雑音重畳音声からの
窓関数の特性を用いた
音声信号スペクトルの推定

電気通信大学大学院 情報理工学研究科
総合情報学専攻 吉田研究室所属

学籍番号 1130042
高崎 雅也

主任指導教員　　西野 哲朗　教授
指導教員　　吉田 利信　教授

平成 26 年 1 月 30 日

概要

音声を収録する場合、周囲が騒がしいと目的の音声以外に余計な雑音が入ってきてしまう。こういった場合に雑音低減の技術が用いられる。複数のマイクや指向性のマイクによる雑音低減の方法は実用化されている。しかし、単一マイクでサンプルされた信号の雑音低減は難しい。本研究の目的は単一マイクで収録した雑音が混じった観測信号スペクトルから音声信号スペクトルを推定することである。

本研究では、先行研究である統計的モデルと決定論的モデルを組み合わせた音声スペクトルの MMSE 推定 [4](以下、SD 法) を実装し、その決定論的モデルの問題点を考察した。また、その問題点に対する改善案として窓関数の特性を用いて音声信号の周波数を推定し、推定した周波数から音声を再構成する方法 [7] を採用した。この方法と先行研究における統計的モデルを組み合わせた音声スペクトル推定システムを提案した。

そして、SD 法によるシステムと提案システムとの性能比較実験を行った。実験では評価尺度にセグメンタル SNR 改善値と対数スペクトル歪みを用いた。その結果、雑音が音声よりも大きな信号ではほとんどの場合で提案システムの方が良い結果が得られた。特にレストラン雑音はすべての SNR で SD 法によるシステムを上回った。一方、元々SNR が高いときやバス雑音では、音声の周波数推定の精度が良くなかったため提案システムの方が悪い結果となった。

今後の課題としては、窓関数の特性を用いた音声スペクトル推定システムの中で行われる周波数推定の精度向上が挙げられる。

目 次

1　はじめに

音声を収録する場合、周囲が騒がしいと目的の音声以外に余計な雑音が入ってきてしまう。こういった場合に雑音低減の技術が用いられる。複数のマイクや指向性のマイクによる雑音低減の方法は実用化されている。しかし、単一マイクでサンプルされた信号の雑音低減は難しい。本研究では単一マイクで収録された信号を対象とする。

雑音低減は多くの場合「観測信号スペクトルから目的音声信号スペクトルを推定する問題」として研究されている。大きく分けて、雑音や音声の統計的性質を元に雑音低減する方法と、音声をモデル化し、その性質から音声を推定していく方法がある。

統計的性質を元にした雑音低減方法の例として、SS 法 [1] やウィーナーフィルタ、MMSE STSA[2] などが挙げられる。これらの雑音低減方法は音声スペクトル振幅の推定を行うが、スペクトル位相の推定は行わずに観測信号の位相を用いる。

音声モデルを用いた雑音低減の例の一つに、統計的モデルと決定論的モデルを組み合わせた音声モデルを用いた音声スペクトルの MMSE 推定 (以下、SD 法)[4] がある。

本研究の目的は、観測信号スペクトルから音声信号スペクトルの推定を行うことである。本研究では、この SD 法によるシステムを実装し、その決定論的モデルの問題点を考察する。また、その問題点に対する改善案として窓関数の特性を用いて音声信号の周波数を推定し、推定した周波数から音声を再構成する方法 [7] を採用した。この方法と統計的モデルを組み合わせた音声スペクトル推定システムを提案する。

2　先行研究

雑音低減について数多くの研究が行われている。本章では、先行研究としてSD法 [4] について説明する。

2.1　雑音重畳音声

観測信号 $y(t)$ は音声信号 $x(t)$ と加法性雑音 $v(t)$ の和で表されるものとする。t はサンプル点である。ただし、音声と雑音はそれぞれ平均が0で互いに無相関であるとする。

$$y(t) = x(t) + v(t) \tag{1}$$

2.2　短時間フーリエ変換

短時間フーリエ変換を用いて、音声信号を周波数領域のスペクトルに変換する。

$$X(t,k) = \sum_{\tau=-N/2}^{N/2} x(t+\tau)w(\tau)e^{-j\frac{2\pi\tau}{N}k} \tag{2}$$

τ はフレーム内の時刻、k は周波数ビン、$w(\tau)$ は窓関数、N はSTFTのフレーム長を表す。

周波数領域から時間領域に変換する際には離散逆フーリエ変換を用いる。これらは、それぞれ以下の式で定義される。

$$w(\tau)x(t+\tau) = \frac{1}{N}\sum_{k=-N/2}^{N/2} X(t,k)e^{j\frac{2\pi\tau}{N}k} \tag{3}$$

また、再構成の際に足し合わせを行うOverlap-add法 (付録A参照) を用いる。

観測信号と音声信号、雑音を短時間フーリエ変換したものをそれぞれ$Y(t,k), X(t,k), V(t,k)$とする。それぞれのスペクトルでは式 (1) より次の関係を持つ。

$$Y(t,k) = X(t,k) + V(t,k) \tag{4}$$

本研究の目標は、$Y(t,k)$ から $X(t,k)$ を推定することである。

2.3 SD 法

SD 法は、統計的 (Stochastic) モデルと決定論的 (Deterministic) モデルを用いて $Y(t,k)$ から $X(t,k)$ をそれぞれ推定し、これらを確率を用いて組み合わせる方法である。

2.3.1 統計的モデル

統計的モデルによる推定法は、音声と雑音がそれぞれ平均 0 で互いに無相関であるという仮定と音声と雑音の統計的分布を元に音声スペクトルを推定していく方法である。

統計に基づく音声スペクトル推定法としてよく知られている方法に、Wiener フィルタがある。これは、推定音声スペクトル $\hat{X}(t,k)$ と音声信号スペクトル $X(t,k)$ の最小平均二乗誤差 (MMSE) 推定に基づく雑音低減法である。

推定音声スペクトル $\hat{X}(t,k)$ を観測信号スペクトル $Y(t,k)$ とフィルタ係数 $H(t,k)$ の積で求める。

$$\hat{X}(t,k) = H(t,k)Y(t,k) \tag{5}$$

このとき、$\hat{X}(t,k)$ と $X(t,k)$ の平均二乗誤差 $J[H(t,k)]$ は次のように表される。

$$J[H(t,k)] = E[|X(t,k) - H(t,k)Y(t,k)|^2] \tag{6}$$

$J[H(t,k)]$ が最小になるように $H(t,k)$ を決める。$J[H(t,k)]$ を $H(t,k)$ について微分を行い、その値が 0 になるように $H(t,k)$ を計算すると、次のようになる (付録 B 参照)。

$$\begin{aligned}
H(t,k) =& \frac{E[|X(t,k)|^2]}{E[|Y(t,k)|^2]} && (7)\\
=& \frac{E[|X(t,k)|^2]}{E[|X(t,k)|^2] + E[|V(t,k)|^2]} && (8)\\
=& \frac{\sigma_X^2(t,k)}{\sigma_X^2(t,k) + \sigma_V^2(t,k)} && (9)\\
\sigma_X^2(t,k) =& E[|X(t,k)|^2] && (10)\\
\sigma_V^2(t,k) =& E[|V(t,k)|^2] && (11)\\
& && (12)
\end{aligned}$$

ここで $\sigma_X^2(t,k), \sigma_V^2(t,k)$ はそれぞれ音声スペクトルと雑音スペクトルの分散である。音声と雑音が互いに無相関でそれぞれ平均 0 のとき、$E[|Y(t,k)|^2] = E[|X(t,k)|^2] + E[|V(t,k)|^2]$ となる。そのため、上の式の一行目から二行目への式変換ができる。

事前 SNR $\xi(n,k)$ を次のように定義する。

$$\xi(t,k) = \frac{\sigma_X^2(t,k)}{\sigma_V^2(t,k)} \tag{13}$$

このとき、式 (9)(13) よりフィルタ係数 $H(t,k)$ は次のようになる。

$$H(t,k) = \frac{\xi(t,k)}{1+\xi(t,k)} \tag{14}$$

事前 SNR $\xi(t,k)$ が分かれば、フィルタ係数 $H(t,k)$ を計算でき、推定音声 $\hat{X}(t,k)$ が得られる。スペクトルの分散 $\sigma_X^2(t,k), \sigma_V^2(t,k)$ は実数であるため、フィルタ係数 $H(t,k)$ も実数となる。ここから、このフィルタは観測信号スペクトル $Y(t,k)$ の位相は変えずに振幅のみを変えるフィルタと言える。

