報のみからパワーエンベロープを回復する手法を提案している [10]。しかしその手法には解決すべき問題が残されている。１つは帯域分割幅の検討の際、MTF 理論成立/不成立の検討がなされていないこと、もう１つは低帯域の回復効果が小さいことである。

また時間波形としての回復を考えると、キャリアに関しての処理も考える必要がある。本研究では、室内伝達特性を測定せず、観測した残響音声の情報のみから残響音声の回復処理を行う手法を提案する。音声信号をエンベロープとキャリアでモデル化し、それぞれに分けて回復処理を行う。エンベロープの処理では、古川らのパワーエンベロープ回復法の問題点を解決し、改善した手法を用いる。キャリアの処理では、残響音声中から F0 が推定されたと仮定して、その F0 情報を基にキャリアを再合成する処理を提案する。

パワーエンベロープ回復法の適切な帯域分割幅の決定に関する問題では、狭帯域内でのパワーエンベロープの共変調、MTF 理論成立/不成立の二点に着目し、その結果、300 から 400 Hz が適切な帯域分割幅であるとみなした。また低帯域の回復効果に関する問題では、低帯域では音声間の無音区間が長い場合が多く存在し、従来の回復法はその場合に適用できないことを明らかにした。そして音声間の無音声区間が長い場合でも適用できる回復法を、提案し、低帯域の回復効果を上げることができた。キャリア再合成処理では、F0 からキャリアを作成する手法を提案し、音声の特徴をもつキャリアを再合成することができた。また提案モデルの評価のためのシミュレーションを行い、その有効性を示した。

目 次

図 目 次

表 目 次

第1章　序論

1.1　背景

部屋の中やコンサートホールのように壁や天井で囲まれた空間内で音が放射されると、受音点では音源から直接伝搬される音 (直接音) の他に、壁、天井などの障害物により反射された音 (反射音) も含まれる。この現象は残響と呼ばれる。そして直接音と残響による反射音が重なり合って受音された音は残響音と呼ばれる。 残響は、音声に歪みを与える原因となる。ハンズフリーマイクロフォンを用いた遠隔会議システムでは、話者とマイクロフォン間にある程度の距離があると、受音された音声は残響の影響を受け、歪みを生じる。人間はその歪んだ音声を聴くと不明瞭に感じる。また残響は音声認識器の認識精度を低下させる原因の一つに挙げられている [1, 2]。故に、音声を歪ませる原因である残響の影響を抑圧することは大きな課題である。また残響は、室温、障害物、音源と受音点の位置や数など、空間内のあらゆる状況に依存した時変的な特性を持つ。残響の影響を抑圧する手法としては、この特性に適応できることが望まれる。

これまで残響の影響を抑圧する手法が多く提案されている。まず室内伝達特性の逆フィルタを用いた手法が挙げられる。S.T.Neely, J.B.Allen らは単一マイクロフォ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くことで、回復信号を求める手法を提案している。しかし室内伝達系が最小位相特性を有していないと、提案法の回復精度が低下する [3]。また三好, 金田らは、音源の数に対しマイクロフォンを一つ以上多く配置し、室内伝達特性の零点が重複しない条件ならば、室内伝達系が非最小位相特性の場合でも残響音声を回復できる手法 (MINT 法) を提案している [4]。しかしこれらの手法は、時間変動する室内伝達特性をその都度正確に計測しなければ回復精度が下がる。時間変動する度に処理を行う必要があるこれらの手法の実用化は困難である。

一方、室内伝達特性の測定を必要としない手法が提案されている。広林らは、MTF の理論 [6],[7] に基づき信号をエンベロープとキャリアでモデル化し、室内伝達特性を測定せずにパワーエンベロープのみの回復を行う、パワーエンベロープ逆フィルタ法を提案している [8],[9]。しかしこの手法は、パワーエンベロープをどう抽出するか、残響時間などのパラメータをどう推定するか、音声信号に適用できるか、などの問題があった。古川らはこれらの問題に対する検討および、改善方法を提案した。古川らが改善したパワーエンベロープ回復法は、観測した残響音声の情報のみから音声のパワーエンベロープの回復を行うことができる。しかしその手法には、解決すべき問題が残されている。帯域分割処理で帯域分割幅を決定する際に、パワーエンベロープ間の共変調に対する検討はされてい

るが MTF 理論成立に対する検討はされていない、また低帯域内でパワーエンベロープの改善効果が得られないてないという問題が挙げられる。また、これらの問題が解決されたとしても、パワーエンベロープのみの回復であるこの手法は、音声認識器の場合にしか適用できない。人間が聴くという点で考えると、時間波形として回復する手法を考える必要がある。そのためにはエンベロープだけでなく、キャリアに関しての処理も考える必要がある。

1.2 目的

本研究では、室内伝達特性を測定せず、観測した残響音声の情報のみから残響音声の回復処理を行う手法を提案する。音声信号をエンベロープとキャリアでモデル化し、それぞれに分けて回復処理を行う。エンベロープの処理では、古川らのパワーエンベロープ回復法の問題点を解決し、改善した手法を用いる。キャリアの処理では、残響音声中から F0 が推定されたと仮定して、その F0 情報を基にキャリアを再合成する処理を提案する。そのために以下の仮定を設ける。

- 元の音声信号の F0 および、F0 の存在する有音声区間と F0 の存在しない無音声区間は既知とする。

最後に、それぞれの処理を行ったエンベロープとキャリアから、音声信号を合成することで、時間波形としての残響音声の回復を行う。

本手法を実現できれば、室内伝達特性を測定する必要がなく、残響音声の回復処理を行うことができ、音声認識器の認識精度向上や遠隔会議システムなどでの音声明瞭度向上に貢献することができる。

1.3 本論文の構成

本論文は全 6 章により構成される。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部で用いるパワーエンベロープ回復法の原理とその問題点について説明し、問題点の検討、解決策について述べる。第 4 章ではキャリア再合成法の原理について述べる。 第 5 章では提案モデルの有効性を示すために、評価シミュレーションについて述べる。第 6 章では、本研究のまとめと今後の課題について説明する。

第2章　提案モデルの概要

提案モデルの概要図を図 2.1 に示す。 モデルへの入力は観測された残響音声のみである。そして本モデルは大きく二つにわけられる。一つはエンベロープ回復部、もう一つはキャリア再合成部である。この二つの部からそれぞれ出力されるエンベロープとキャリアを用いて、元の音声の再合成処理を行う。

2.1　エンベロープ回復部

エンベロープ回復部では、観測した残響音声のパワーエンベロープの回復処理を行うことを目的としている。

回復方法としては、古川らが提案しているパワーエンベロープ回復法を用いる [10]。パワーエンベロープ回復法の原理および、回復法の問題点の詳しい説明は第 3 章で行う。

2.2　キャリア再合成部

キャリア再合成部では、残響音声中から推定された F0 の情報を基に、元の音声信号のキャリアの再合成処理を行うことを目的としている。

キャリアの再合成処理には、F0 の情報が必要不可欠である。一般に音声のキャリアの大部分は周期的な構造である。その周期の単位秒あたりの変動数は基本周波数 (F0) と呼ばれる。F0 は STRAIGHT[11] などの音声合成の分野で、自然な合成音を生成するための重要な特徴量として扱われている。

残響音声から元の音声の F0 を推定できれば、元の音声のキャリアを作成できる可能性がある。近年、雑音に頑健で高精度な F0 推定法が盛んに提案されており [12,]、これらの成果から将来は残響音声中から基本周波数を推定できる手法が提案されると期待できる。

本研究では F0 および有声音/無声音区間は既知と仮定した上で、キャリア再合成処理法を提案する。その処理法の原理については第 4 章で説明する。

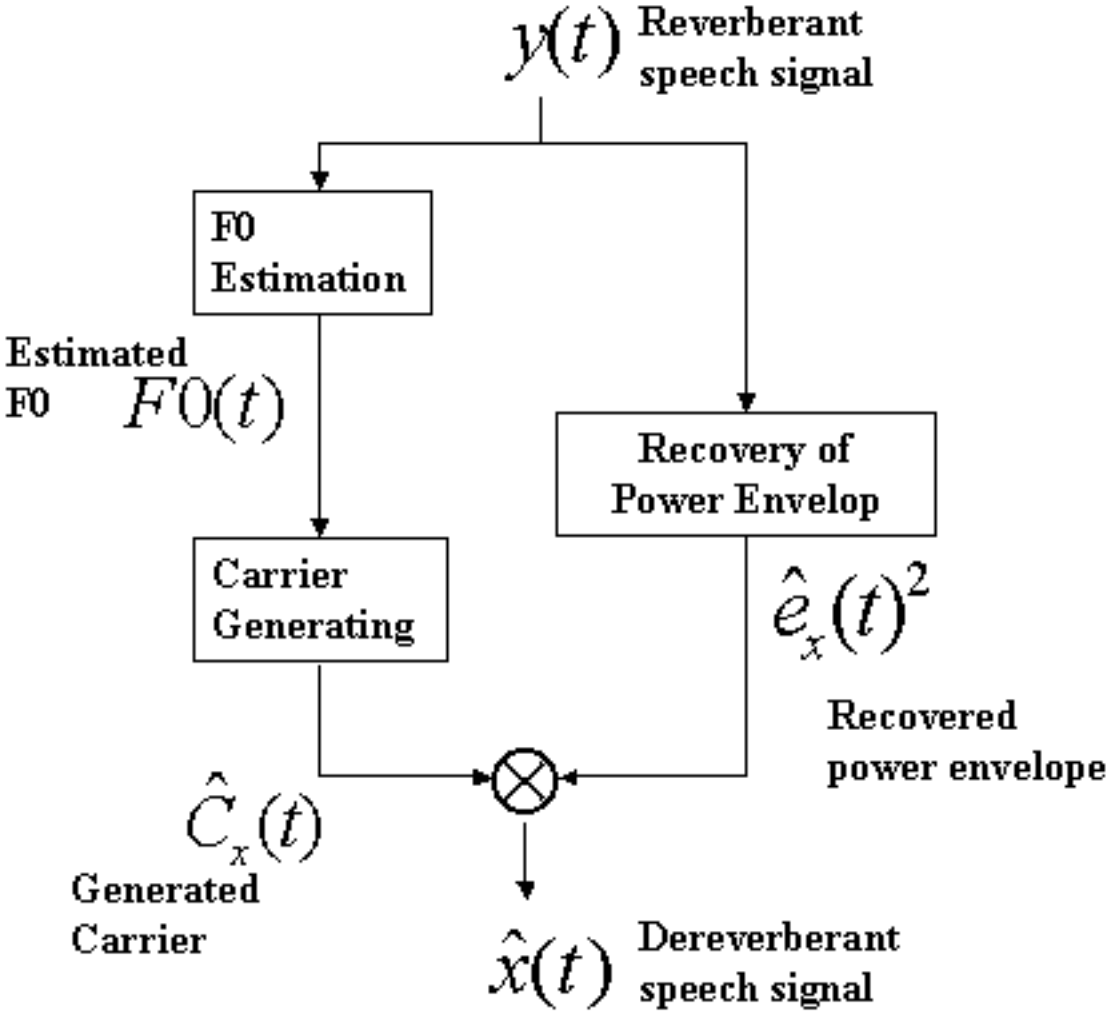

図 2.1: 提案モデルの概要

第3章　パワーエンベロープ回復法

この章では、本研究で提案する残響音声モデルのエンベロープ回復部で用いられる、パワーエンベロープ回復法の原理について説明する。またパワーエンベロープ回復法の問題点を挙げる。そして各問題点に対する検討を行う。

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)

変調伝達関数 (MTF) は室内の残響特性による音声明瞭度の低下を測定する尺度として用いられている [5],[6],[7]。その尺度は音声のパワーエンベロープに着目し、その歪みを評価している。変調伝達関数 $m(\omega)$ は以下の式で表される。

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

$h(t)$ はインパルス応答である。ここで $h(t)$ を残響特性を表す式として、指数減衰するエンベロープと白色ガウス過程から生起した雑音 $n(t)$ から成るキャリアをもつ室内インパルス応答とすると、以下のように定義する。

$$h(t) = e^{-\frac{t}{\tau}} n(t) = e^{-\frac{6.9t}{T_R}} n(t) \tag{3.2}$$

この式を MTF の式 3.1 に代入すると、以下の式が得られる。

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

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。この式は、残響の影響に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。

ここで一例を示す。図 3.1(a) の変調周波数 10 Hz の正弦波に白色雑音を振幅変調した信号 (変調度 m=1) を図 3.1(c) の残響インパルス応答 (残響時間 T_R=0.5) に畳み込んで得られた残響信号を図 3.1(e) に表す。各信号のパワーエンベロープを図 3.1(b),(d),(f) に表す。残響の影響を受けることで信号の変調度は 1 から 0.4 と減少している。一方、式 (3.3) から得られる値は m=0.402 であり、残響の影響を受けた後の変調度とほぼ等しい。

このように、パワーエンベロープの歪みが室内の残響特性による場合、MTF よりにどれだけ残響の影響を受けたかを知ることができる。

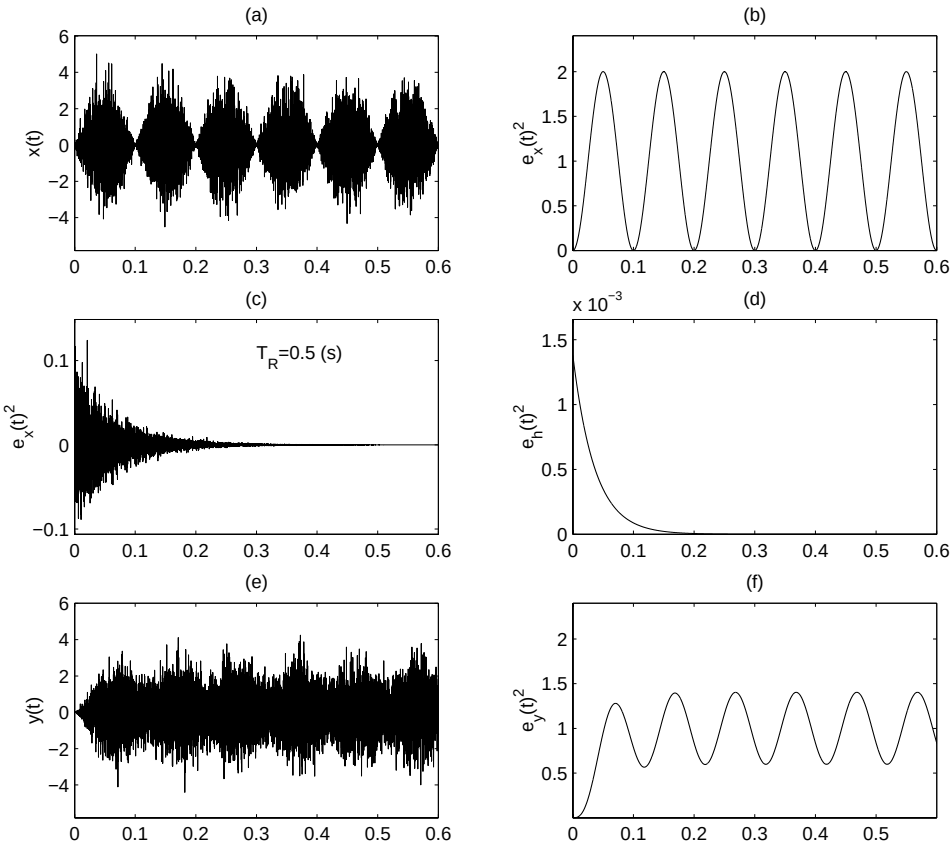

図 3.1:　MTF に基づいたパワーエンベロープの関係。(a) 残響の影響を受ける前の信号 $x(t)$、(b)$x(t)$ のパワーエンベロープ $e_x(t)^2$、(c) 残響インパルス応答 $h(t)$、(d)$h(t)$ のパワーエンベロープ $e_h(t)^2$、(e) 残響の影響を受けた信号 $y(t)$、(f)$y(t)$ のパワーエンベロープ $e_y(t)^2$

3.1.2　パワーエンベロープ逆フィルタ法

広林らは MTF 理論に基づき、残響信号のパワーエンベロープを回復するパワーエンベロープ逆フィルタ法を提案している [8]。

MTF 理論に基づき音源信号 $x(t)$、インパルス応答 $h(t)$、残響信号 (観測信号) をエンベロープとキャリアに分け、以下の式のように定義できる。

$$x(t) = e_x(t)n_1(t) \tag{3.4}$$

$$h(t) = e_h(t)n_2(t) = e^{-\frac{6.9t}{T_R}}n_2(t) \tag{3.5}$$

$$y(t) = x(t) * h(t) \tag{3.6}$$

$$< n_k(t), n_k(t-\tau) > = \delta(\tau) \tag{3.7}$$

$*$ は畳み込み積分、$e_x(t), e_h(t)$ は各信号のエンベロープ、$< \cdot >$ は集合平均、$n_1(t)$, $n_2(t)$ は互いに無相関なキャリアである。a, T_R は室内インパルス応答のパラメータ振幅項と残

響時間である。

ここで $y(t)$ の集合 2 乗平均を求める。

$$< y(t)^2 > \;=\; < \{\int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau\}^2 >$$

$$\begin{aligned} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e_x(\tau_1)e_x(\tau_2)e_h(t-\tau_1)e_h(t-\tau_2) \\ &\quad < n_1(\tau_1)n_1(\tau_2) >< n_2(t-\tau_1)n_2(t-\tau_2) > d\tau_1 d\tau_2 \\ &= \int_{-\infty}^{\infty} e_x(\tau)^2 e_h(t-\tau)^2 d\tau \\ &= e_x(t)^2 * e_h(t)^2 \end{aligned} \tag{3.8}$$

また $< y(t)^2 >$ は $< y(t)^2 >=< e_y(t)^2 n(t)^2 >= e_y(t)^2$ となることから、以下の式が得られる。

$$e_y(t)^2 \;=\; e_x(t)^2 * e_h(t)^2 \tag{3.9}$$

この式は残響信号のパワーエンベロープがインパルス応答のパワーエンベロープと音源信号のパワーエンベロープの畳み込みで得られることを意味する。$e_h(t)^2$ を z 変換したパワーエンベロープ伝達特性を $P_h(z)$ とおくと、$P_h(z)$ は以下のように定義できる。

$$\begin{aligned} P_h(z) &= a^2 + a^2\alpha z^{-1} + a^2\alpha^2 z^{-2} + a^2\alpha^3 z^{-3} + ... \\ &= \frac{a^2}{1-\alpha z^{-1}} \end{aligned} \tag{3.10}$$

$\alpha = e^{-\frac{13.8T_s}{T_R}}$、$T_s$ はサンプリング周期である。この式から元の音源信号のパワーエンベロープ特性 $P_x(z)$ は、以下の式で求めることができる。

$$\begin{aligned} P_x(z) &= \frac{P_y(z)}{P_h(z)} \\ &= \frac{1-\alpha z^{-1}}{a^2} P_y(z) \end{aligned} \tag{3.11}$$

$P_x(z)$ の逆変換を求めることで音源信号のパワーエンベロープ $e_x^2(t)$ を得ることができる。

以上から、観測した残響信号のパワーエンベロープ $e_y(t)^2$ は入力信号、残響パラメータ a, T_R を決定できれば、式 3.11 を用いて回復処理を行うことができる。

図 3.1(f) の残響信号パワーエンベロープに逆フィルタ処理を行った後の概形を図 3.2 に示す。この逆フィルタ処理は残響時間パラメータ T_R の値を大きくするほどパワーエンベロープのピークとディップを強調させる働きがある。

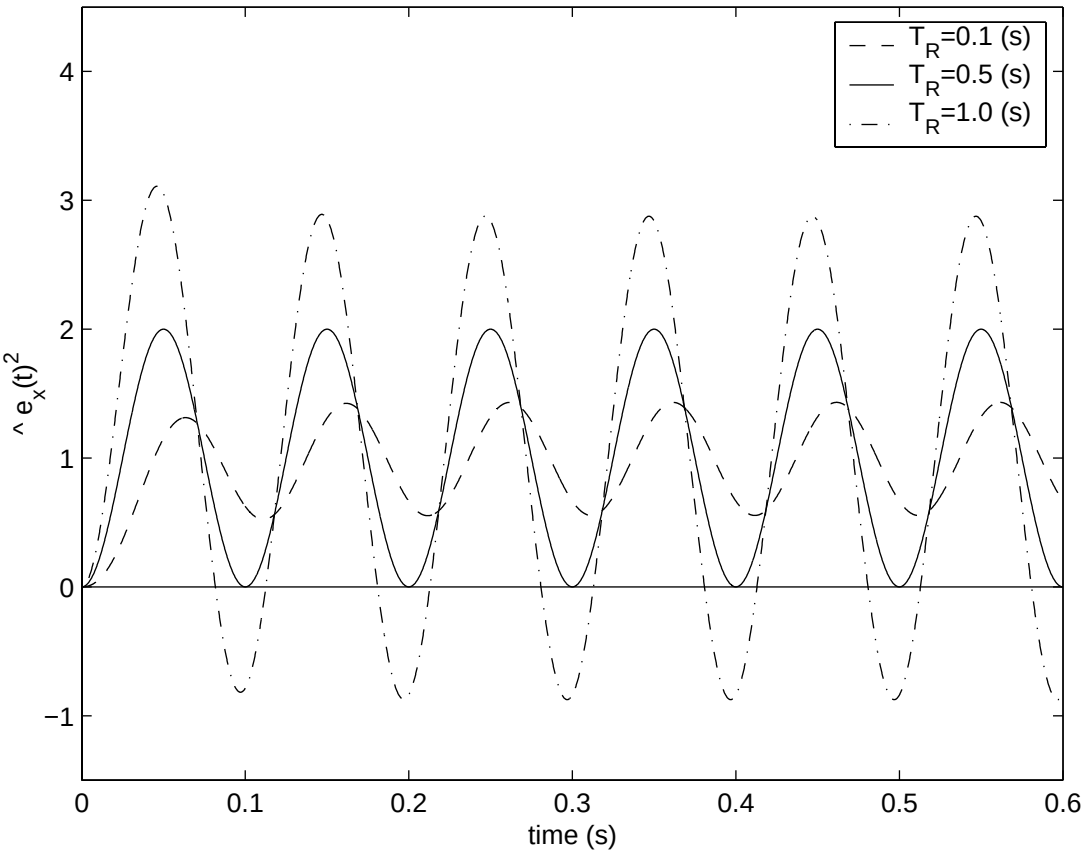

図 3.2: 図 3.1(f)T_R=0.5 のパワーエンベロープ逆フィルタ処理後のパワーエンベロープ。各線は、処理に用いた残響時間パラメータ T_R が、破線:0.1、実線 0.5、一点差線:1.0 のとき

3.1.3 パワーエンベロープ抽出法

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 $y(t)$ から残響信号の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。この手法は古川らにより有効性が示されている [10]。

この手法の原理としては、まず観測信号のキャリアが偶関数あるいは奇関数で構成されているものと仮定すれば、Hilbert 変換を利用して瞬時振幅を得ることができる。そして得られた 2 乗瞬時振幅をローパスフィルタをかけることで、観測信号のパワーエンベロープ $\hat{e}_y(t)^2$ を抽出することができる。

$$\hat{e}_y(t)^2 = \mathrm{LPF}\left[\mathrm{Hilbert(y(t))}^2\right] \tag{3.12}$$

ここでローパスフィルタのカットオフ周波数は 20 Hz とした。これは金寺ら [14],[15] によって報告された、音声知覚と音声認識における変調周波数は主に 1〜16 Hz の帯域が重要であるという結果に基づいて設定されたものである。

3.1.4 残響パラメータ推定法

パワーエンベロープ逆フィルタ処理を行うには、残響インパルス応答のパラメータである振幅項 a および、残響時間 T_R の値を決定する必要がある。古川らは抽出した残響信号パワーエンベロープから、これらの値を決定する手法を提案している。[10]、その決定法の原理を説明をする。

1. 振幅項 a の決定法

   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 を決定する。

2. 残響時間 T_R の決定法

   図 3.2 から分かるように逆フィルタ処理は、回復処理のパラメータ、残響時間 T_R の値が大きくするほど、パワーエンベロープのピークとディップ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、式 (3.13) 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で、回復処理に適した残響時間パラメータを推定できる。

$$\hat{T}_R = \max\left(\underset{0\leq T_R\leq T_{R,\max}}{\arg\min}\left\{\int_0^T \min\left(\hat{e}_{x,T_R}(t)^2, 0\right) dt\right\}\right) \tag{3.13}$$

   ここで，$\hat{e}_{x,T_R}(t)^2$ は、T_R を関数として回復されたパワーエンベロープ、$T_{R,\max}$ は T_R の上限である。

以上、パワーエンベロープ逆フィルタ処理の前処理として、残響信号のパワーエンベロープ $e_y(t)^2$ の抽出、残響パラメータ a, T_R の決定することで、室内インパルス応答を測定することなく、観測した残響信号のみからパワーエンベロープの回復処理を行うことができる。

3.1.5 帯域分割処理

上記で述べたパワーエンベロープ逆フィルタ法を音声信号に適用させる場合を考える。パワーエンベロープ逆フィルタ法は全帯域にてパワーエンベロープが共変調として処理を

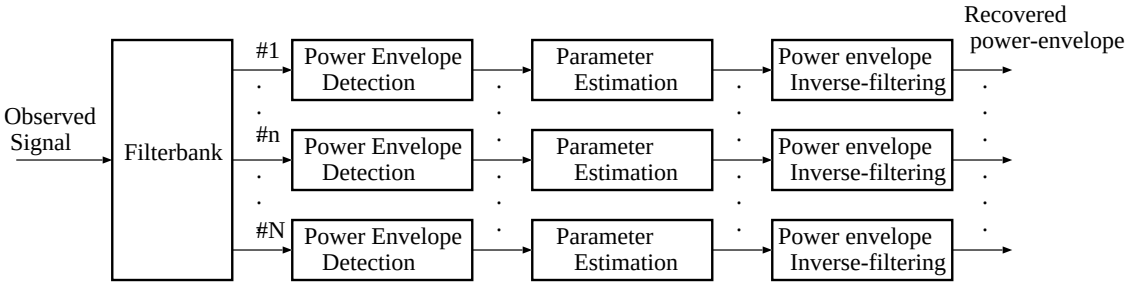

図 3.3: パワーエンベロープ回復法のブロック図

行う。この処理を共変調でない状況が多い音声信号に適用させる場合、音声信号のパワーエンベロープが共変調とみなせる帯域毎に分割を行い、そして各帯域内にてパワーエンベロープの回復処理を行なう必要がある。

そしてこのときの帯域分割処理において、適切な帯域分割幅を設定する必要がある。

以上、パワーエンベロープ回復法の各処理部について説明した。そしてパワーエンベロープ回復法のブロック図を図 3.3 に表す。

回復法の手順について説明する。まず観測した残響音声を定帯域フィルタバンクを用いて、設定された帯域分割幅毎に帯域分割処理を行う。そして各帯域毎に残響音声中から残響音声パワーエンベロープ $e_y(t)^2$ を抽出する。抽出したパワーエンベロープから、残響パラメータ a, T_R を決定する。求めた $e_y(t)^2$ および a, T_R を用いて、パワーエンベロープ回復処理を行う。

3.2　パワーエンベロープ回復法の問題点

古川らが提案しているパワーエンベロープ回復法には、以下の問題点が挙げられる。

1. 帯域分割処理において適切な帯域分割幅の決定する際、MTF 成立/不成立の検討がされていない。

2. 音声の重要な特徴を有する低帯域での回復効果が小さい。

まず、適切な帯域分割幅の検討の問題点について説明する。帯域分割幅を狭くするにつれ、帯域幅内でのパワーエンベロープが共変調とみなすことができると予測される。しかしその一方、キャリア間の無相関性である仮定が大幅に崩れる可能性がある。それに伴い式 3.9 が成り立たない、すなわち MTF 理論が適用されないことが予測される。これまでに提案されている帯域分割型パワーエンベロープ逆フィルタ法では、広林らは定 Q/定帯域フィルタバンクの構成と回復精度の関係のみ調べている [8],[9]。また古川らは音声のパワーエンベロープが共変調とみなせる帯域幅を調べている [10]。しかしいずれの方法も、狭帯域内の MTF 理論成立/不成立の議論はなされていない。

また低帯域の回復効果が小さい問題では、これまで低帯域の回復効果についての検討がなされていない。

よって、これら二つの問題点に対して、検討を行った。

3.3　適切な帯域分割幅の検討

適切な帯域分割幅を設定するとき、パワーエンベロープの共変調、MTF 理論成立/不成立の二点を考慮する必要がある。そこで本研究では、帯域分割処理における適切な帯域分割幅を検討する。その際に以下の二点がトレードオフの関係にあると予測し、この二点の調査を行う。

- 帯域分割幅内でのパワーエンベロープの共変調
- 帯域分割幅内での MTF 理論成立/不成立

3.3.1　パワーエンベロープの共変調についての調査

パワーエンベロープの共変調に関する調査方法として、古川らによって検証された方法を利用する [10]。

パワーエンベロープを帯域毎に見たとき、各帯域のパワーエンベロープ同士の相関が高いほど、共変調とみなせると考える。そして帯域幅を狭くするにつれ、帯域内のパワーエンベロープ同士の相関がどれだけ高いかを調べる。

まず音声信号を帯域分割幅 40 Hz の定帯域フィルタバンクで帯域分割を行なう。次に各チャネルのパワーエンベロープ同士の相関値を計算する。一例として、ATR 音声デー

タベースの音声データ（Mau,/aikawarazu/) について各帯域毎のパワーエンベロープ同士の相関を調べた結果を図 3.4 に表す。各図内の等高線は相関の分布を表しており、相関が 0.98, 0.95, 0.9, 0.85, 0.8 以上の範囲を、図 3.4(a),(b),(c),(d),(e) にそれぞれ示している。相関が高くなるほど分布の範囲が狭くなる。すなわち帯域幅を狭くするほど、その帯域内のパワーエンベロープ同士の相関が高いことがわかる。また低域のチャネルでは、他の帯域と比べて狭帯域で設定しないと、高い相関が得られないことがわかる。本研究では定帯域フィルタバンクを用いるため、図 3.4 の各結果から、全帯域で一定で、与えられた相関値の分布範囲内にある幅を決定する。決定方法としては、相関の分布範囲よりも狭いチャネルが全体のチャネル数の大体 9 割になるときの幅を推定することにした。この決定方法の理由は、低域のチャネルでは相関がどの値でもその分布範囲が狭いためである。図 3.4 (a),(b),(c),(d),(e) の結果から、それぞれ 90, 130, 170, 210, 250 Hz の幅が得られた。例えば図 3.4(b) では、全帯域を帯域幅分割 130 Hz に帯域分割すれば、その帯域内のパワーエンベロープ同士の相関 0.95 以上であるとみなしている。