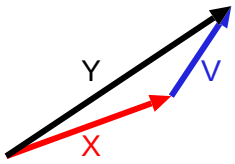

図 1: 観測信号スペクトル

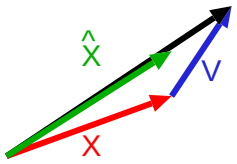

図 2: 推定音声スペクトル

しかし、真の事前 SNR は得ることが出来ないため、”Decision-Directed” 法 [2] を用いて観測信号から事前 SNR の推定値 $\hat{\xi}(t,k)$ を推定する。ただし、L はフレームシフト幅である。

$$\hat{\xi}(t,k) = (1-\alpha)P[\gamma(t,k)-1] + \alpha \cdot \frac{|\hat{X}(t-L,k)|^2}{\sigma_V^2(t-L,k)}, 0 \le \alpha < 1 \tag{15}$$

$$\gamma(t,k) = \frac{|Y(t,k)|^2}{\sigma_V^2(t,k)} \tag{16}$$

$P[\cdot]$ は次の半波整流を行なう関数である。

$$P[x] = \begin{cases} x & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

式 (15) の第一項は注目フレームの雑音スペクトルの分散と観測信号スペクトルより求められる SNR である。注目フレームの観測信号スペクトルのパワーから雑音スペクトルの分散を引くことで音声信号スペクトルのパワーを推定しているが、$X(t,k)$ と $V(t,k)$ のパワーの和は必ずしも $Y(t,k)$ のパワーとはならない。そのため、場合によっては $\gamma(t,k)-1$ が負の値をとってしまうこともある。そのため、式 (17) のように半波整流を行なう。

第二項は一つ前のフレームの雑音スペクトルの分散と推定音声スペクトルから求められる事前 SNR を表している。これは、第一項の推定に誤りがある可能性があるので、1 フレーム前の推定音声を用いた事前 SNR で平滑化をおこなっている。

この二種類の計算方法で求めた事前SNR をパラメータ α で制御することで事前 SNR$\xi(t,k)$ を求める。

雑音スペクトルの分散の推定値 $\hat{\sigma}_V^2(t,k)$ は音声ファイル先頭の「音声信号が入っていないフレーム」の観測信号のパワーの平均をとることで求める。これにより、$\hat{\sigma}_V^2(t,k)$ は t に依存しない。

2.3.2 決定論的モデル

決定論的モデルは統計的モデルとは違い、音声を調波成分の和で表し、そこから各調波成分の位相も同時に推定している。

SD 法では、音声を次のような P 個の正弦波の和で表されるものとしてモデル化を行う。

$$x(t) = \sum_{p=1}^{P} a_p e^{j\phi_p} e^{(-d_p + j f_p)t} \tag{18}$$

a_p, ϕ_p, d_p, f_p はそれぞれ第 p 調波の振幅、位相、減衰係数、周波数である。

フレーム長が長くなるほど、周波数ビン k の間隔は狭くなる。そのため、フレーム長が十分長いときは、窓関数のメインローブの周波数幅が小さくなり、その中に含まれる調波の周波数 $f_{p'}$ が高々一つだけになる。

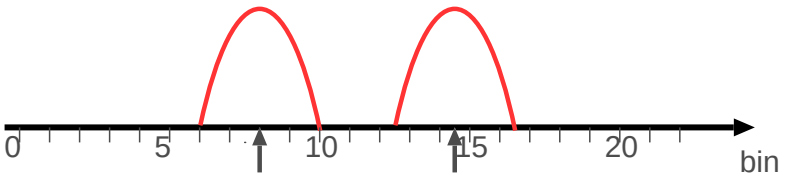

図 3: 周波数ビン k の間隔が狭いと窓関数のメインローブが狭くなり、正弦波の存在する周波数が他の正弦波の存在する周波数ビンやそのメインローブの範囲に影響を与える可能性が減る

式 (18) を短時間フーリエ変換して式変形すると、次のようになる。

$$X(t,k) = \sum_{\tau=-N/2}^{N/2} \sum_{p=1}^{P} a_p e^{j\phi_p} e^{(-d_p+jf_p)(t+\tau)} w(\tau) e^{-j\frac{2\pi\tau}{N}k} \tag{19}$$

$$\sim \sum_{\tau=-N/2}^{N/2} a_{p'} e^{j\phi_{p'}} e^{(-d_{p'}+jf_{p'})\tau} w(\tau) e^{-j\frac{2\pi\tau}{N}k} e^{(-d_{p'}+jf_{p'})t} \tag{20}$$

$$= X(0,k) e^{(-d_{p'}+jf_{p'})t} \tag{21}$$

サイドローブにある調波 p は窓関数の特性により振幅が小さくなるので、メインローブ内の調波 p' だけが残り、式 (18) の短時間フーリエ変換は式 (20) のようになる。また、式 (20) は式変形することで、式 (21) が得られる。これにより、音声モデルスペクトルがフレームシフト幅 L で 1 フレーム進むごとに $e^{(-d_{p'}+jf_{p'})L}$ の割合で位相や振幅が変化していくことがわかる。逆に言えば、$d_{p'}, f_{p'}$ が分かれば注目しているフレームの前後フレームから注目フレームの音声スペクトルを計算することができる。

式 (21) より、$X(t,k)$ と i フレーム先の音声信号 $X(t+iL,k)$ は次の関係をもつ。

$$X(t,k) = X(t+iL,k) e^{-(-d_{p'}+jf_{p'})iL} \tag{22}$$

減衰係数 d_p が 0 のときについて議論する。

式 (22) から、次のように位相揃え周波数 f を用いた平均を考える。

$$\tilde{A}(t,k;f) = \frac{1}{n_1+n_2+1} \sum_{i=-n1}^{n_2} A(t+iL,k) e^{-jfiL} \tag{23}$$

式 (25) のことを位相揃え平均と呼ぶ。

観測信号スペクトル内の正弦波スペクトル成分 $X(t+iL,k)$ は、位相揃え周波数 f が正弦波の周波数 $f_{p'}$ と一致している場合、この周波数で位相を回転するとそれぞれは式 (22) より $X(t,k)$ となり、これらの平均も $X(t,k)$ となる。

一方、雑音が白色雑音でフレームシフト幅 L が十分大きい場合、雑音スペクトル成分 $V(t+iL,k)$ は相互に無相関なので、$V(t+iL,k)$ の位相を回転したものも相互に無相関になり、これらの平均は $V(t,k)$ よりも振幅が低減することが期待できる。

以上より、観測信号に対して式 (25) を用いて注目フレームの前 n_1 フレーム、後 n_2 フレームの位相を揃えて平均をすることで音声スペクトルをそのままに雑音スペクトルの低減が期待できる。

$$\hat{X}(t,k) = \tilde{Y}(t,k;f) \tag{24}$$

$$= \frac{1}{n_1+n_2+1} \sum_{i=-n1}^{n_2} Y(t+iL,k)e^{-jfiL} \tag{25}$$

先行研究では、ESPRIT アルゴリズム [5] を用いて f を推定している。

2.3.3 確率による二つの方法の組み合わせ

音声スペクトルの状態を次の S,D,A の三状態であるとする。

S 無声子音などのように調波構造を持たないフレーム・周波数ビンの状態

D 調波構造を持つフレームの音声スペクトルが存在する周波数ビンの状態

A 調波構造を持つフレーム内で音声スペクトルが存在しない周波数ビンの状態

この節では $P_{D|Y}(y(t,k))$ を $P_{D|Y}(y)$ のように t,k を省略して表す。また、小文字は実測値、大文字は確率変数を表す。

$$\begin{aligned}
\hat{X} &= E[X|y] \\
&= \int_x x p_{X|Y}(x|y)dx \\
&= \int_x x\{p_{X|Y,D}(x|y)P_{D|Y}(y) + p_{X|Y,S}(x|y)P_{S|Y}(y) + p_{X|Y,A}(x|y)P_{A|Y}(y)\}dx \\
&= \int_x x p_{X|Y,D}(x|y)P_{D|Y}(y)dx \\
&\quad + \int_x x p_{X|Y,S}(x|y)P_{S|Y}(y)dx + \int_x x p_{X|Y,A}(x|y)P_{A|Y}(y)dx
\end{aligned} \tag{26}$$

$p_{X|Y}(x|y)$ は観測信号 y が観測されたときに音声信号が x である確率、$P_{D|Y}(y)$, $P_{S|Y}(y)$, $P_{A|Y}(y)$ は y がそれぞれの状態である確率、$p_{X|Y,D}(x|y)$, $p_{X|Y,S}(x|y)$, $p_{X|Y,A}(x|y)$ は y がそれぞれの状態であるときの x の確率である。

ここで、式 (12) の最後の項に注目する。A の状態は音声スペクトルが存在しない状態なので、このとき $x=0$ である。そのため、

$$\begin{aligned}
\hat{X} &= E[X|y] \\
&= \int_x x p_{X|Y,D}(x|y)P_{D|Y}(y)dx + \int_x x p_{X|Y,S}(x|y)P_{S|Y}(y)dx \\
&= E[X|Y,D]P_{D|Y}(y) + E[X|Y,S]P_{S|Y}(y)
\end{aligned} \tag{27}$$

と表せる。$E[X|Y,D]$ は決定論的モデルで推定した音声スペクトル、$E[X|Y,S]$ は統計的方法で推定した音声スペクトルである。それぞれ次の式で求める。

$$E[X|Y,S] = H(t,k)Y(t,k) \tag{28}$$

$$E[X|Y,D] = \tilde{Y}(t,k;f) \tag{29}$$

$$= \frac{1}{n_1+n_2+1}\sum_{i=-n1}^{n_2} Y(t+iL,k)e^{-jfiL} \tag{30}$$

ここで、$H(t,k)$ は 2.3.1 章で求めたフィルタ係数である。

それぞれの方法で推定した音声スペクトルに、観測信号がそれぞれの状態である確率 $P_{D|Y}(y)$, $P_{S|Y}(y)$ を掛けることで、より近い状態の推定値に重みを置くようにしている。