そして音声信号 mau ,/aikawarazu/において帯域幅とその帯域内でのパワーエンベロープの相関の関係を表したものを図 3.5 に表す。横軸が全帯域に一定に分割した帯域幅、縦軸がその各帯域幅内でのパワーエンベロープ同士の相関の高さを示している。帯域幅を狭くするにつれ、帯域内のパワーエンベロープ同士の相関が高くなる傾向にあることがわかる。

以上の帯域幅とエンベロープの相関の関係の調査を、ATR 音声データベースにある 30 話者 (男性 5 名: Mau, Mtm, Mnm,Mtm, Mtt, 女性 5 名: Faf, Ffs, Fkn, Fsu, Fyn) の 3 単語 (/aikawarazu/,/sinbun/,/joudan/) の音声データを対象に同様に行った。その結果を図 3.6 に表す。各音声に対する平均データの結果を実線、他の音声と比べて狭帯域でないと相関が高くならない音声のサンプル例を点線、また比較的広い帯域幅で相関が高い場合の音声のサンプル例を隔線に示す。この結果から、どの音声も帯域幅を狭くするにつれ、帯域内のパワーエンベロープ同士の相関が高くなる傾向にあり、200 Hz から 300 Hz あたりの帯域幅で相関が 0.8 以上をもつことがわかった。

表 3.1: MTF 理論成立/不成立の調査のシミュレーション条件

入力信号	キャリア:100 種類の白色雑音 パワーエンベロープ:下記の三種類の $e_x(t)^2$
インパルス応答	残響時間 T_R=0.1, 0.3, 0.5, 1.0, 2.0 キャリア:一種類の白色雑音
フィルタバンク	定帯域フィルタバンク 帯域分割幅:10, 5, 2, 1, 0.5, 0.4, 0.2 0.1 (kHz)
評価尺度	$e_y(t)^2$ と $\hat{e}_y(t)^2$ に対する SNR, 相関値

3.3.2 MTF 理論成立/不成立についての調査

まず調査方法について説明する。MTF 理論に基づけば図 3.1(b) のパワーエンベロープ $e_x(t)^2$ が残響の影響を受けることで、図 3.1(f) のように変調度が下がり、時間方向へシフトするパワーエンベロープ $e_y(t)^2$ が得られる。

これは式 3.9 のパワーエンベロープ同士の畳み込み積分の式に $e_x(t)^2$ と $e_h(t)^2$ を代入すれば、同様の $e_y(t)^2$ が得られる。つまり MTF 理論が成立すれば観測した残響音声 $y(t)$ から抽出した $\hat{e}_y(t)^2$ は $e_y(t)^2$ と同等とみなすことができる。

そこで帯域分割幅を関数 (10, 5, 2, 1, 0.5, 0.4, 0.2, 0.1 kHz) として、各帯域内で $\hat{e}_y(t)^2$ と $e_y(t)^2$ が近似的にどこまで等しいかシミュレーションを行う。

シミュレーションの条件は表 3.1 の以下のとおりである。

各帯域幅内の $e_x(t)^2$ には以下の三種類のパワーエンベロープを用いた。

1. 正弦波信号：$e_x(t)^2 = 1 - \cos(2\pi F t)$
2. 調波複合音：$e_x(t)^2 = 1 + \frac{1}{K}\sum_{k=1}^{K} \sin(2\pi k t + \theta_k)$
3. 帯域制限されたランダム信号：$e_x(t)^2 = \mathrm{LPF}[n(t)]$

ここで、$F = 10$ Hz、$K = 20$、θ_k はランダム位相、$n(t)$ は白色雑音である。

また評価尺度の SNR は以下の式のように、S を $e_y(t)^2$、N を $e_y(t)^2$ と $\hat{e}_y(t)^2$ の差とした。

$$\mathrm{SNR(dB)} = 10\log_{10}\frac{\int_{-\infty}^{\infty}\{e_y(t)^2\}^2 dt}{\int_{-\infty}^{\infty}\{e_y(t)^2 - \hat{e}_y(t)^2\}^2 dt} \tag{3.14}$$

残響時間が 0.1, 0.3, 0.5 のときの調査結果を図 3.7, 残響時間が 1.0, 2.0 のときの調査結果を図 3.8 に示す。各帯域分割幅における各チャネルの SNR と相関の平均を表している。

この結果には、各帯域内におけるパワーエンベロープの抽出誤差が含まれているため、各分割帯域幅毎に平等に判断するのは難しいが、帯域幅を狭くするにつれ、SNR と相関値が低下していることがわかった。

またどの帯域分割幅の場合でも各チャネルに大きなばらつきは見当たらなかった。

表 3.2: 従来の回復法によるシミュレーション条件

サンプリング周波数	f_s=20000 Hz
入力音声	ATR 音声データベース (mau /aikawarazu/)
残響時間	T_R=0.5
フィルタバンク	定帯域フィルタバンク (帯域分割幅 400 Hz)

3.3.3 適切な帯域分割幅の決定

以上、帯域分割幅内におけるパワーエンベロープの共変調の調査結果と MTF 理論成立/不成立の調査をそれぞれ行った。そしてパワーエンベロープの共変調の調査結果と MTF 理論成立/不成立の調査結果を比べた図を図 3.9 に表す。点線がパワーエンベロープの共変調 (Research1)、実線が MTF 理論の成立/不成立の相関値 (Research2) である。これらの図から、どの残響時間においてもパワーエンベロープの共変調と MTF 理論の成立/不成立の二点はトレードオフの関係にある。また適切な帯域分割幅は 300 Hz から 400 Hz の範囲にあるとみなすことができる。

3.4 低帯域における回復効果についての検討

3.4.1 低帯域の回復効果が小さい原因

まず低帯域の回復効果が小さい原因について説明する。

今、表 3.2 の条件のように、ATR 音声データベース (mau /aikawarazu/) の音声信号が残響時間 0.5 秒の残響の影響を受けた残響信号を考える。この残響信号を帯域分割幅 400 Hz で従来のパワーエンベロープ回復処理を行う。評価尺度は SNR の改善度、相関の改善度とし、それぞれ以下の式で求める。

$$\text{ImprovedSNR(dB)} = 10\log_{10}\frac{\int_{-\infty}^{\infty}\{e_x(t)^2-\hat{e}_x(t)^2\}^2dt}{\int_{-\infty}^{\infty}\{e_x(t)^2-\hat{e}_y(t)^2\}^2dt} \tag{3.15}$$

$$\text{ImprovedCorrelation} = \text{Correlation}(e_x(t)^2,\hat{e}_x(t)^2) \tag{3.16}$$

$$-\text{Correlation}(e_x(t)^2,\hat{e}_y(t)^2) \tag{3.17}$$

各チャネルの改善度を図 3.10、オリジナルパワーエンベロープ、残響音声パワーエンベロープ $\hat{e}_y(t)^2$、回復処理後のパワーエンベロープ $\hat{e}_x(t)^2$ の概形を図 3.11 にそれぞれ示す。

図 3.10 からわかるように、低域 3,4 チャネル目 (800 Hz から 1600 Hz の範囲) は全く改善されていないことがわかる。一方、図 3.11 の低域 3,4 チャネル目に着目すると、0.3 秒から 0.45 秒の間に、長い無音区間が存在することがわかる。

このような場合の残響音声のパワーエンベロープは変調度 1 もしくは限りなく 1 に近いものである。回復法で用いられる残響時間推定法は、オリジナル音声の変調度が 1 と仮定

の基で逆フィルタ処理後のパワーエンベロープの変調度が 1 となる最も小さい残響時間パラメータを推定する手法であり、この推定法を長い無音区間が存在する残響音声に適用させたとき、残響時間パラメータとして 0 に近い値 (パワーエンベロープの回復処理を殆んど行わない値) を推定する。そのため、全く回復されない結果となる。実際、図 3.11 を見るとわかるように、中, 高域ではパワーエンベロープが回復されてのに比べて、低域から 3,4 チャネル目は $\hat{e}_y(t)^2$ と $\hat{e}_x(t)^2$ の位置は全く同一であり、回復が全くされてないことがわかる。

以上、低域の回復精度が悪い原因としては、音声の低帯域成分では音声間の長い無音区間が多く存在し、この場合は回復に適した残響時間パラメータを推定することができないことが挙げられる。この例だと 21,22 チャネル目も改善が全くされていないが、これらの場合も同様の理由であった。

また他の音声に対しても、低域には長い無音区間が存在する場合が多いことがわかった。

よって長い無音区間が存在する場合でも正しい残響時間パラメータを推定できる手法を提案した。

3.4.2 長い無音区間に対応した残響時間決定法の検討

無音区間に対応した残響時間推定法の原理について説明する。MTF 理論に基づけば、パワーエンベロープは残響の影響を受けることで図 3.12(a) のように時間方向へ伸びていく傾向にある。一方残響音声パワーエンベロープに広林らが提案している逆フィルタ処理を行うことで回復処理後のパワーエンベロープ $e_x(t)^2$ は時間方向とは逆方向に移動していく傾向にある。また、逆フィルタ処理は T_R パラメータの値が大きくなるにつれ、パワーエンベロープのピークを強調させる働きがある。これは図 3.2 から明らかである。ピークがより強調されれば、パワーエンベロープの概形の変化は振幅方向に対して大きくなり、一方、逆時間方向への変化は小さくなる。このトレードオフの関係は、逆フィルタ処理のパラメータ a が室内インパルス応答のパワーエンベロープの面積が 1 して決定していることに起因する。つまり、逆フィルタ処理によるパワーエンベロープの逆時間方向への移動量は T_R の値を大きくするにつれ、次第に減少していく。またパワーエンベロープのピークが過剰に強調されることで、パワーエンベロープ概形が歪む可能性が大きい。以上の見解から、パワーエンベロープの逆時間方向への移動量の変化が減少する点がパワーエンベロープが歪むことなく適切な回復処理が行える境界条件であると定義できる。

そこでこの境界条件を推定することで、回復に適切な残響時間 $\hat{T}_R$ パラメータを推定する方法を提案する。

例として、この推定法を無音区間が存在するパワーエンベロープの場合に適用させる。図 3.12 の点線で表される、後ろ部分で無音区間が長く続く変調周波数 10Hz の正弦波のパワーエンベロープ $e_x(t)^2$ と、残響時間が 0.5 秒の $e_h(t)^2$ を式 3.9 のパワーエンベロープ畳み込み積分の式で代入してで得られる $e_y(t)^2$ が図 3.12 の実線で示している。この $e_y(t)^2$ から提案した残響時間推定法で T_R を推定する。

T_R の値を増やしながら逆フィルタ処理を行い、$\hat{e}_y(t)^2$ の一番後ろのピークの地点 t_1 からパワーが 0 となる尾の先端部分 t_2 の範囲内での $e_x(t)^2$ の全面積 S を計算する。また T_R に対して S を微分することで、$e_x(t)^2$ の移動変化量 D を求める。D が急激に減少する地点を調べることで T_R を推定する。

$$S = \int_{t_1}^{t_2} \hat{e}_x(t, T_R)dt \tag{3.18}$$

$$D = -\frac{dS}{dT_R} \tag{3.19}$$

以上の処理を残響時間が $e_x(t)^2$に0.1, 0.3, 0.5, 1.0, 2.0 の $e_h(t)^2$ それぞれを畳み込んで得られる 5 種類の残響音声パワーエンベロープ $e_y(t)^2$ に対して行った。それぞれの結果を図 3.13(a),(b) に表す。図 3.13(a) から T_R の変化に対応して、S が減少、つまり回復処理後のパワーエンベロープ $e_x(t)^2$ が逆時間方向へ移動していることを意味する。また S が直線的に減少していることから、D の値は T_R に対してほぼ一定である。また T_R のある境界を越えると、S の減少が緩やかに、D の値が急激に低下する。図中の丸印は、この境界を推定した地点を示す。この推定位置から得られた $\hat{T}_R$ は、どの場合も e_h^2 の残響時間と一致した。

以上から、長い無音区間が存在する場合でも、パワーエンベロープ移動変化量の境界条件を推定することで残響時間パラメータ T_R を正確に推定できる推定法を提案した。

3.4.3 パワーエンベロープ移動変化量による残響時間推定法の評価のためのシミュレーション

提案したパワーエンベロープ移動変化量による残響時間推定法を用いたパワーエンベロープ回復法の有効性を示すためシミュレーションを行う。シミュレーション条件は、表 3.2 の条件で行い、従来の回復法と提案した回復法の結果を比較する。

従来法による結果を図 3.10,3.11 提案法による結果を図 3.14, 3.15 に示す。低域 2,3 チャンネル目の結果から、従来の推定法では長い無音区間の場合に回復法が適用されないため、相関、SNR ともに改善が得られないのがわかる。一方、提案した推定法を用いた結果では、3 チャネル目では相関が 0.05, SNR が 1 dB、4 チャネル目では相関が 0.14、SNR が 1.8 dB の改善度が得られた。図 3.15 から、提案した推定法が無音区間が長い 2,3 チャネル目で回復処理が行われているのがわかる。また 21, 22 チャネルでも同様の理由から、提案した方法で回復効果が得られた。

この結果から提案した推定法の有効性が示せ、低域での回復精度を上げることができた。

3.4.4 低域の回復効果が小さい他の原因

今回、低域の回復効果が小さい原因として、従来の回復法では音声間の無音区間が長い場合に適用できないことを挙げた。これとは別に存在する低域の悪さの原因について説明

する。

図 3.16 のように抽出した残響音声パワーエンベロープがオリジナルパワーエンベロープよりも高い周波数成分を多く含み、MTF 理論が適用できない場合が挙げられる。

先述の狭帯域による MTF 理論成立/不成立の調査では、どの帯域においてもほぼ同一の結果が得られた。しかし図 5.3 からわかるように、低域は中, 高域に比べて、パワーエンベロープの概形が複雑である。今回の狭帯域による MTF 理論成立/不成立の調査では三種類のパワーエンベロープで検討を行ったが、一つの単調な山や、周波数 10 Hz 以下の緩やかな山が長く続くパワーエンベロープを対象にして検討を行う必要がある。また多数の実音声信号を対象に対しても検討を行う必要がある。

3.5 まとめ

提案モデルのエンベロープ回復部で用いるパワーエンベロープ回復法について問題点を挙げた。一つは、適切な帯域分割幅の検討について、もう一つは低帯域での回復効果について、である。適切な帯域分割幅の検討については、狭帯域内におけるパワーエンベロープの共変調、MTF 理論成立/不成立の二点に調査を行った。それらの結果から適切な帯域分割幅は 300 から 400 Hz の範囲と決定した。

低帯域での回復効果については、回復効果が小さい原因は、音声の低帯域内で音声間の無音区間が長い場合が多く、従来の回復法だとこの場合に適用できてないことがわかった。

そして無音区間が長い場合にも適用できる回復処理法を提案し、低域の回復精度を挙げることができた。

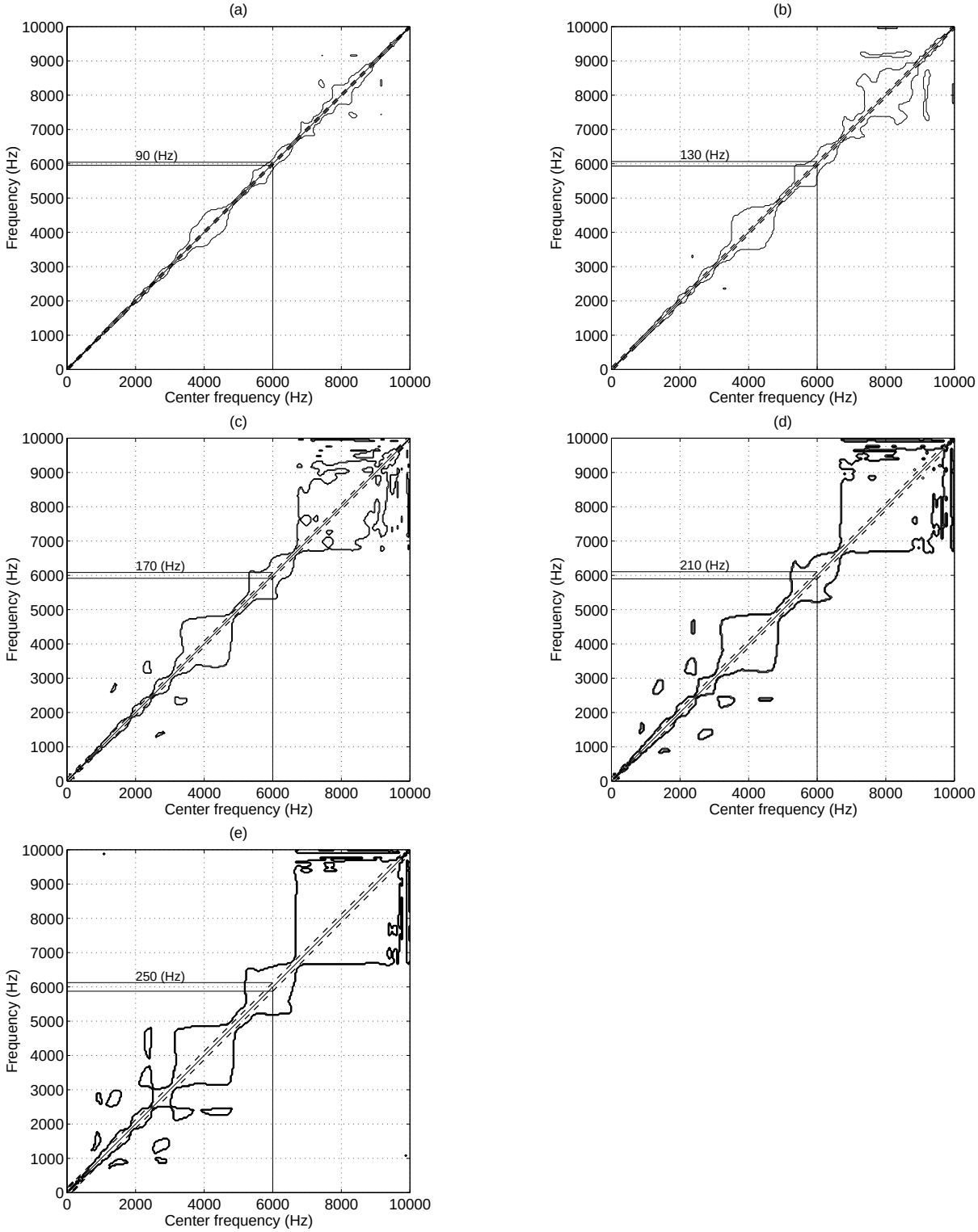

図 3.4: 音声データ (mau,/aikawarazu/) に対する、パワーエンベロープ間の相関関係。等高間隔は相関値が (a)0.98 (b)0.95 (c)0.9 (d)0.85 (e)0.8 以上の範囲。

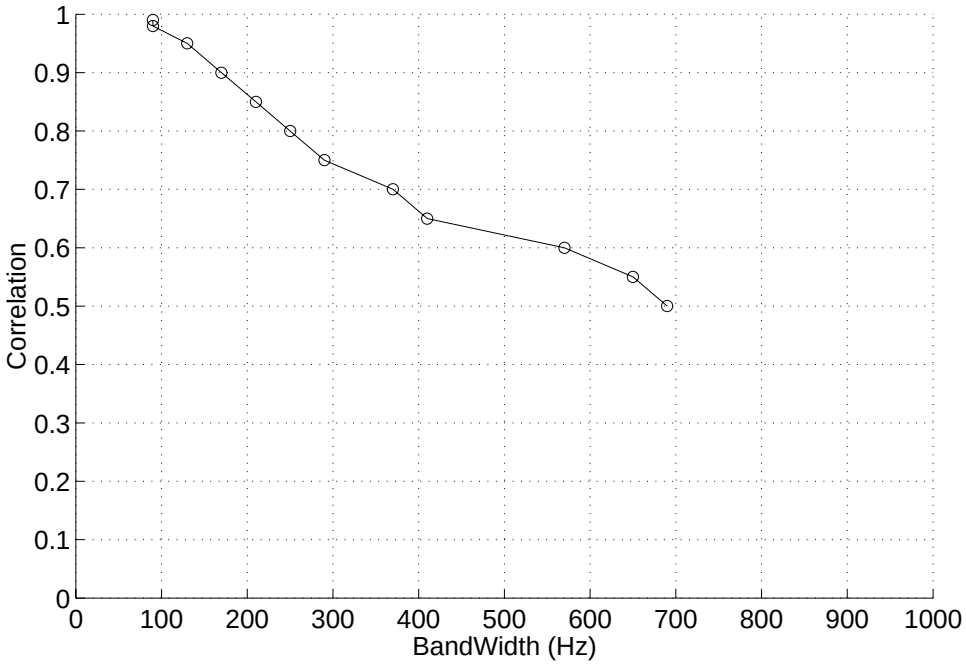

図 3.5: 音声データ (mau,/aikawarazu/) の帯域間のパワーエンベロープの相関と帯域幅の関係の調査結果

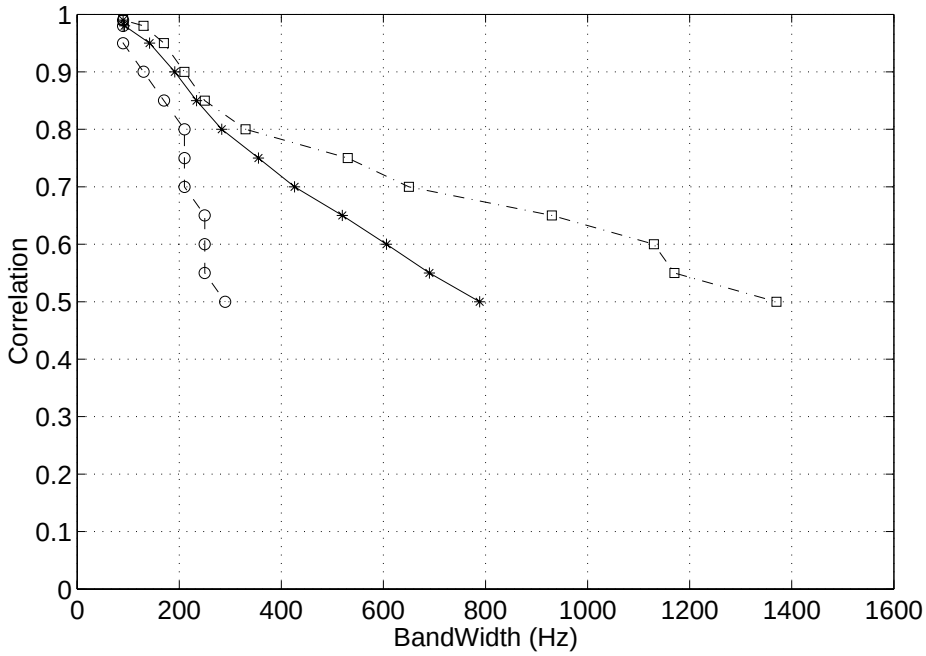

図 3.6: 30 種類の音声データの帯域間のパワーエンベロープの相関と帯域幅の関係の調査結果。実線:全音声の平均のデータ結果、破線, 点線: 標準偏差に大体位置する音声の場合の結果。

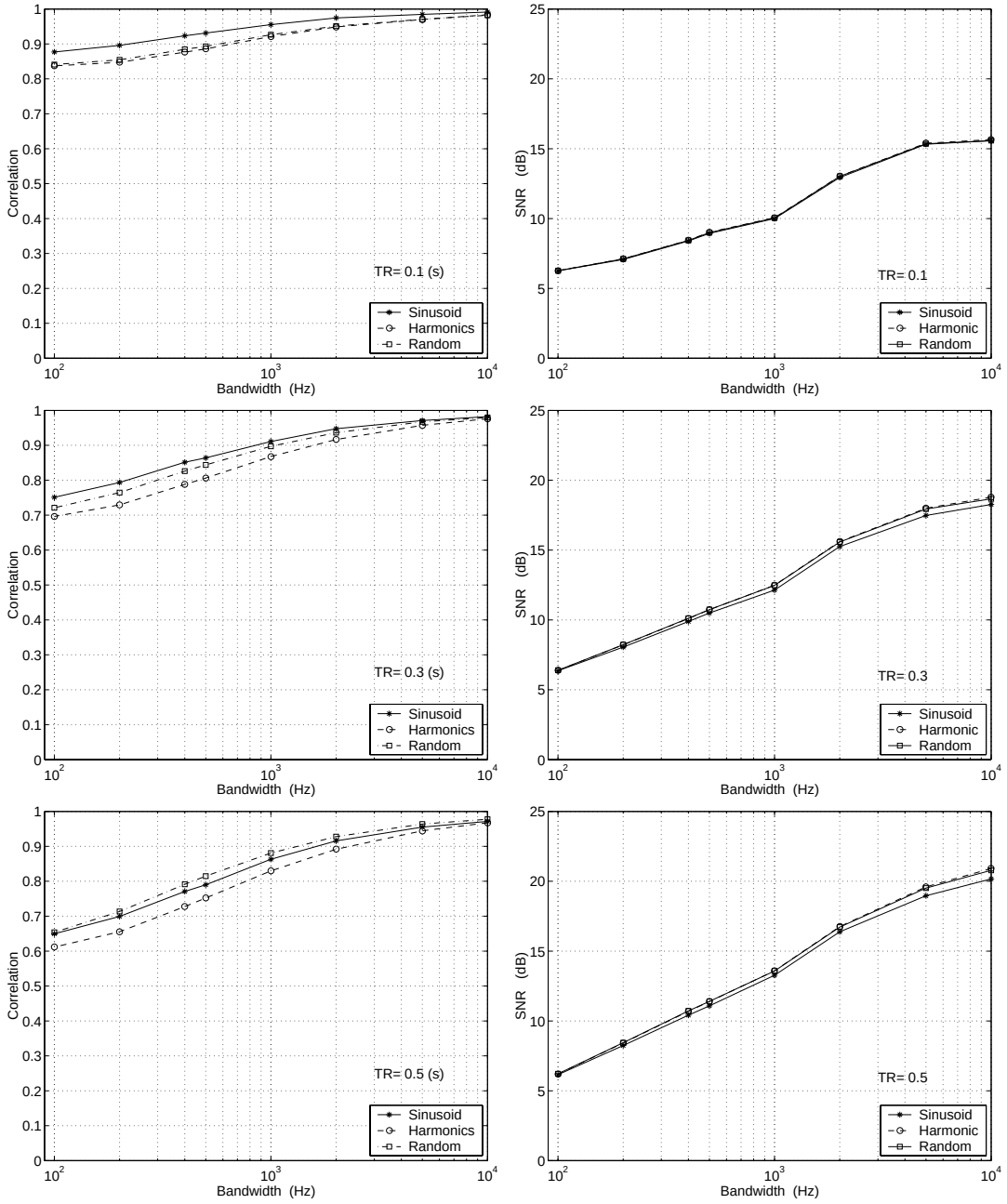

図 3.7: 上図から残響時間が 0.1, 0.3 0.5 の場合の、$e_y(t)^2$ と $\hat{e}_y(t)^2$ の右図:相関、左図:SNR の結果。