$P_{D|Y}(y), P_{S|Y}(y)$ はベイズの定理を用いて次のように変形できる。

$$P_{D|Y}(y) = \frac{p_{Y|D}(y)P_D}{p_{Y|D}(y)P_D + p_{Y|S}(y)P_S + p_{Y|A}(y)P_A} \tag{31}$$

$$= \frac{\Lambda_D}{\Lambda_D + \Lambda_S + 1} \tag{32}$$

$$P_{S|Y}(y) = \frac{p_{Y|S}(y)P_S}{p_{Y|D}(y)P_D + p_{Y|S}(y)P_S + p_{Y|A}(y)P_A} \tag{33}$$

$$= \frac{\Lambda_S}{\Lambda_D + \Lambda_S + 1} \tag{34}$$

$$\Lambda_D = \frac{p_{Y|D}(y)P_D}{p_{Y|A}(y)P_A} \tag{35}$$

$$\Lambda_S = \frac{p_{Y|S}(y)P_S}{p_{Y|A}(y)P_A} \tag{36}$$

$P_{D|Y}(y), P_{S|Y}(y)$ の分母は、それぞれの状態における観測信号 y の出現確率の合計で、分子はそれぞれ D,S の状態における観測信号 y の出現確率を表している。

観測信号の出現確率 $p_{Y|D}(y), p_{Y|S}(y), p_{Y|A}(y)$ は以下の式で与えられる。図 4、5、6、はそれぞれ状態 D,S,A での音声の分布を表す。

$$p_{Y|D}(y) = \frac{1}{\pi\sigma_V^2} exp\left(- \frac{|y - E[X|Y,D]|^2}{\sigma_V^2} \right) \tag{37}$$

$$p_{Y|S}(y) = \frac{1}{\pi\sigma_Y^2} exp\left(- \frac{|y|^2}{\sigma_Y^2} \right) \tag{38}$$

$$p_{Y|A}(y) = \frac{1}{\pi\sigma_V^2} exp\left(- \frac{|y|^2}{\sigma_V^2} \right) \tag{39}$$

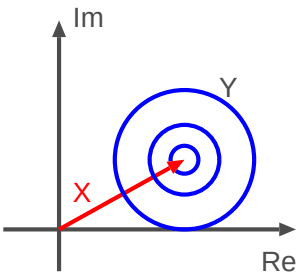

図 4: 状態 D のときの観測信号の分布

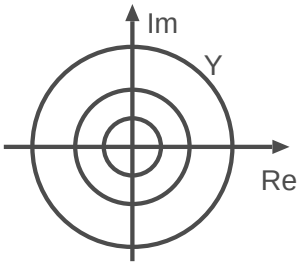

図 5: 状態 S のときの観測信号の分布

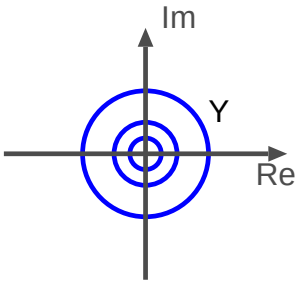

図 6: 状態 A のときの観測信号の分布

P_D, P_S, P_A はそれぞれの状態の出現確率である。先行研究では以下の式で求めている。

$$P_D = \zeta * \frac{f_c}{f_0} / \frac{N}{2} \tag{40}$$

$$P_S = 1 - \zeta \tag{41}$$

$$P_A = 1 - P_D - P_S \tag{42}$$

ζ は音声の有声音の存在確率で、先行研究では英語音声を対象として $\zeta = 0.78$ としていた。f_c は音声のエネルギーが存在する上限の周波数で、先行研究では $f_c = 2000Hz$ としている。f_0 は音声の推定基本周波数である。音声の周波数は一定でないため f_0 はフレームごとに値が変わっていく。そのため P_D, P_A は t に依存する。、P_D は、$\frac{N}{2}$ 個のビンの中に音声の周波数が $\frac{f_c}{f_0}$ 個含まれていることから上の式のように計算している。もし f_0 が推定できなかった場合、そのフレームは調波構造を持っていないとして $P_D = 0$ とする。

2.4 決定論的モデルの問題点

この先行研究の決定論的モデルにはいくつか問題点がある。

問題点 1 文献 [4] ではフレームシフト幅 L はフレーム長 N の半分で式 (25) の n_1, n_2 をそれぞれ 2 とし、前後 2 フレーム、計 5 フレームの平均を用いていた。フレームシフトがフレーム長の半分だと、1 フレームずれてもフレーム長の半分のデータが共通のものとして用いられている。そのため、各フレームの雑音スペクトル成分の位相を揃えたものは相互に無相関とならず、それらの平均は、相互に無相関の場合よりも雑音が残ってしまう。

問題点 2 位相揃え平均において、位相揃え周波数 f が正弦波の周波数 $f_{p'}$ と一致しない場合、正弦波スペクトル成分の位相揃え平均 $\hat{X}(t,k)$ は正弦波スペクトル成分 $X(t,k)$ と比べ、振幅が減衰してしまう。

問題点 3 通常では、音声スペクトルの周波数は一定ではなく時間と共に変化してる。しかし、位相揃え平均は注目フレームの推定周波数のみを用いて行われる。時間と共に周波数が変化していく場合、注目フレームの推定周波数 $f_{p'}$ のみを用いて位相揃え平均したスペクトルは真のスペクトルと異なる値となってしまう。

問題点 4 式 (18) によって定義した音声モデルを短時間フーリエ変換する際に、式 (19) から式 (20) への変換はフレーム長が十分でないと成り立たないという問題がある。例えば、サンプリング周波数が 8kHz、フレーム長が 256 点の場合では、ビンの間隔は 31.25Hz となる。一般的によく用いられるハミング窓のメインローブは前後 2 ビンであるため、メインローブの範囲は 125Hz となる。しかし、男性音声の基本周波数は 100Hz を下回ることもあり、メインローブの範囲に他の調波成分が入ってしまう可能性が高い。

3　正弦成分抽出方法

本研究では、前章で述べたSD法の問題点1,2の改良システムの構築とその性能評価を行った。この改良システムは、吉田[7]が特許出願中の原理に基づいたシステムである。

3.1　正弦波の短時間フーリエ変換

信号$x(t)$が以下のように周波数f、振幅A、初期位相ϕの正弦波信号であるとする。

$$x(t) = Ae^{j2\pi f't+\phi} \tag{43}$$

この正弦波に対し、次のように短時間フーリエ変換を行う。

$$\begin{aligned} X(t,k) =& \int_{-\infty}^{\infty} x(t+\tau)w(\tau)e^{-j2\pi k\frac{\tau}{N}}d\tau & (44)\\ =& x(t)\int_{-\frac{N}{2}}^{\frac{N}{2}} w(\tau)e^{j2\pi(fN-k)\frac{\tau}{N}}d\tau & (45)\\ W(\xi) =& \frac{1}{N}\int_{-\frac{N}{2}}^{\frac{N}{2}} w(\tau)e^{j2\pi\xi\frac{\tau}{N}}d\tau & (46)\\ X(t,k) =& x(t)W(f'N-k)N & (47) \end{aligned}$$

3.2　窓関数の特性

音声分析の際にはハミング窓やハニング窓がよく用いられるが、他にも種類があり、minimum 3-term窓[6]もその一つである。図7に示すようにハミング窓は前後2ビンのメインローブ、minimum 3-term窓は前後3ビンのメインローブとなっていて、メインローブはハミング窓の方が狭い。しかし、サイドローブの減衰を見るとハミング窓が-42dBくらいしか減っていないのに対してminimum 3term窓は-71.48dBとなり、より減衰していることがわかる。サイドローブに出てくる振幅の影響を極力減らすため、本研究ではminimum 3-term窓を用いる。

Minimum 3-term窓は次の式で定義される。

$$\begin{aligned} w(t) =& 1+\frac{a_1}{a_0}\cos\left(\frac{2\pi}{N}t\right)+\frac{a_2}{a_0}\cos\left(\frac{4\pi}{N}t\right) & (48)\\ |t| <& \frac{N}{2} & \\ a_1 =& 0.4973406, a_2 = 0.0782793, a_0 = 1-a_1-a_2 & \end{aligned}$$

また、$W(\xi)$ は以下のように表される (付録 C 参照)。

$$W(\xi) = \frac{\sin(\pi\xi)}{\pi\xi}\left[1 - \frac{a_1}{a_0}\frac{\xi^2}{\xi^2 - 1^2} + \frac{a_2}{a_0}\frac{\xi^2}{\xi^2 - 2^2}\right] \tag{49}$$

窓関数が実数で偶関数であることから、式 (46) のように $W(\xi)$ を定義するとこの値は実数となる。このため、$X(t,k)$ は k によらず位相は $x(t)$ と同相であり、振幅は $W(f'N-k)$ 倍となる。

minimum 3-term 窓の場合、周波数 f の正弦波の短時間フーリエ変換は $f'N-3 < k < f'N+3$ の範囲の周波数ビン k の $X(t,k)$ 上に現れるといえる。