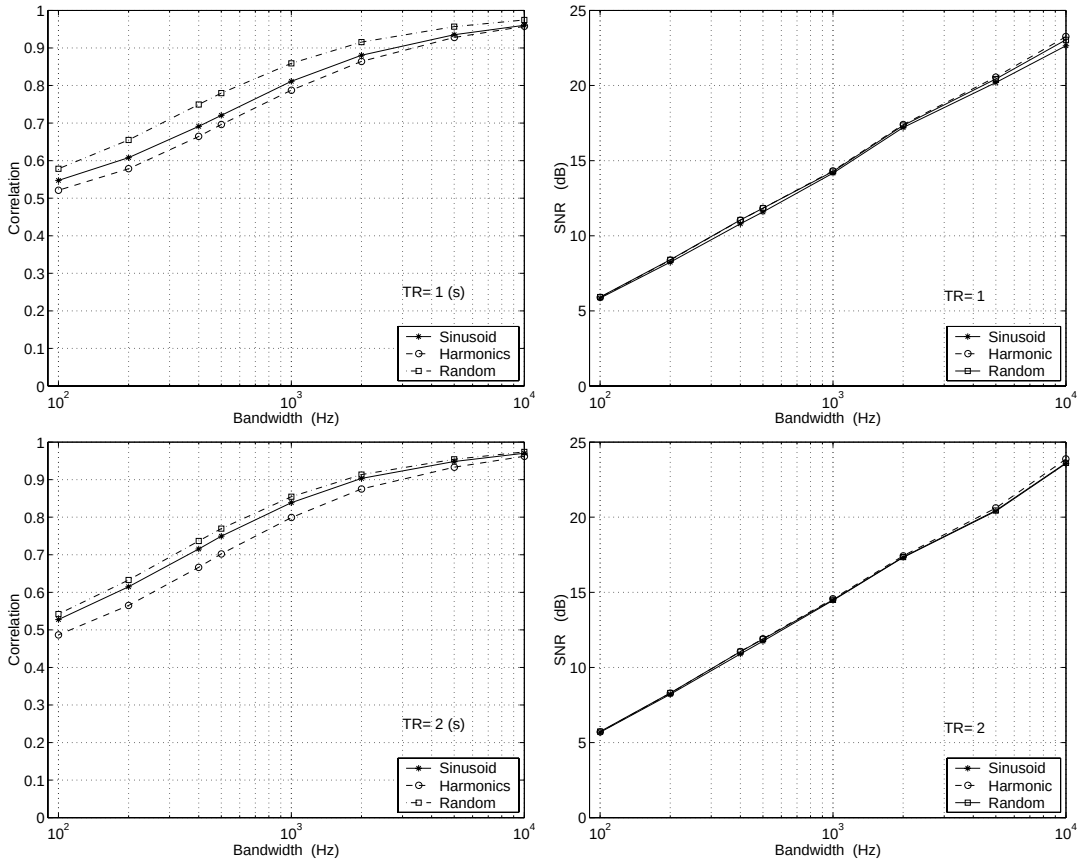

図 3.8: 上図から残響時間が 1.0, 2.0 の場合の、$e_y(t)^2$ と $\hat{e}_y(t)^2$ の右図:相関、左図:SNR の結果。

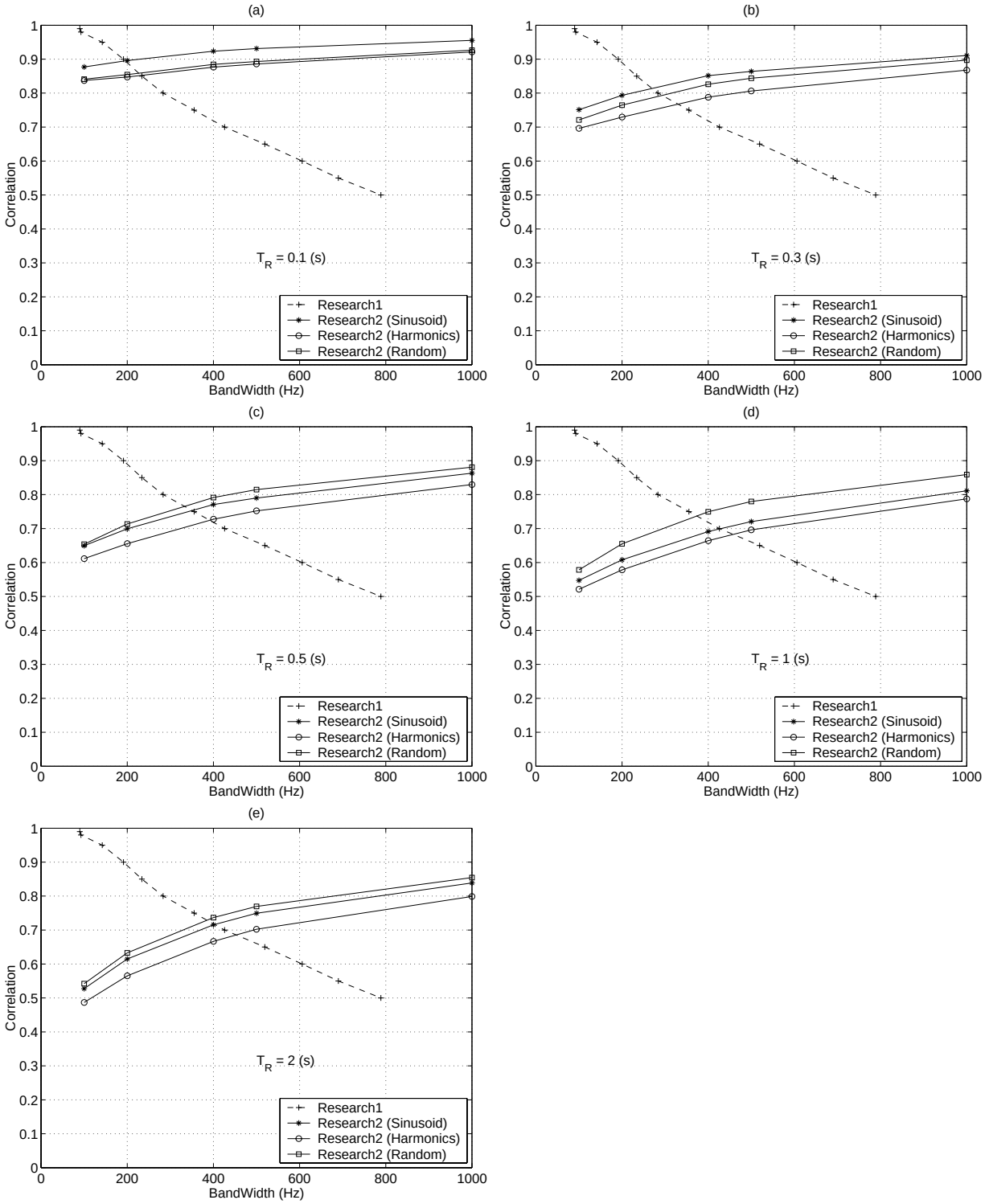

図 3.9: 帯域内のパワーエンベロープの共変調 (Research1) と残響時間が (a)0.1, (b)0.3, (c)0.5, (d)1.0, (e)2.0 のときの MTF 理論成立/不成立の相関値 (Research2) のそれぞれの結果の比較図

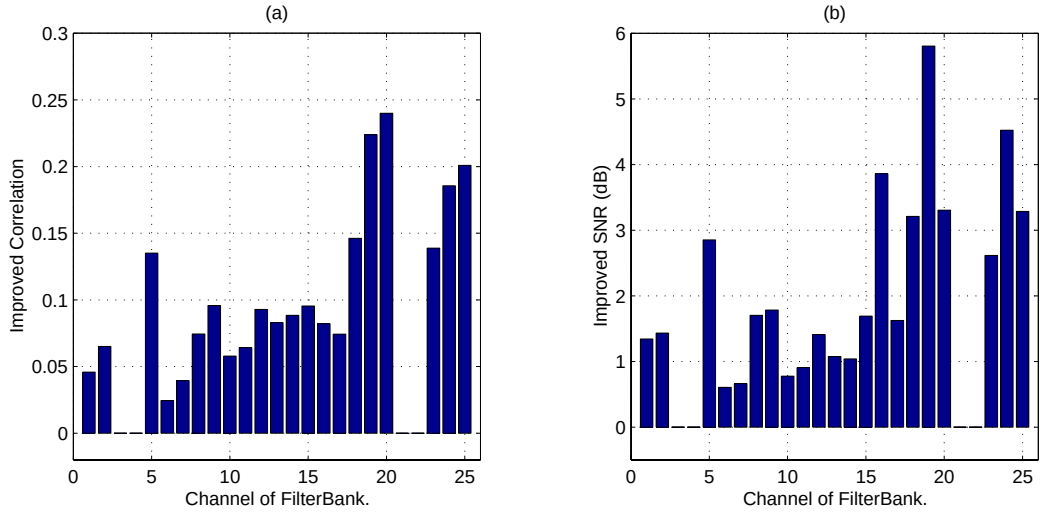

図 3.10: 従来の回復法による結果。左が低域で各チャネルの改善度を示す。(a) 相関の改善度、(b)SNR の改善度

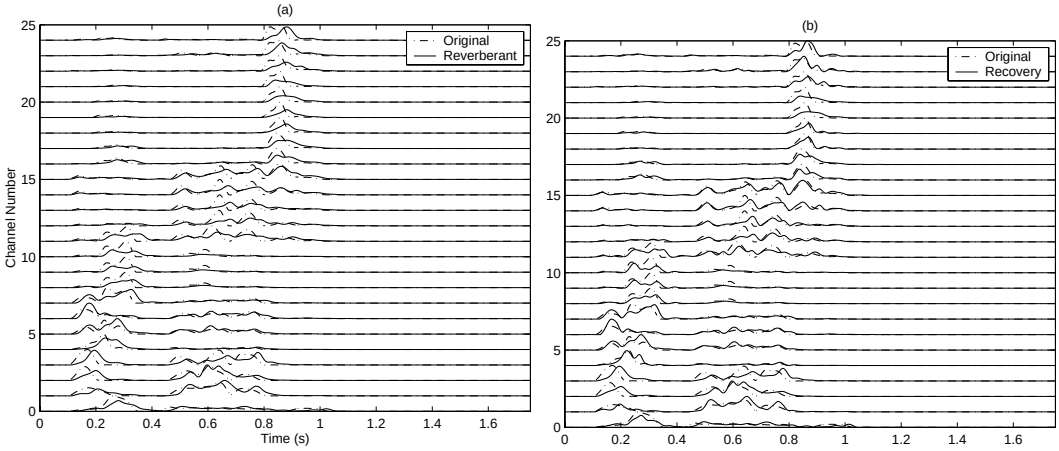

図 3.11: 帯域分割幅 400 Hz で帯域分割した各チャネルのパワーエンベロープの概形。縦軸は下が低域で各チャネルのパワーエンベロープ概形を示す。(a) 隔線:$e_x(t)^2$, 実線:$\hat{e}_y(t)^2$、(b) 隔線:$e_x(t)^2$, 実線:$\hat{e}_x(t)^2$

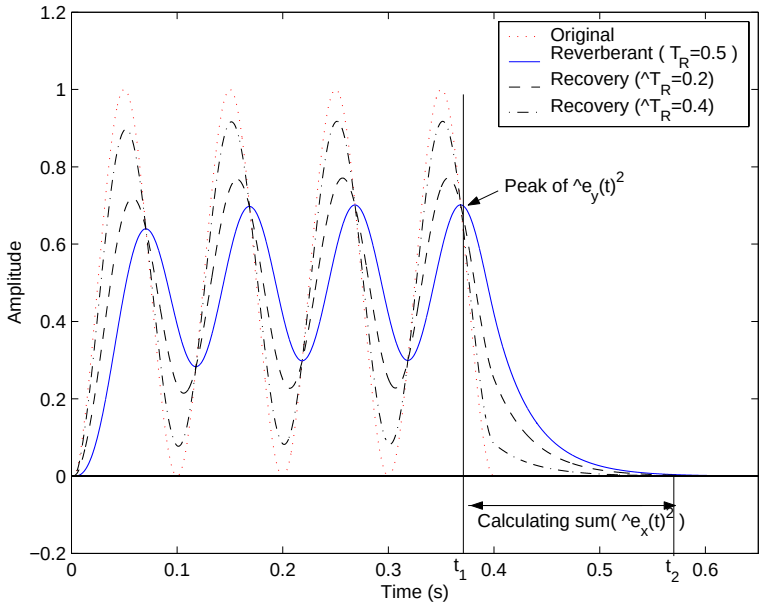

図 3.12: オリジナル (点線)、残響 (実線)、パワーエンベロープ逆フィルタ処理後 (破線)、それぞれのパワーエンベロープの関係

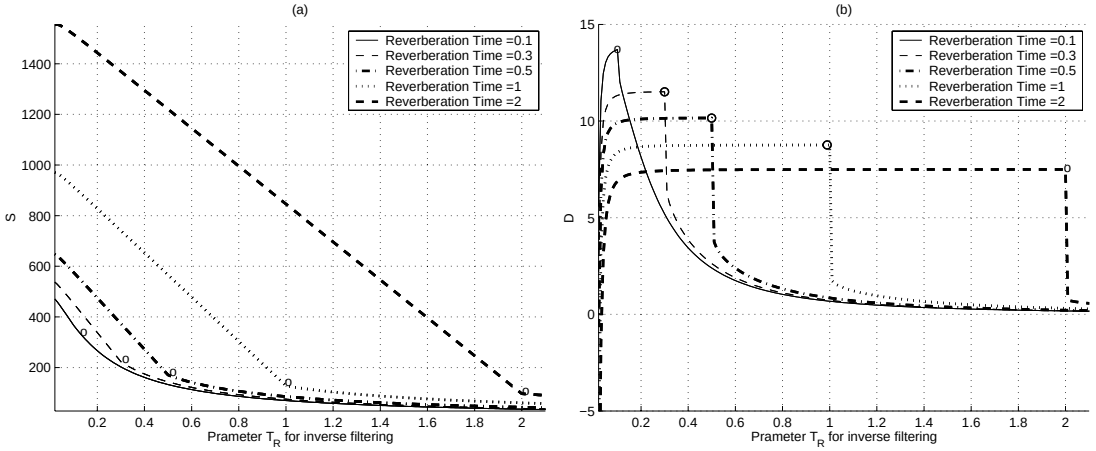

図 3.13: $\hat{T}_R$ とパワーエンベロープ移動変化量の関係。横軸は逆フィルタ処理で用いるパラメータ T_R、縦軸は (a)S、(b)D の値を示す。丸印はパワーエンベロープ移動変化量から $\hat{T}_R$ を推定した地点を示す。

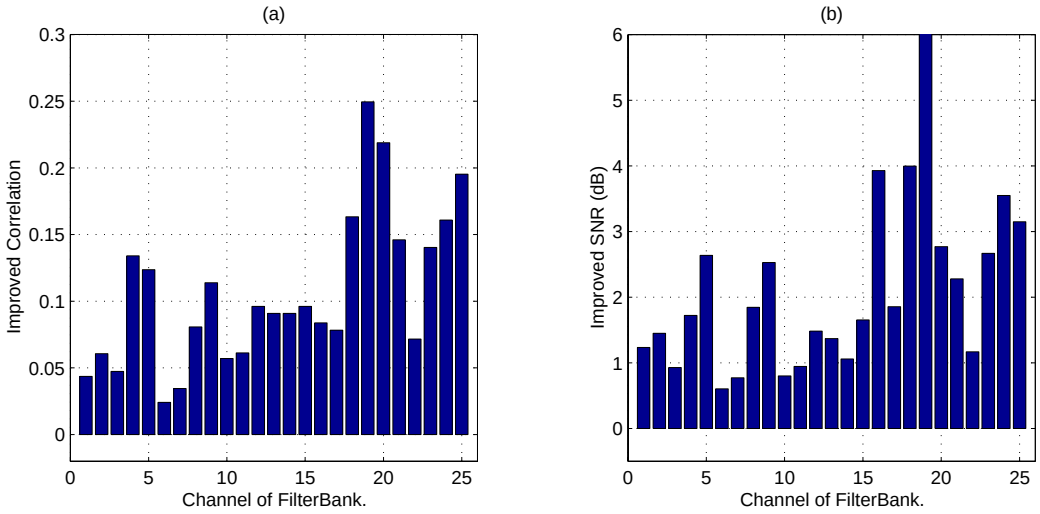

図 3.14: 提案した推定法による結果。左が低域で各チャネルの改善度を示す。(a) 相関の改善度、(b)SNR の改善度

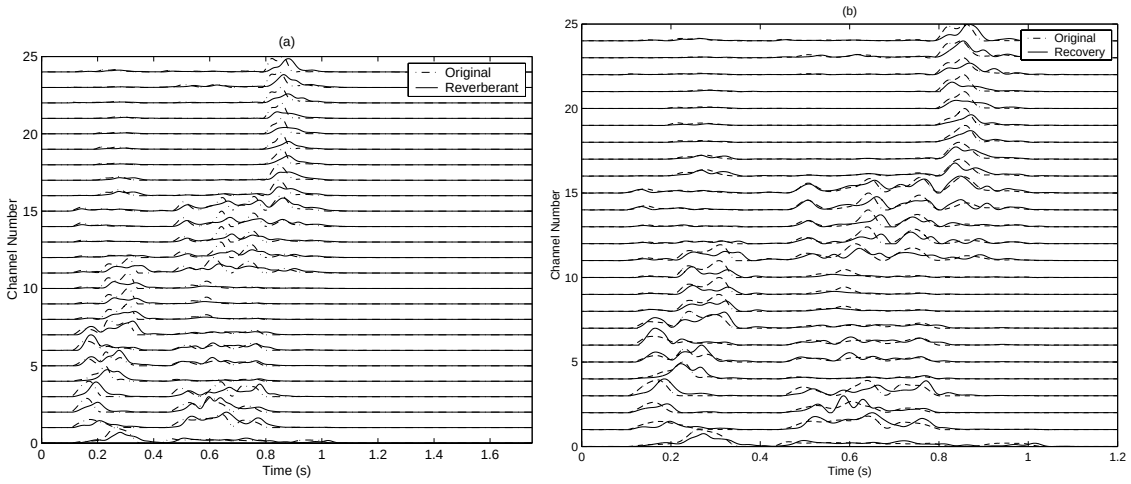

図 3.15: 提案した回復法による帯域分割幅 400 Hz で帯域分割した各チャネルのパワーエンベロープの概形。縦軸は下が低域で各チャネルのパワーエンベロープ概形を示す。(a) 隔線:$e_x(t)^2$, 実線:$\hat{e}_y(t)^2$、隔線:$e_x(t)^2$, 実線:$\hat{e}_x(t)^2$

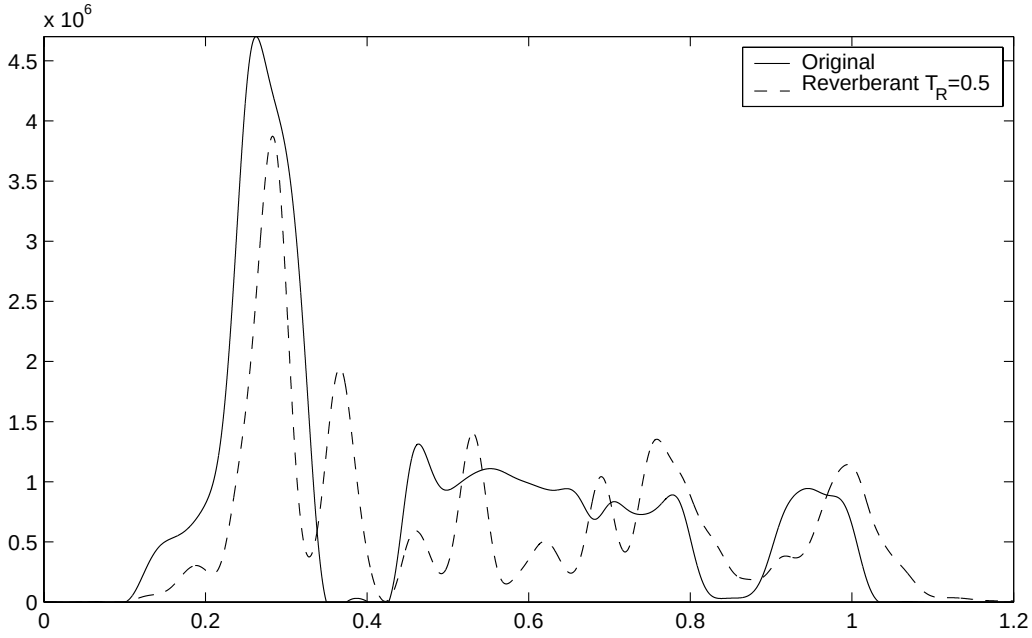

図 3.16: MTF 理論が適用できないパワーエンベロープ。実線:オリジナルパワーエンベロープ、破線:残響信号パワーエンベロープ

第4章　キャリア再合成法

この章では、本研究が提案したキャリア再合成法の原理の説明および、提案した手法の評価について論じる。

4.1　キャリア再合成法の原理の説明

本研究が提案するキャリア作成法では、F0 が存在する有声音の区間と、F0 が存在しない無声音区間で別々に処理を行う。

4.1.1　有声音区間におけるキャリア再合成法の原理

有声音区間におけるキャリアの再合成法について説明する。濱上が提案している PIFM 音源のモデル [16] を参考に、周期的な構造を持つ音声、すなわち F0 の情報を持つ音声である有音声は、F0 と同値の周波数の基本波と、F0 の倍数から成る複数の調波を足し合わせた調波複合音の波形構造である。また音声の基本波は時間的に緩やかに変化し、調波もその基本波に対応しながら緩やかに変化する。

以上の音声の性質を利用して、有声音区間におけるキャリア再合成法は、既知とした元の音声の各時間における F0 の情報 (瞬時周波数）を基に、それに対応する基本波およびその倍音である複数の調波を作成し、それらを足し合わせた調波複合音のキャリアを作成する。この作成モデルを離散表現した式を以下に表す。

$$\hat{c}_v(nT_s) = \frac{1}{K(nT_s)} \sum_{k=1}^{K(nT_s)} \hat{c}_h(k, nT_s) \tag{4.1}$$

$$\hat{c}_h(k, nT_s) = \sin(k \sum_{\tau=0}^{nT_s} \omega_0(\tau T_s) T_s) \tag{4.2}$$

ここで各変数の意味は以下のとおりである。

- $\hat{c}_v(nT_s)$: 有声音区間のキャリア
- $\hat{c}_h(k, nT_s)$: 基本波および基本波の k 倍の調波
- n: サンプル点

- $T_s = 1/f_s$: サンプリング周期
- $\omega_0(\tau T_s) = 2\pi f_0(\tau T_s)$: τ サンプル目における瞬時基本角周波数 (rad/s)
- $K(nT_s) = \frac{f_s}{f_0(nT_s)}$-1: n サンプル目における高調波の本数
- k: 高調波番号 $(1 \leq k \leq K(nT_s))$

4.1.2 無声音区間におけるキャリア再合成法の原理

無声音区間のキャリア再合成の説明をする。F0 の情報を持たない音声である無音声は、非周期的で不規則な波形構造である。

そこで無声音区間におけるキャリア再合成法では、白色ガウス過程から生起した雑音による、不規則な波形構造のキャリア $\hat{c}_u(nT_s)$ を作成する。

$$\hat{c}_u(nT_s) = \omega(nT_s) \tag{4.3}$$

$\omega(nT_s)$ は白色雑音である。

以上、有声/無声音それぞれの区間の場合のキャリア再合成法の原理について説明した。そして各区間で作成したキャリアを足し合わせることで、有声音/無声音を含んだキャリアを作成する。そのモデル図を図 4.1 に表す。そして上記で述べた手法により再合成したキャリアを帯域分割処理した後、各帯域毎にエンベロープと掛け合わせることで音声を合成することができる。

4.2 キャリア再合成法の主観的評価

今回提案した手法の主観的評価を行う。評価方法としては、まず、評価に用いる音声（入力音声）を一括処理、およびフィルタバンクに通して各帯域毎にエンベロープを抽出する。エンベロープに関しては何も処理は行わない。また入力音声から、TEMPO[11] を用いて F0 の抽出、STRAIGHT の有声音/無声音の判定を行うアルゴリズムを利用して [11]、有声/無声音区間を検出した。

そして得られた F0 および有声/無声区間の情報からキャリアを作成する。最後に作成したキャリアをフィルタバンクで各帯域毎に分割した後、各帯域でエンベロープとキャリアを掛け合わせ、合成音声を出力する。この出力音声と入力音声のスペクトログラムを比較して評価を行う。実験条件は表 4.1、入力音声の時間波形、スペクトログラム、F0 を図 4.2(a),(b),(c) にそれぞれ示す。

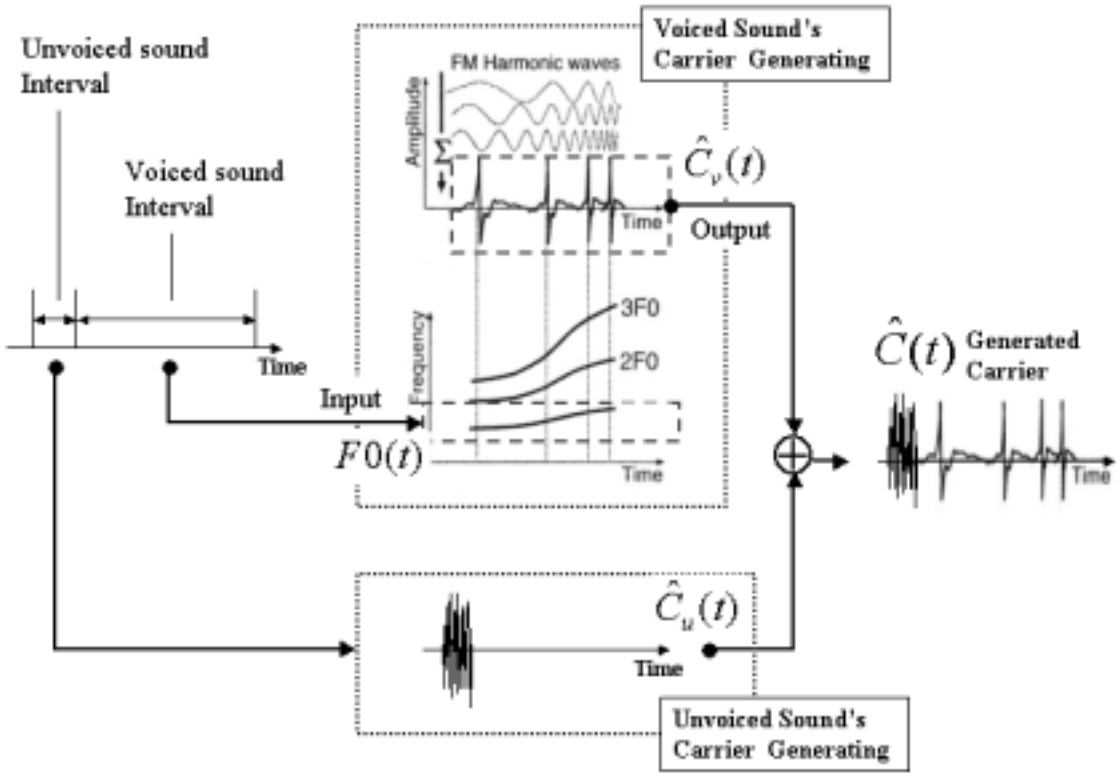

図 4.1: キャリア再合成法のモデル図

4.2.1 評価および考察

一括処理および各帯域分割幅で行った場合に得られた再合成音声の時間波形およびサウンドスペクトログラムを図 4.3,4.4,4.5, 4.6,4.7 にそれぞれ表す。どの帯域分割幅においても、再合成音声は、有声音区間 (0.3 秒から 1 秒の範囲) では入力音声と同様の周期的な構造をもつ時間波形が得られた。またサウンドスペクトログラムは、各時間における瞬時周波数に相当する周期的な濃淡模様が見られた。

無声音区間 (0.1 秒から 0.3 秒の範囲) では非周期的で不規則な時間波形が得られた。またサウンドスペクトログラムにおいても不規則な模様が見られた。一括処理の場合では、サウンドスペクトログラムは周波数領域でほぼ同一であり、スペクトル情報が存在しないことがわかる。一方、帯域分割幅を狭くするにつれ、スペクトル情報が得られた。また、それぞれの再合成音声を聴感上で評価した。その結果、帯域幅を狭くする、特に 200 ,100 Hz では入力音声に近い音声が得られた。

表 4.1: キャリア再合成法評価の実験条件

サンプリング周波数	20000 Hz
入力音声	ATR 音声データベース mau /sinbun/
フィルタバンク	定帯域フィルタバンク (帯域分割幅:1000, 400, 200, 100 Hz)

4.3 まとめ

この章では音声信号の F0 情報を基に、キャリアを再合成する処理を提案し、その原理について説明した。そして提案した処理法の有効性を示すために主観的評価実験を行った。その結果、帯域分割処理での帯域分割幅を狭帯域にするほど、スペクトル情報が得られ、それに伴い元の音声に近い再合成音声が得られた。

以上から、元の音声の F0 の情報および、有声/無声音区間が既知であれば、提案したキャリア再合成処理を用いることで、音声としての特徴を持つキャリアを作成できた。元の音声に近い再合成音声が得られることがわかった。