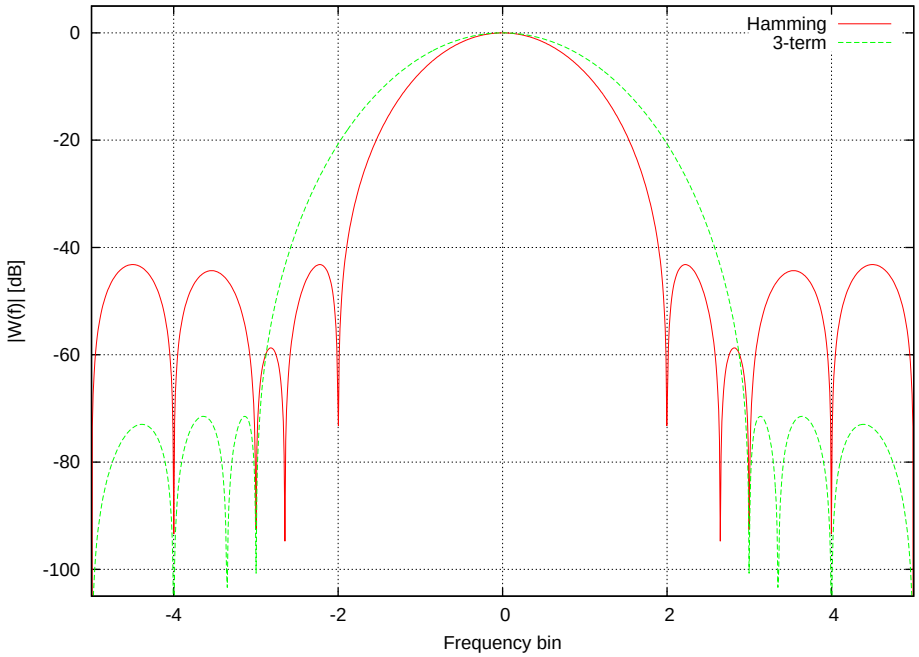

図 7: 窓関数の周波数特性

3.3 窓関数が Minimum 3-term 窓のときの正弦波推定

正弦波の周波数 f' にもっとも近い中心周波数を持つ周波数ビンを k とする。このとき、$\xi = f'N - k$ とすると ξ は $-0.5 < \xi < 0.5$ の範囲にある。

付近に他の正弦波の成分がない場合は、付近の周波数ビン $k+i(i = \cdots, -2, -1, 0, 1, 2. \cdots)$ での値は

$$X(t,k+i) = x(t)W(\xi-i)N \tag{50}$$

$$X(t,k) = x(t)W(\xi)N \tag{51}$$

となる。$W(\xi-i)$ も $W(\xi)$ も実数であるため、$x(t,k+i)$ と $X(t,k)$ は同相であり、また振幅比は $W(\xi-i)$ と $W(\xi)$ の比となる。

$$\frac{|X(t,k+i)|}{|X(t,k)|} = \frac{|W(\xi-i)|}{|W(\xi)|} \tag{52}$$

この式の右辺は信号に依存せず、窓関数のみで求まる。この式の右辺の dB 値を以下のように ξ の関数として定義する。

$$r_i(\xi) = 10\log_{10}\left|\frac{W(\xi-i)}{W(\xi)}\right|^2 \tag{53}$$

図 8 は minimum 3-term 窓のときの ξ と $r_i(\xi)$ の関係である。

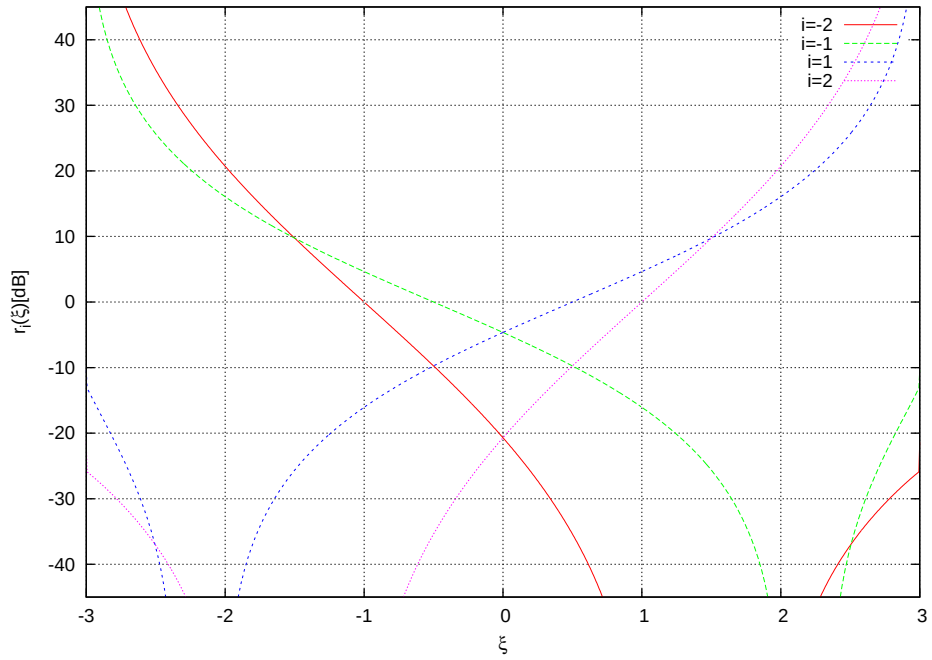

図 8: k 番目のビンに対する k+i 番目のビンとの比

$r_i(\xi)$ の逆関数 $r_i^{-1}(r)$ も窓関数だけで定まる。そのため、$X(t,k+i)$ と $X(t,k)$ の比の dB 値 r を

$$r = 10\log_{10}\frac{|X(t,k+i)|^2}{|X(t,k)|^2} \tag{54}$$

のように求め、正弦波の周波数 f を逆関数 $r_i^{-1}(r)$ を用いて

$$\xi = r_i^{-1}(r) \tag{55}$$

$$f' = (k + \xi)/N \tag{56}$$

のように求めることが出来る。

窓関数が Minimum 3-term の場合、逆関数 $r_i^{-1}(r)$ は以下のように近似できることが示されている。

$$\begin{aligned} r_{-2}^{-1} &= 2.118824157 * \tanh(-0.024833631 * r) - 1 \\ r_{-1}^{-1} &= 2.498565584 * \tanh(-0.043232123 * r) - 0.5 \\ r_{1}^{-1} &= 2.498565584 * \tanh(0.043232123 * r) + 0.5 \\ r_{2}^{-1} &= 2.118824157 * \tanh(0.024833631 * r) + 1 \end{aligned} \tag{57}$$

式 (56) で求めた f' と式 (58) を用いて $x(t)$ を求めることができるため、式 (43) より正弦波 $x(t)$ の振幅 A, 初期位相 ϕ を求めることができる。

$$x(t) = \frac{X(t,k)}{W(f'N - k)N} \tag{58}$$

3.4 雑音がある場合

観測信号には雑音が混じっているため、必ずしも $Y(t,k)$ と近傍のビン $Y(t,k+i)$ は同相ではない。そこで、これらのビン $k+i$ のそれぞれに対して同一の位相揃え周波数 f を用いて位相揃え平均を行なうことによって観測信号中の雑音成分を低減させ、それぞれ $X(t,k+i)$ の推定値を得る。これらから相互に同位相の組み合わせに対して上記の方法を用いて正弦波成分の周波数 f' を推定し、その正弦波成分の振幅と位相を推定する方法が示されている。

位相揃え平均するフレーム数を $n = n1 = n2$ とした場合、正弦波の位相揃え平均の式

は次のように変形できる。

$$\begin{aligned}
\tilde{X}(i,k;f) =& \frac{1}{2n+1}\sum_{m=i-n}^{i+n} X(m,k)e^{j(f'+\Delta f)(i-m)L} \\
=& \frac{1}{2n+1}\sum_{m=i-n}^{i+n} X(i,k)e^{j(\Delta f)(i-m)L} && (59)\\
=& X(i,k)\frac{1}{2n+1}\sum_{m=i-n}^{i+n} e^{j(\Delta f)(i-m)L} && (60)\\
=& X(i,k)\alpha(f) && (61)\\
\alpha(f) =& \frac{1}{2n+1}\sum_{m=i-n}^{i+n} e^{j(\Delta f)(i-m)L} && (62)
\end{aligned}$$

$\alpha(f)$ を描いたのが図 9 である。式 () より正弦波スペクトル成分の位相は k によらず同相である。また、位相揃え周波数 f が正弦波の周波数 f' と Δf ずれているときの振幅低減率は

$$\begin{aligned}
\frac{|\tilde{X}(t,k+i;f)|}{|\tilde{X}(t,k;f)|} &= \frac{|X(t,k+i)\alpha(f)|}{|X(t,k)\alpha(f)|} && (63)\\
&= \frac{|X(t,k+i)|}{|X(t,k)|} && (64)
\end{aligned}$$