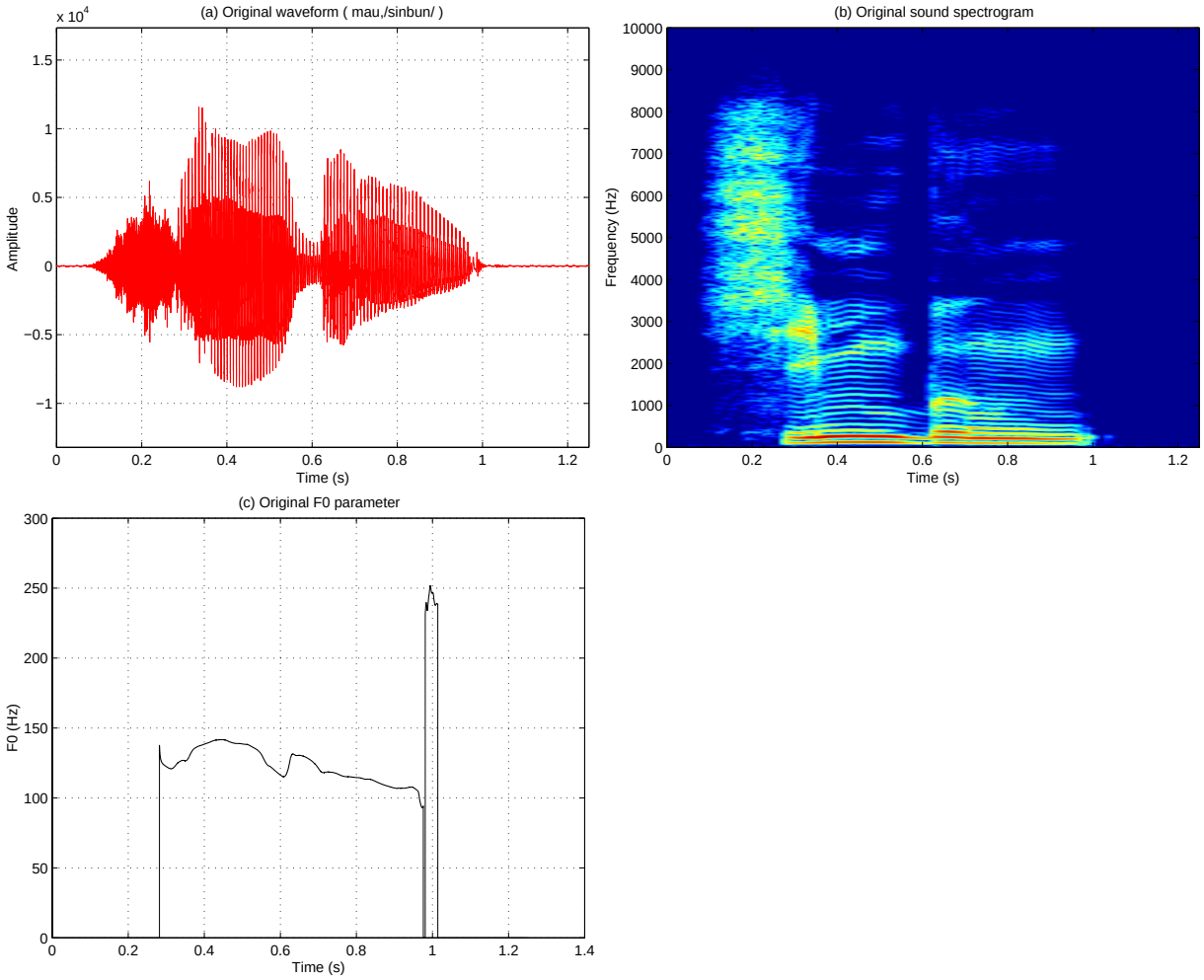

図 4.2: 入力音声 (mau /sinbun/)。(a) 時間波形、(b) サウンドスペクトログラム、(c)F0

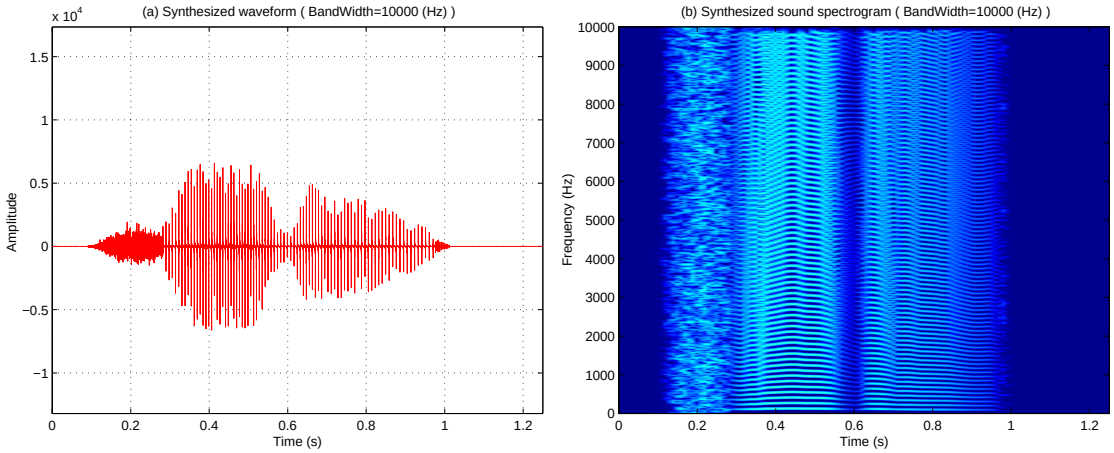

図 4.3: 一括処理の場合の再合成音声。(a) 時間波形、(b) サウンドスペクトログラム

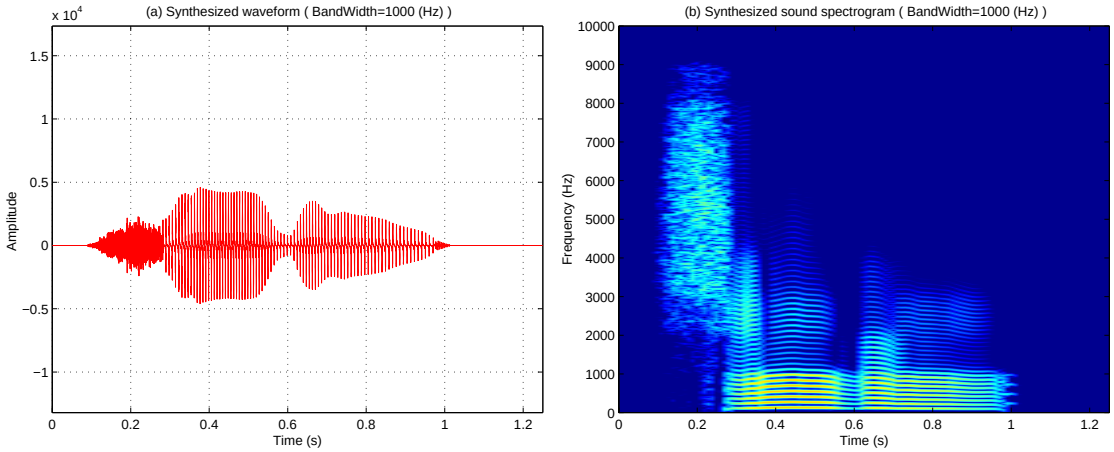

図 4.4: 帯域分割幅 1000 Hz の場合の再合成音声。(a) 時間波形、(b) サウンドスペクトログラム

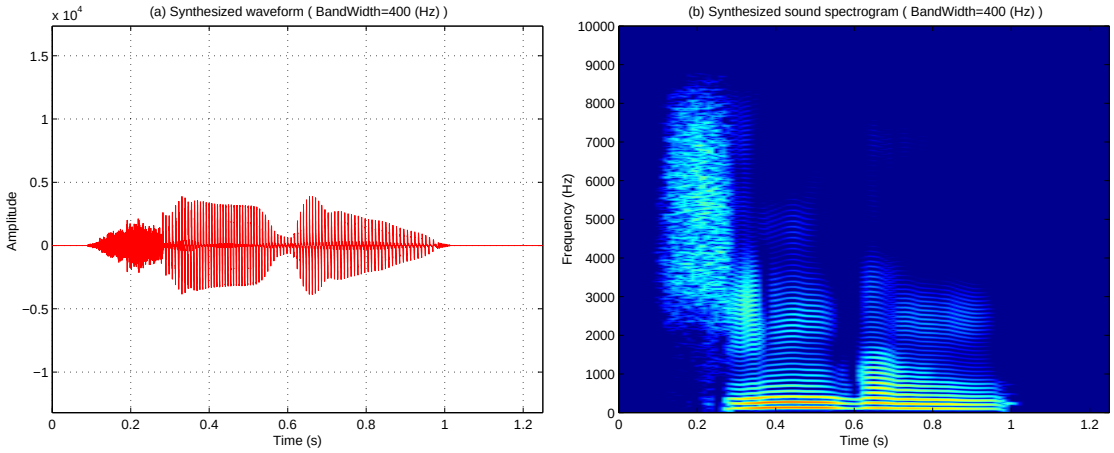

図 4.5: 帯域分割幅 400 Hz の場合の再合成音声。(a) 時間波形、(b) サウンドスペクトログラム

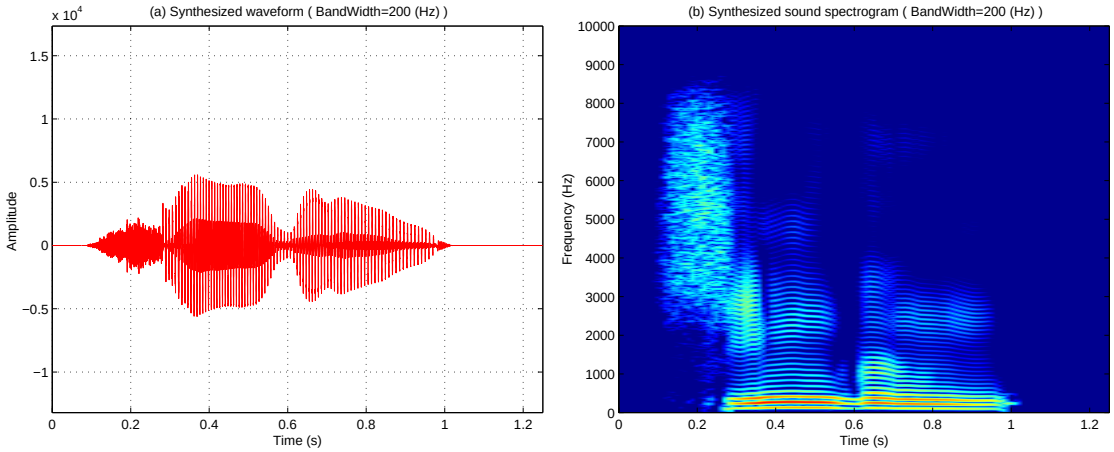

図 4.6: 帯域分割幅 200 Hz の場合の再合成音声。(a) 時間波形、(b) サウンドスペクトログラム

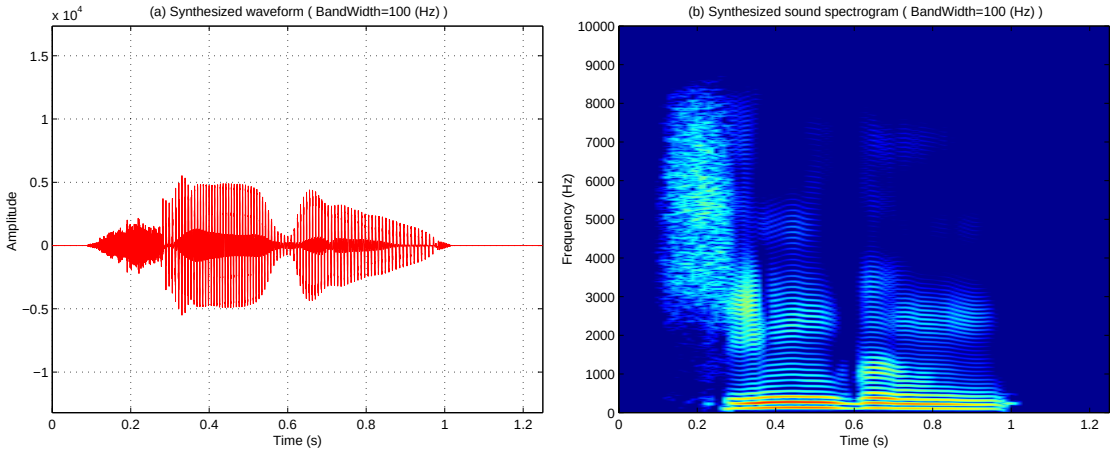

図 4.7: 帯域分割幅 100 Hz の場合の再合成音声。(a) 時間波形、(b) サウンドスペクトログラム

第5章　提案モデルの評価のためのシミュレーション

提案した提案法の評価のため、シミュレーションを行う。今回はパワーエンベロープ回復部の評価に着目する。また従来の各帯域の時間方向の SNR と相関の評価に加え、音声の特徴としての評価を考え、周波数スペクトル方向の評価も行う。

5.1　提案モデル評価のシミュレーション条件

シミュレーション条件を表 5.1 に示す。フィルタバンクの帯域分割幅は前々章の帯域分割幅の決定によるものである。

パワーエンベロープ回復部の評価尺度として、SNR の改善度および相関の改善度を用いる。また回復したパワーエンベロープと再合成したキャリアから得られた音声信号 $\hat{x}(t)$ が音声の周波数スペクトルとしての評価を行うため、オリジナル、残響、回復処理後のそれぞれのパワーエンベロープを、本研究で提案したキャリア再合成処理で作成したキャリアに掛けた 3 つの音声信号 (Original, Reverberant, Dereverberant) を作成し、Original と Reverberant の対数スペクトル距離 (LSD)、Original と Dereverberant の LSD をそれぞれ求め、評価する。尚、Original の無音区間はキャリアを作成することができないので、その区間は評価の対象外とする。

$$\mathrm{LSD(dB)} = \sqrt{\frac{1}{W}\sum_{\omega}^{W}(20\log_{10}|S(\omega)| - 20\log_{10}|\hat{S}(\omega)|)^2} \tag{5.1}$$

$$\tag{5.2}$$

$|S(\omega)|$ は Original の振幅スペクトル、$|\hat{S}(\omega)|$ は Reverberant, Dereverberant の振幅スペクトルそれぞれの場合で LSD を測定する。W は対象周波数の上限を表す。

表 5.1: シミュレーション条件

サンプリング周波数	f_s=20000 Hz
入力音声	ATR 音声データベース (mau /sinbun/)
残響時間	T_R=0.5
フィルタバンク	定帯域フィルタバンク (帯域分割幅 400 Hz)

5.2 シミュレーション結果および考察

5.2.1 相関、SNR の改善度の評価

シミュレーション結果を示す。パワーエンベロープの SNR, 相関の改善度を図 5.1、LSD による結果を図 5.2、Original と Reverberate、Original と Dereverberant の各チャネルのパワーエンベロープの概形を表したものを図 5.3(a)、(b) にそれぞれ示す。

図 5.1 から、低域のチャネルでは、1 チャネル目を除いては低帯域のチャネルに相関が 0.1、SNR が 2 dB の大きな改善が得られた。低域 1 チャネルの回復効果が得られないのは、図 5.3 の低域 1 チャネルのパワーエンベロープの概形からわかるように、残響音声のパワーエンベロープが MTF に適用できないためである。また中、高域では回復効果が得られた。

5.2.2 LSD による評価

また LSD による結果では、0.2 秒から 0.4 秒あたりの区間では平均して 1 dB の改善が得られた。しかし、0.6 から 0.8 秒では回復効果が小さい。その原因としては、2 から 4 チャネル目の 0.7 秒あたりを見ると、パワーエンベロープの尾の部分に対する回復がまだ不十分であるからと考えられる。

5.2.3 聴感上による評価

また Original、Reverberant、Dereverberant の 3 つの音声信号を比較して聴くことで、パワーエンベロープ回復処理前後による聴感上の評価を行った。Original と Reverberant と比較したところ、Reverberant では特に調音部分と音声の尾の部分に歪みが生じて聴こえた。これは残響の影響によりパワーエンベロープが歪んでいるため、音声の明瞭度が低下したと考えられる。調音部分と残響の尾の部分に特に歪みが生じるのは残響の影響を受けたパワーエンベロープの尾の部分が直後のパワーエンベロープの山をマスキングしたためと考えられる。

また Reverberant には残響感は感じられなかった。この原因は作成したキャリアを用い

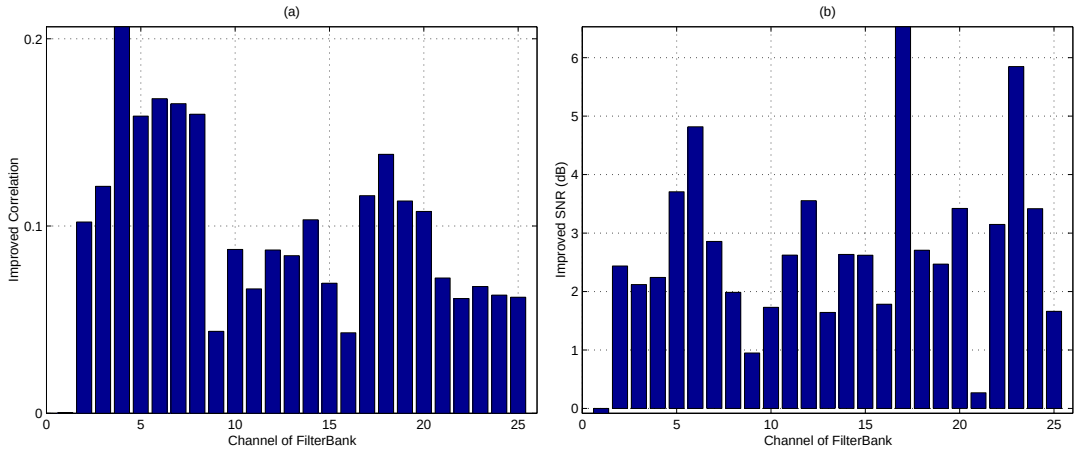

図 5.1: (a) 相関の改善度、(b)SNR の改善度

たためである。これは残響感を感じる情報は主にキャリアに含まれ、また残響音声中から F0 が抽出できれば、本研究で提案したキャリア再合成法を用いることで、残響感のない音声信号を合成できることを意味する。

また Reverberant と Dereverberant を比較したところ、大きな違いはみられなかった。これは、各帯域の音声信号の尾の部分での回復が不十分であることが原因だと考えられる。調音部分に相当する図 5.2(a),(b) の低域の 5 から 8 チャネル目の 0.4 秒、1 チャネル目の 0.6 秒、2 から 4 チャネル目の 0.7 秒の区間に着目すると、パワーエンベロープの尾の部分に対する回復がまだ不十分であるのがわかった。

5.3 まとめ

提案モデルのエンベロープの回復処理部の評価のためシミュレーションを行った。時間方向による評価では低域 1 チャネル目を除いては回復効果が得られた。また周波数スペクトル方向でも回復効果が得られた。しかしパワーエンベロープの尾の部分が十分に回復されない区間では、周波数スペクトルの回復効果が小さいことがわかった。また聴感上の評価では大きな改善による大きな差は感じられかった。

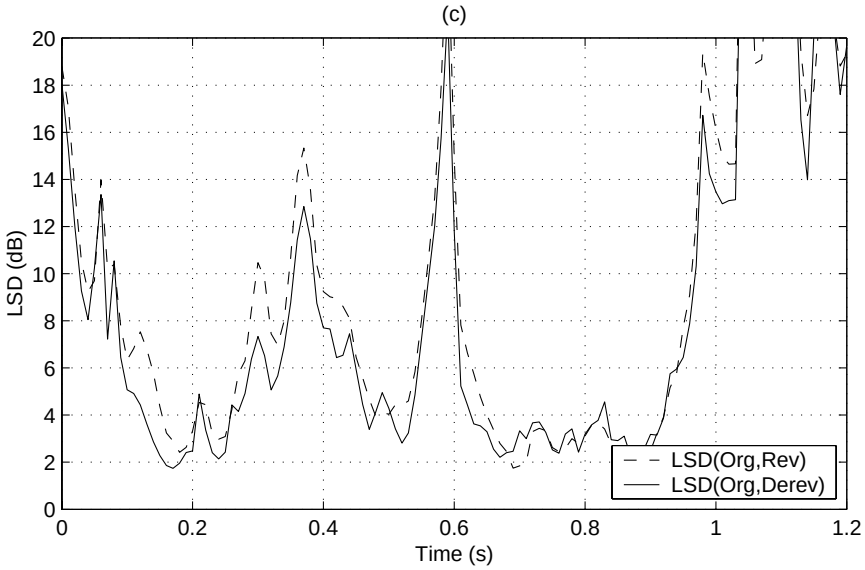

図 5.2: LSD の改善度。実線:Original と Reverberant の LSD、破線:Original と Dereverberant の LSD

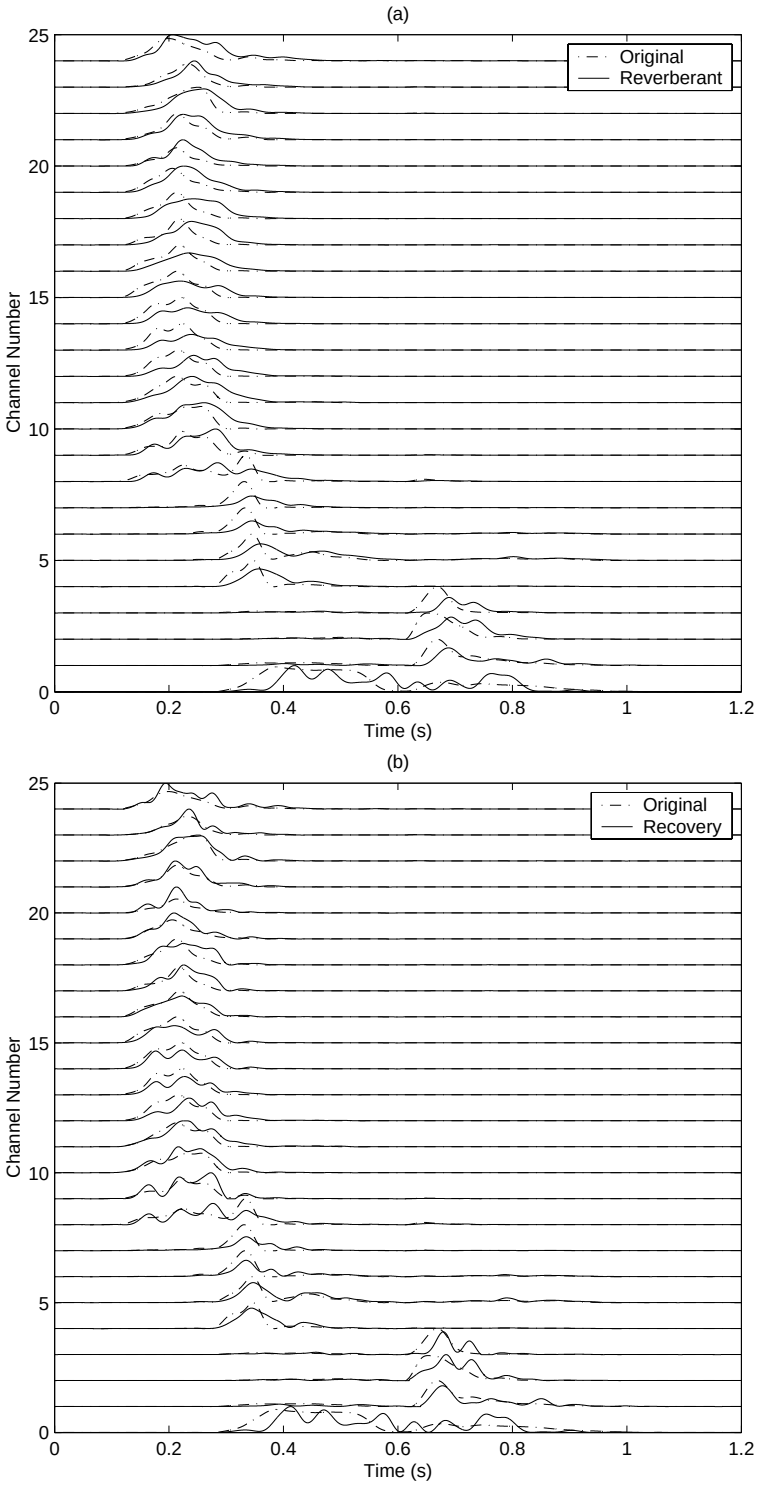

図 5.3: (a) 各チャネルのパワーエンベロープの概形 (original,Reverberant) (b) 各チャネルのパワーエンベロープの概形 (Original,Dereverberant)

第6章　まとめ

6.1　本研究で明らかにしたこと

本研究では、観測した残響音声の情報のみから残響音声の回復処理を行う残響回復処理のモデルを提案した。提案モデルのブロック図を図??に示す。エンベロープ回復部では、適切な帯域分割幅の検討と低帯域での回復精度の二つの問題点に取り組んだ。適切な帯域分割幅の検討については、狭帯域内におけるパワーエンベロープの共変調、MTF 理論成立/ 不成立の二点に着目し、適切な帯域分割幅は 300 から 400 Hz の範囲と決定した。低帯域での回復精度の不十分の問題については、音声間の無音区間が長いため、十分な回復処理が行われていないことを明らかにした。そして無音区間が長い場合にも適用した回復処理法を提案し、低域の回復精度を上げることができた。

キャリア再合成部では、有声音/無声音のそれぞれの区間でキャリアを作成する手法を提案し、残響音声中から F0 が推定できれば、音声が得られることを示した。

提案モデルの評価のためのシミュレーションを行い、各区間において周波数スペクトルとしての回復効果が見られた。よって提案モデルの有効性が示された。

6.2　本研究における課題

本研究で残った課題について説明する。

6.2.1　エンベロープ回復部での課題

- 各帯域毎に適切な帯域幅を適応的に決定する手法の提案

  音声のパワーエンベロープの共変調とみなせる帯域幅は各帯域によって異なる。これは適切な帯域幅は各音声、各周波数毎に異なることを意味している。今回は定帯域フィルタバンクを用いていたが、今後は各帯域毎に適切な帯域幅を適応的に行う帯域分割処理法を構築する必要がある。

- 時間方向による分割処理

音声は、単音よりも連続音で発話される場合が多い。全区間で同じ度合いの回復処理を行うパワーエンベロープ逆フィルタ処理をこの連続音声に対して適用させるために、時間分割処理を適応的に行う処理を構築する必要がある。

- 低域の回復精度の向上

音声の特徴量を多く含む低域の回復精度を上げる必要がある。その原因の一つとして、低帯域では、抽出した残響音声パワーエンベロープが MTF 理論を適用できない場合が多いことが挙げられる。何故、低帯域にこの場合が多く発生するか検討する必要がある。例えば狭帯域における MTF 理論成立/不成立の調査では、一つの単調な山や、10 Hz 以下の緩やかな山が長く続くパワーエンベロープを対象にして検討を行う必要がある。また多数の実音声信号を対象に検討を行う必要がある。

6.2.2 キャリア再合成処理部での課題

- 残響音声中からの F0 の抽出

キャリア再合成処理を行うには、音声の F0 の情報は不可欠である。残響音声に対しても頑健かつ精度の高い F0 抽出法を提案する必要がある。

- 作成したキャリアの各調波成分の初期位相の制御

今回作成したキャリア作成法では初期位相が全て同一であり、その結果、キャリアがパルス状の波形となっている。今後より自然性の高い音声を合成するために、STRAIGHT の SPIKES などを参考に、各調波成分の初期位相を分散させる処理を提案する。

- 有声音/無音区間の検出

残響音声から有声音/無声音の各区間を検出する必要がある。また、無声音区間と有声音区間へ遷移する区間で、どれだけの割合で調波複合音および白色雑音を含ませるか、を適用的に判定する処理を、STRAIGHT などを参考に提案する。

6.2.3 その他の課題

- フィルタバンクの構成の検討

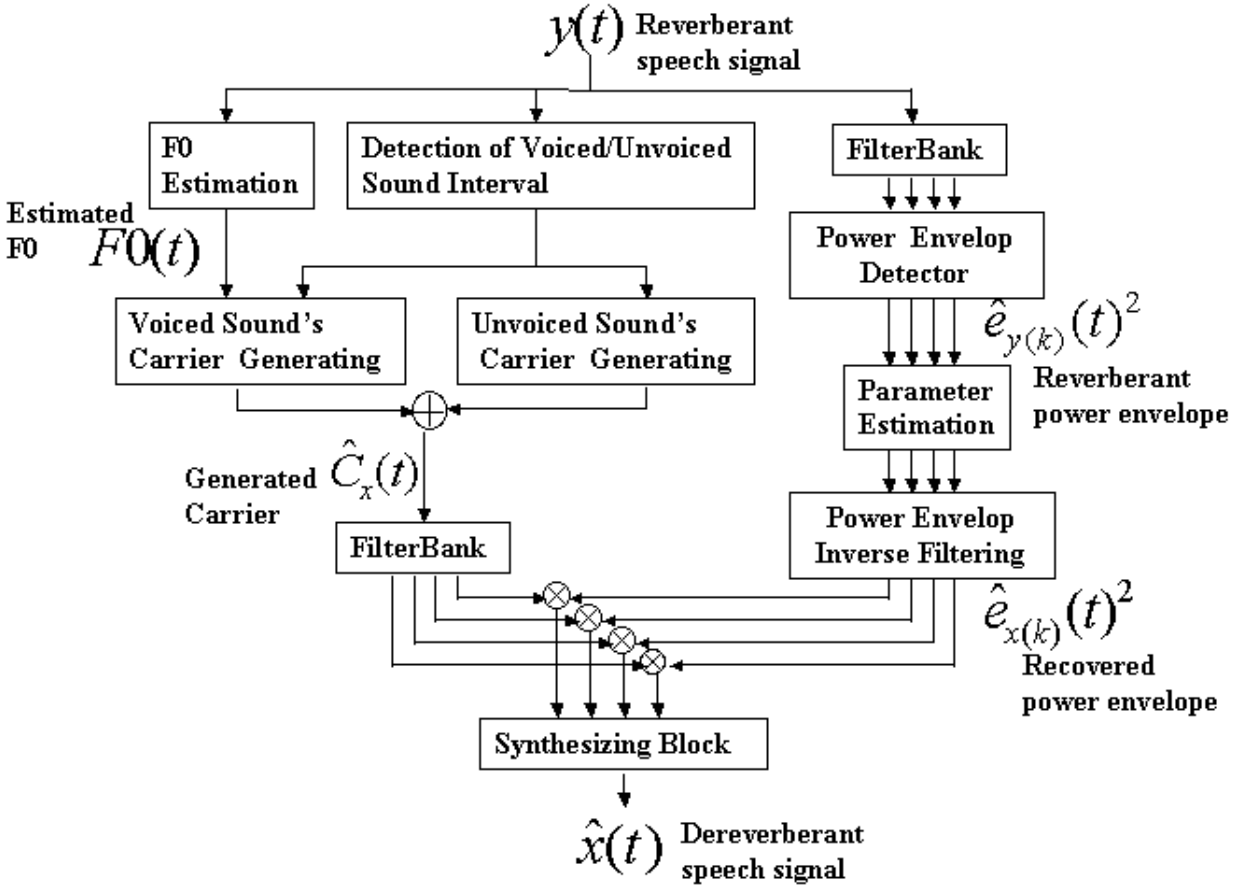

図 6.1: 提案したモデルのブロック図

本研究ではエンベロープ回復部の適切な帯域分割幅の検討を行った結果、300 から 400 Hz の帯域分割幅が適切とみなした。しかしより自然性の高い合成音声を得るにはこの帯域分割幅よりも更に狭くする必要がある。この対応として、フィルタバンクの構成の検討を提案する。本研究で用いたフィルタバンクは、隣りのチャネルのフィルタの帯域と重複しないように構成する。各フィルタを帯域分割幅よりも狭い間隔でずらしながら構成することで、音声の振幅スペクトルの情報がより多く得られると考えられる。

謝辞

本研究を進めるにあたり、多大な助言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助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鵜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博士後期課程の伊藤一仁氏、石本祐一氏、西本博則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心より感謝致します。

参考文献

[1] M.Tobita, N.Sugamura, and R.Nakatsu. "Improvement methods for effects of acoustic transmission characteristics upon word recognition performance(in Japanese)". IEICE Trans. , Vol. J73-DII, No. 6, pp. 781-787, 1990.

[2] H.Wang and F.Itakura. "An Implementation of Multi-microphone Dereverberation Approach as a Preprocessor to the Word Recognition System". J.Acoust. Soc. Japan, Vol. 13, No. 5, pp. 285-293, 1992.

[3] Neely S. T. and Allen J. B., "Invertibility of a room impulse response, "J. Acoust. Soc. Am. Vol. 66, No. 1, July 1979.

[4] Miyoshi, M. and Kaneda, Y., "Inverse filtering of room acoustics," IEEE Trans. ASSP, Vol. 36, No. 2, pp. 145-152, Feb 1988.

[5] Schroeder, M.R., "Modulation Transfer Functions:Definition and Measurement", Acoustics, Vol. 49, pp.179-182, 1981.

[6] Houtgast, T., Steenken, H. J. M., and Plomp, R., "Predicting speech intelligibility in room acoustics," Acoustica, Vol. 46, pp. 60-72, 1980.

[7] Houtgast,T.,Steenken,H.J.M., "A review of the MTF concept in room acoustic and its use for extimating speech intelligibility in audioria," J.Acoust.Soc.Am Vol.77, No.3, March 1985.

[8] 広林, 野村, 小池, 東山 "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," 信学論 A, Vol. J81-A, No.10, pp. 1323-1330, 1998.

[9] 広林, 山淵, "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," 信学論 A, Vol. J83-A, No. 8, pp. 1029-1033, 2000.

[10] 古川, 鵜木, 赤木 "MTF に基づいた残響音声パワーエンベロープの回復方法, " 信学技報, SP2002-15, pp. 49-54, 2002.

[11] 河原 " 高品質音声分析変換合成法 STRAIGHT," , ATR 人間情報通信研究所, 和歌山大学, 平成 13 年 1 月 14 日

[12] 中谷, 入野 ” 占有度を用いた耐雑音性の高い基本周波数推定法,” 信学技報.

[13] 石本 , 石塚 , 相川 , 赤木 ” 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” , 信学技法, SP2002-53, pp. 13-18, 2002.

[14] Kanedera, N. *et al.*, “On the importance of various modulation frequency for speech recognition,” Proc. EuroSpeech, pp. 1079–1082, Rhodes, Greece, Sept. 1997.

[15] 金寺, 荒井, 船田, “変調スペクトルの重要な成分のみを選択的に用いた雑音に強い音声認識,” 信学論 D-II Vol. J84-D-II, No. 7, pp. 1261-1269, 2001.

[16] 濱上 “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.

本研究に関する研究業績

- 酒田, 鵜木, 古川, 赤木 ” 帯域分割型残響音声パワーエンベロープ回復法の提案と帯域分割幅の検討”, 音講論集, pp. 507-508, (2002.9).
- 酒田, 鵜木, 赤木, ”MTF に基づいた残響音声の回復法の検討”