となり、k に依存せず、隣のビンとの振幅比は変化しない。

一方、雑音成分の位相揃え平均は次の式のように近似できることが示されている。

$$\begin{aligned}
\tilde{V}(t,k;f) &= \frac{1}{2n+1}\sum_{i=-n}^{n} V(t+iL,k)e^{-jfiL} \\
&\sim V(t,k;f)\beta(f) && (65)\\
\beta(f) &= \frac{1}{2n+1}\sum_{i=-n}^{n} w(iL)e^{-jfiL} && (66)
\end{aligned}$$

$\beta(f)$ を描いたのが図 10 である。この図から、ビンの中心周波数で位相揃え平均を行うと k ビンよりも前後の $k+i$ ビンの方が雑音が低減される。

位相揃え平均したスペクトルは式 (1) と同様に式 (67) の関係を持つ。

$$\tilde{Y}(t,k;f) = \tilde{X}(t,k;f) + \tilde{V}(t,k;f) \quad (67)$$

そのため、次のように観測信号の位相揃え平均を行い、$k\pm2$ ビンと $k\pm1$ ビンで式 (68) のように比をとる。このとき、雑音は位相揃え平均を行なうことで k ビンから離れるほど低減されるので、式 (69) のように近似でき、式 (64) によって式 (70) のようになる。

$$\frac{|\tilde{Y}(t,k\pm2;f)|}{|\tilde{Y}(t,k\pm1;f)|} = \frac{|\tilde{X}(t,k\pm2;f)+\tilde{V}(t,k\pm2;f)|}{|\tilde{X}(t,k\pm1;f)+\tilde{V}(t,k\pm1;f)|} \tag{68}$$

$$\sim \frac{|\tilde{X}(t,k\pm2;f)|}{|\tilde{X}(t,k\pm1;f)|} \tag{69}$$

$$= \frac{|X(t,k\pm2)|}{|X(t,k\pm1)|} \tag{70}$$

以上より、式 (71) のようになることに気をつけながら r を式 (72) のように定義する。

$$\frac{|X(t,k\pm2)|}{|X(t,k\pm1)|} = \frac{|W((\xi\mp1)\mp1)|}{|W(\xi\mp1)|} \tag{71}$$

$$r = 10\log_{10}\frac{|\tilde{Y}(t,k\pm2;f)|}{|\tilde{Y}(t,k\pm1;f)|} \tag{72}$$

$$\tag{73}$$

このとき、式 (74) より $\xi\mp1$ を求められ、式 (75) から推定周波数 f' が得られ、式 (58)(43) により正弦波の振幅、位相が推定できる。

$$\xi\mp1 = r_{\pm1}^{-1}(r) \tag{74}$$

$$f' = (k+(\xi\mp1))/N \tag{75}$$

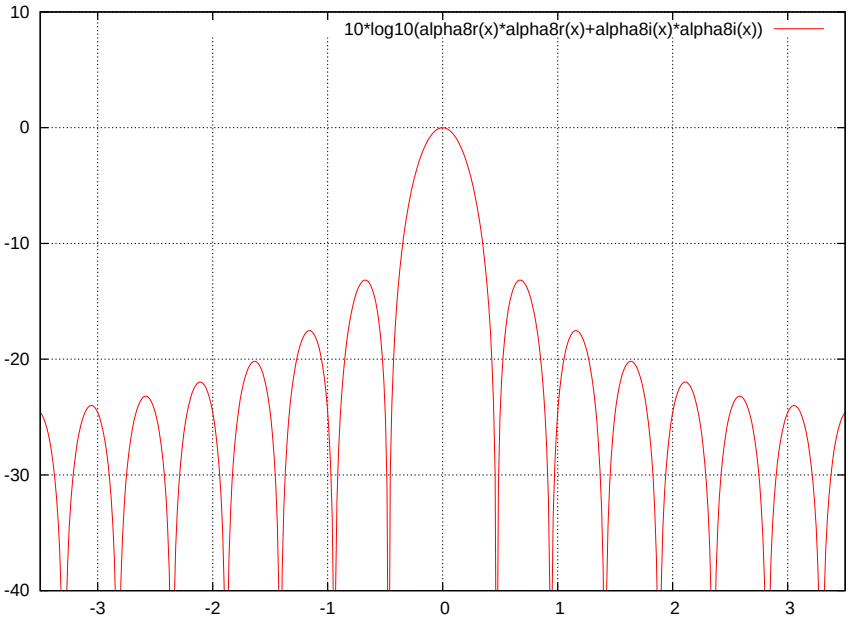

図 9: ビン周波数と位相揃え周波数との差を横軸とした場合の正弦波の振幅低減率 [dB]

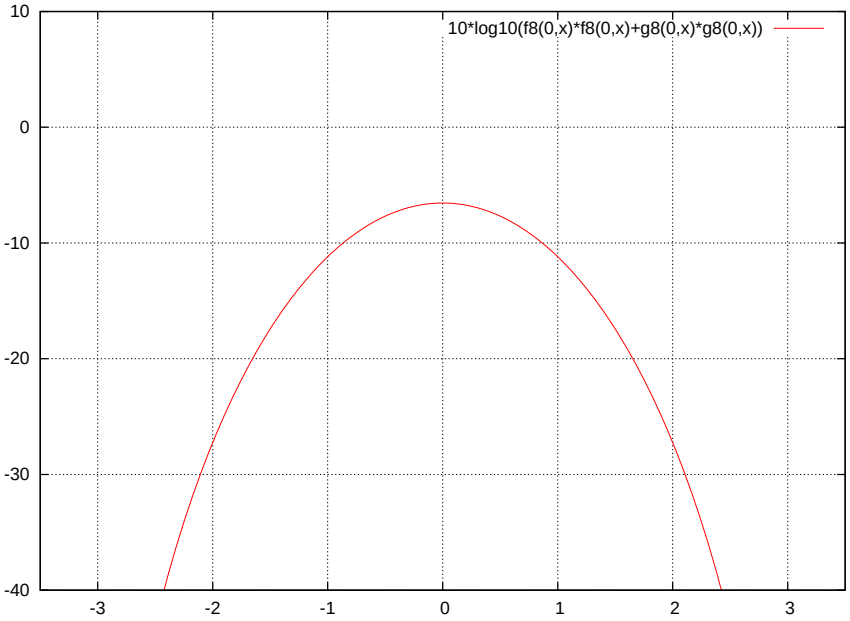

図 10: ビン周波数と位相揃え周波数との差を横軸とした場合の雑音低減率 [dB]

4　提案方法

SD 法の決定論的モデルの問題点の改善案として窓関数の特性を用いた正弦成分抽出方法を採用した。しかし、雑音の影響によって本来音声の調波成分が存在しない周波数を誤って推定してしまう可能性がある。そのため、次のように二段階で周波数の推定を行なった。

(1) それぞれの周波数ビンにおいて、ビンの中心周波数で位相揃え平均をとる

(2)50Hz か 8000Hz の範囲で、次の条件を満たすビンを音声の周波数に近いビンの候補とする

- 一つ上のビンと二つ上のビン、一つ下のビンと二つ下のビンの位相差のどちらかが 15 °以下である
- 観測信号スペクトルの振幅が観測信号先頭フレームから推定した雑音スペクトルの分散の推定値 $\hat{\sigma}_V^2(t,k)$ の 2 倍よりも大きい

(3) それぞれの周波数ビン候補 k に対して、$\hat{Y}(t,k+1)$ と $\hat{Y}(t,k+2)$、$\hat{Y}(t,k-1)$ と $\hat{Y}(t,k-2)$ の比を計算し、そこから式（57）を用いて一段階目の周波数 ξ を求める

(4) 周波数ビン候補 k とその周辺 2 ビンで、周波数 ξ を用いて位相揃え平均を求める

(5) 周波数ビン候補 k の一つ上のビンと二つ上のビン、一つしたのビンと二つ下のビンの位相揃え平均したスペクトルの位相差のどちらかが 10 °以下であるビンをさらに厳選する

(6) (5) で選んだ各ビン候補について $\hat{Y}(t,k+1)$ と $\hat{Y}(t,k+2)$, $\hat{Y}(t,k-1)$ と $\hat{Y}(t,k-2)$ の比をそれぞれ計算し、式（57）からそれぞれ周波数 ξ を求める

(7) 最後に求めた 2 つの ξ のそれぞれがビン候補から 0.5 ビン以内で、かつ ξ の平均がビン候補から 0.5 ビン以内であるとき、平均した ξ から音声の推定周波数を求める

(8) 式 (51) を用いて $x(t)$ を求め、また式 (51) を用いて音声スペクトルの再構成を行う

こうして再構成した音声スペクトルと統計的推定法で推定した音声スペクトルを組み合わせることで、推定音声スペクトルを得る。

5　決定論的方法の問題点の検証

2 章では決定論的方法の問題点をいくつか述べた。その中でも、提案方法と関わりのある問題点 1 と問題点 2 について検証を行なった。

5.1　問題点 1 の検証

決定論的モデルの問題点 1 は次のものである。

問題点 1　文献 [4] ではフレームシフト幅 L はフレーム長 N の半分で式 (25) の n_1, n_2 をそれぞれ 2 とし、前後 2 フレーム、計 5 フレームの平均を用いていた。フレームシフトがフレーム長の半分だと、1 フレームずれてもフレーム長の半分のデータが共通のものとして用いられている。そのため、各フレームの雑音スペクトル成分の位相を揃えたものは相互に無相関とならず、それらの平均は、相互に無相関の場合よりも雑音が残ってしまう。

この問題を検証するため、平均 0、分散 1 の正規乱数による信号に対し位相揃え平均を行い、それにより雑音の低減率を計算した。

5.1.1　実験条件

平均 0、分散 1 の正規乱数による信号に対し、以下の条件でビンの中心周波数で位相揃え平均をし、振幅のパワーの低減率を計算した。全ての場合において、位相揃え平均を行う範囲を 3 フレーム分の長さにした。これは、観測信号で位相揃え平均をとる際に位相揃え平均を行う範囲を広くとると音声の調波周波数が時間と共に変化し、位相回転の周波数がずれてしまうのを防ぐためである。先行研究の実験条件はサンプリング周波数 8000Hz、フレーム長 256 点、フレームシフト 128 点、位相揃え平均のフレーム数は前後 2 フレームであった。本検証ではサンプリング周波数を 2 倍の 16000Hz で行う。そのため、フレーム長 512 点、フレームシフト 256 点、前後 2 フレームによる位相揃え平均が先行研究の条件に対応する。

表 1: 実験条件

サンプリング周波数	16kHz
計算フレーム数	1000
分析窓	ハミング窓
フレーム長	512 点
フレームシフト幅	512 点 (フレーム長),256 点 (フレーム長の 1/2) 128 点 (フレーム長の 1/4),64 点 (フレーム長の 1/8)
位相揃え平均する フレームの範囲	フレームシフト幅 512 点 (フレーム長):前後 1 フレーム フレームシフト幅 256 点 (フレーム長の 1/2):前後 2 フレーム フレームシフト幅 128 点 (フレーム長の 1/4):前後 4 フレーム フレームシフト幅 64 点 (フレーム長の 1/8):前後 8 フレーム

5.1.2 実験結果

表 2 は検証の結果である。3 フレームの区間において雑音を位相揃え平均すると、大体 3 割程度の雑音パワー低減率になることが分かった。各フレームにおいて雑音が無相関である場合には位相揃え平均をとるとパワーが 1/(足したフレーム数) になることを考えると、フレームシフト幅 128 点や 64 点の結果はかなりパワーが残ってしまったことになる。

表 2: 検証結果 1

フレームシフト幅	位相揃え平均した フレーム数	雑音パワー 低減率	無相関のときの の雑音パワー低減率	理論値
512 点 (フレーム長)	3 フレーム (前後 1 フレーム)	0.33	0.33	0.33
256 点 (フレーム長の 1/2)	5 フレーム (前後 2 フレーム)	0.27	0.20	0.20
128 点 (フレーム長の 1/4)	9 フレーム (前後 4 フレーム)	0.27	0.11	0.11
64 点 (フレーム長の 1/8)	17 フレーム (前後 8 フレーム)	0.30	0.06	0.06

5.2 問題点 2 の検証

決定論的モデルの問題点 2 は次のものである。

問題点 2 位相揃え平均において、位相揃え周波数 f が正弦波の周波数 f' と一致しない場合、正弦波スペクトル成分の位相揃え平均 $\hat{X}(t,k)$ は正弦波スペクトル成分 $X(t,k)$ と比べ、振幅が減衰してしまう。

この問題の検証を行なうため、正弦波信号に対して正弦波周波数とずれた周波数で位相揃え平均を行い、どのくらい振幅が減衰するか実験した。また、理論値を計算し比較を行なった。

5.2.1 理論値

問題点 2 を検証するために、理論値を計算した。式 (25) の $A(t,k)$ に $X(t,k)$ を代入し、位相揃え平均 f を正弦波周波数 f' に誤差 Δf を加えた $f'+\Delta f$ として計算すると次のようになる。

$$\begin{aligned}\tilde{X}(i,k;f) =& \frac{1}{n_1+n_2+1}\sum_{n=i-n1}^{i+n_2} X(n,k)e^{j(f'+\Delta f)(i-n)L} \\ =& \frac{1}{n_1+n_2+1}\sum_{n=i-n1}^{i+n_2} X(i,k)e^{j(\Delta f)(i-n)L} \qquad (76) \\ =& X(i,k)\frac{1}{n_1+n_2+1}\sum_{n=i-n1}^{i+n_2} e^{j(\Delta f)(i-n)L} \qquad (77)\end{aligned}$$

5.2.2 実験条件

周波数 250Hz の正弦波信号に対し、以下の条件で位相揃え平均をし、振幅のパワーの低減率を計算した。全ての場合において、位相揃え平均を行う範囲を 3 フレーム分の長さにした。これは、観測信号で位相揃え平均をとる際に位相揃え平均を行う範囲を広くとると音声の調波周波数が時間と共に変化し、位相回転の周波数がずれてしまうのを防ぐためである。位相揃え周波数は 8 ビン (250Hz) から 1/8 ビンずつずれた周波数をそれぞれ用い、最大で 3 ビン離れたところまで実験を行なった。

表 3: 実験条件

サンプリング周波数	16kHz
計算フレーム数	1000
分析窓	ハミング窓
フレーム長	512 点
フレームシフト幅	256 点 (フレーム長の 1/2),128 点 (フレーム長の 1/4), 64 点 (フレーム長の 1/8)
位相揃え平均する フレームの範囲	フレームシフト幅 256 点 (フレーム長の 1/2):前後 2 フレーム フレームシフト幅 128 点 (フレーム長の 1/4):前後 4 フレーム フレームシフト幅 64 点 (フレーム長の 1/8):前後 8 フレーム

5.2.3 実験結果

位相揃え平均による振幅低減率は図 11、12、13 である。横軸が推定周波数の誤差のビン数で、縦軸がパワーの dB 値である。実線が理論値、点が計測値である。図 11 はフレームシフト幅 256 点 (フレーム長の 1/2)、前後 2 フレームでの結果、図 12 はフレームシフト幅 128 点 (フレーム長の 1/4)、前後 4 フレームでの結果、図 13 はフレームシフト幅 64 点 (フレーム長の 1/8)、前後 8 フレームでの結果を表す。

調波成分の周波数推定に誤差が生じると、それぞれの図のように振幅のパワーが小さくなってしまうことが確認できた。また、フレームシフト幅がフレーム長の 1/2, 前後 2 フレームでの位相揃え平均の場合、2 ビンも周波数推定がずれてしまうと本来存在しない 2 ビン離れた周波数に振幅が現れてしまうことも分かった。

ここから、決定論的モデルは周波数の推定が重要であることが言える。

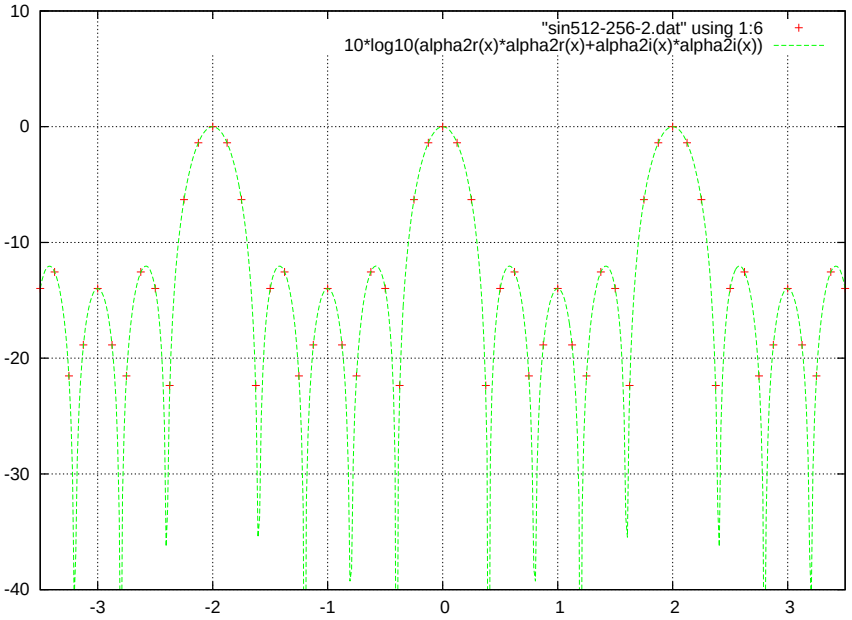

図 11: フレームシフト幅 256 点、前後 2 フレームでの位相揃え平均の、正弦波の推定周波数のずれと振幅低減率の関係

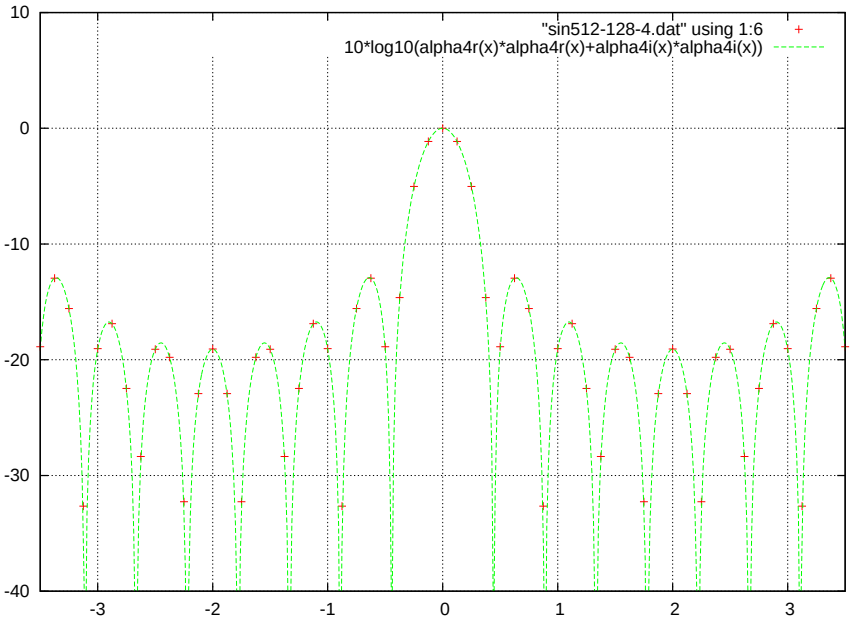

図 12: フレームシフト幅 128 点、前後 4 フレームでの位相揃え平均の、正弦波の推定周波数のずれと振幅低減率の関係

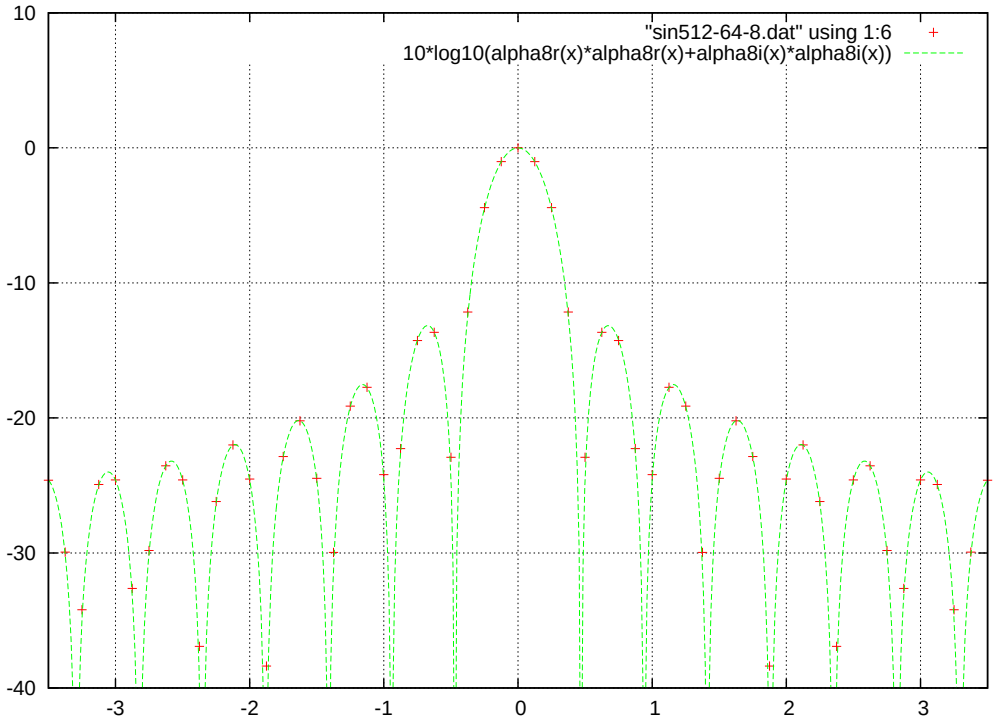

図 13: フレームシフト幅 64 点、前後 8 フレームでの位相揃え平均の、正弦波の推定周波数のずれと振幅低減率の関係

6　正弦成分抽出方法の検証

3 章では正弦成分抽出方法について説明した。そのなかで、雑音が重畳されている場合には位相揃え平均を行うことで正弦波成分抽出が行えることが期待できるとあった。

本章では、位相揃え平均を行うことでどの程度正弦波成分が抽出できるのか検証した。

6.1　位相揃え平均による雑音低減

正弦成分抽出方法では、k ビンの中心周波数で位相揃え平均を行うことで k ビンから離れた周波数での雑音低減が期待できると説明した。

実際にどのくらい雑音が低減されるのか実験を行った。

6.1.1　実験条件

平均 0、分散 1 の正規乱数による信号に対し、以下の条件でビンの中心周波数で位相揃え平均をし、振幅のパワーの低減率を計測した。

表 4: 実験条件

サンプリング周波数	16kHz
計算フレーム数	1000
分析窓	Minimum 3-term 窓
フレーム長	512 点
フレームシフト幅	256 点 (フレーム長の 1/2),128 点 (フレーム長の 1/4), 64 点 (フレーム長の 1/8)
位相揃え平均する フレームの範囲	フレームシフト幅 256 点 (フレーム長の 1/2):前後 2 フレーム フレームシフト幅 128 点 (フレーム長の 1/4):前後 4 フレーム フレームシフト幅 64 点 (フレーム長の 1/8):前後 8 フレーム

6.1.2　結果

横軸が推定周波数の誤差のビン数で、縦軸がパワーの dB 値である。実線が理論値、点が計測値である。図 14 はフレームシフト幅 256 点 (フレーム長の 1/2)、前後 2 フレームでの結果、図 15 はフレームシフト幅 128 点 (フレーム長の 1/4)、前後 4 フレームでの結果、図 16 はフレームシフト幅 64 点 (フレーム長の 1/8)、前後 8 フレームでの結果を表す。

実測値と理論値で誤差がでたものの、おおむね理論値と同じような傾向となった。フレームシフト幅 64 点 (フレーム長の 1/8)、前後 8 フレームのとき、位相揃え平均が注目してるビンの中心周波数から離れた周波数ほど雑音が低減でき、実測値では 2 ビン離れた周波数での位相揃え平均によって-22dB の低減ができることが確認できた。しかしほかの条件では位相揃え周波数が注目しているビンの中心周波数から離れてもそれほど雑音が低減されなかった。

そのため、雑音が重畳された音声から正弦成分抽出方法で正弦成分を抽出するとき、今回の条件の中ではフレームシフト幅 64 点 (フレーム長の 1/8)、前後 8 フレームでの位相揃え平均を用いるのが最良といえる。

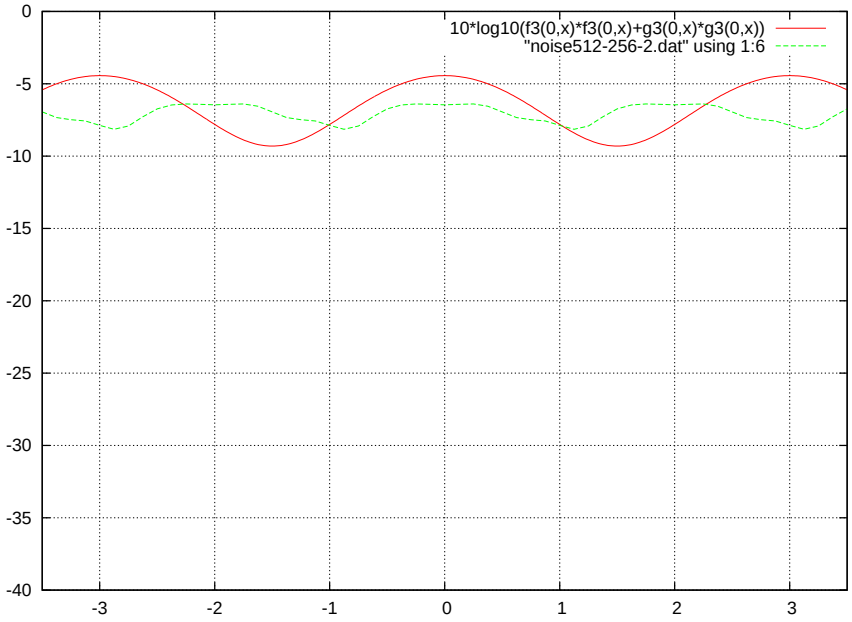

図 14: フレームシフト幅 256 点、前後 2 フレームでの位相揃え平均の、ビン周波数と位相揃え周波数との差と雑音低減率の関係

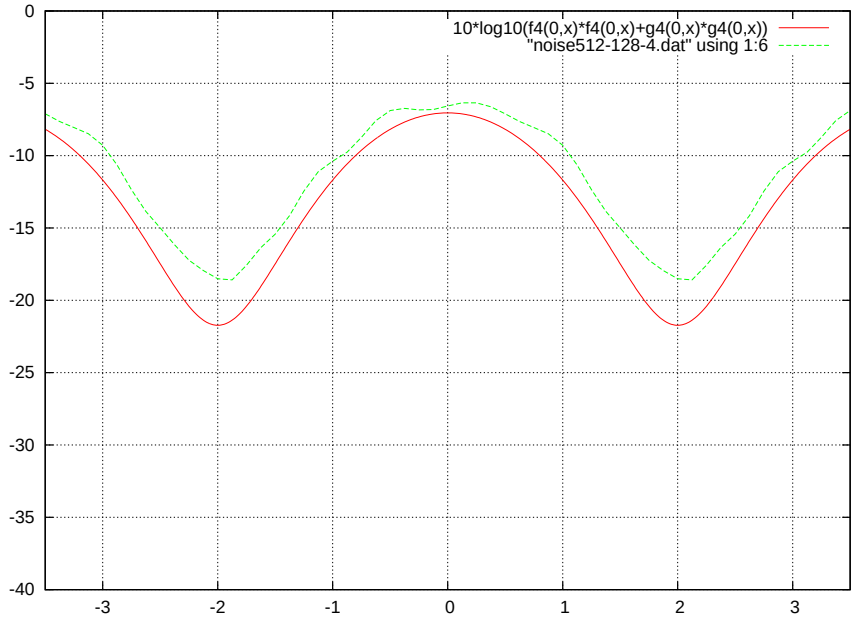

図 15: フレームシフト幅 128 点、前後 4 フレームでの位相揃え平均の、ビン周波数と位相揃え周波数との差と雑音低減率の関係

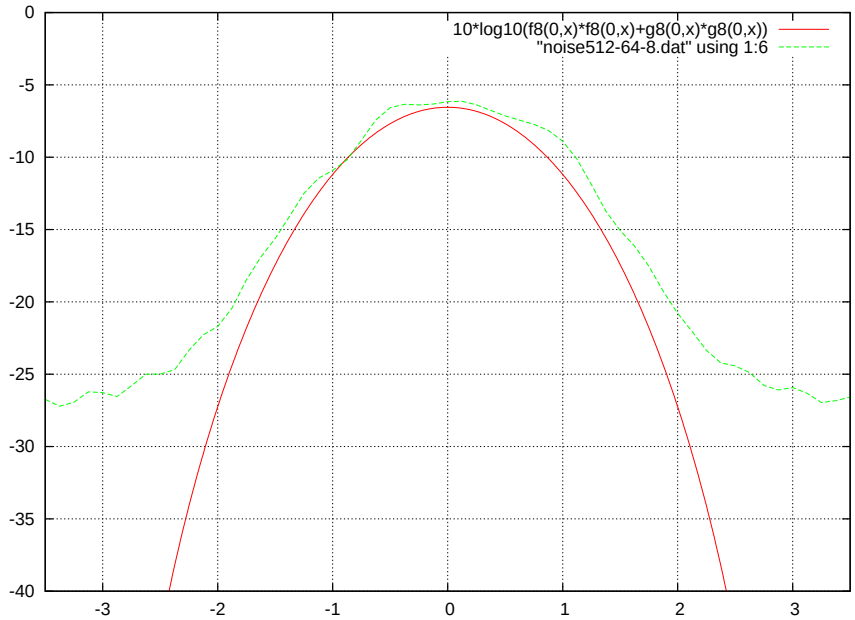

図 16: フレームシフト幅 64 点、前後 8 フレームでの位相揃え平均の、ビン周波数と位相揃え周波数との差と雑音低減率の関係

6.2 位相揃え平均による雑音重畳信号の振幅比

雑音が重畳された正弦波に対して位相揃え平均をを用いた場合、$k-1$ ビンと $k-2$ ビン、$k+1$ ビンと $k+2$ ビンでの振幅比は位相揃え平均をする前より正弦波の振幅比に近づいているのか検証を行った。

6.2.1 実験条件

周波数が 250Hz と 240Hz の二種類の正弦波に白色雑音を 0dB で重畳した信号を観測信号として実験を行った。このとき、k は正弦波に一番近いビンとなるため、$k=8$ となる。フレームシフト幅、位相揃え平均するフレームの範囲は先の実験よりフレームシフト幅 64 点 (フレーム長の 1/8)、前後 8 フレームとした。その他の条件は以下の通りである。

次の二種類の検証を行った。

1 位相揃え平均をすることで位相揃え平均前よりも正弦波の位相に近づいたかどうか

2 位相揃え平均をすることで位相揃え平均前よりも正弦波の振幅比に近づいたかどうか

表 5: 実験条件

サンプリング周波数	16kHz
計算フレーム数	5000
分析窓	Minimum 3-term 窓
フレーム長	512 点
フレームシフト幅	64 点 (フレーム長の 1/8)
位相揃え平均するフレームの範囲	フレームシフト幅 64 点 (フレーム長の 1/8):前後 8 フレーム
正弦波周波数に一番近い周波数ビン k	8 ビン

6.2.2 結果

図 17 から 20 までは位相がどれだけ正弦波に近づいたかを表す図である。横軸が正弦波と観測信号との位相差、縦軸が正弦波と位相揃え平均したスペクトルの位相差である。斜めの線よりも下側にデータが集まっていれば位相揃え平均したスペクトルは観測信号スペクトルよりも正弦波スペクトルの位相に近くなったと言える。図 17 は正弦波の周波

数が 250Hz のときの 6 ビンと 7 ビンでの位相差、図 18 は正弦波の周波数が 250Hz のときの 9 ビンと 10 ビンでの位相差、図 19 は正弦波の周波数が 240Hz のときの 6 ビンと 7 ビンでの位相差、図 19 は正弦波の周波数が 240Hz のときの 9 ビンと 10 ビンでの位相差を表している。赤色は 6 ビンまたは 10 ビンの結果、緑色は 7 ビンまたは 9 ビンの結果である。

図 17,18 より 250Hz の正弦波を入力としたときの結果はどのビンにおいても位相揃え平均後のスペクトルの位相が正弦波のスペクトルの位相に近くなっている。また、図 19,18 より、240Hz の正弦波を入力としたときの位相揃え平均後のスペクトルの位相が 6,7,9 ビンにおいて位相揃え平均前より正弦波の位相に近くなった。しかし、10 ビンでは位相揃え平均による改善が見られなかった。これは、正弦波の周波数が 240Hz で 8 ビンよりも低いところにあるため、10 ビンに出てくる正弦波の振幅がかなり減衰してしまったためと考えられる。

正弦波の周波数に対して 2 ビン以内の範囲であれば、位相揃え平均によって正弦波の位相を近似できると言える。

図 21 から 24 までは振幅比の図である。横軸に試行回数、縦軸に振幅比を表している。図 21 は正弦波の周波数が 250Hz のときの $k-1$ ビンと $k-2$ ビンの振幅比、図 22 は正弦波の周波数が 250Hz のときの $k+1$ ビンと $k+2$ ビンの振幅比、図 23 は正弦波の周波数が 240Hz のときの $k-1$ ビンと $k-2$ ビンの振幅比、図 23 は正弦波の周波数が 240Hz のときの $k+1$ ビンと $k+2$ ビンの振幅比を表している。赤色が観測信号、緑色が位相揃え平均後、青色が正弦波の振幅比を表している。

どの場合においても、位相揃え平均を行うことで正弦波の振幅比に近づいていることがわかる。この結果からも、位相揃え平均が有効であると言える。

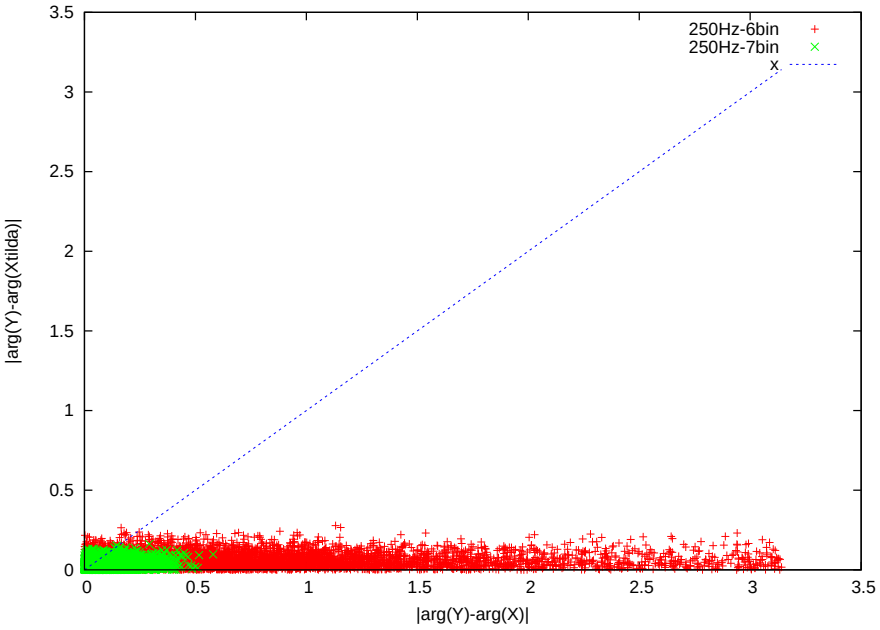

図 17: 位相揃え平均前後における正弦波スペクトル位相との位相差の変化 (正弦波周波数 250Hz、6 ビンと 7 ビンの結果)

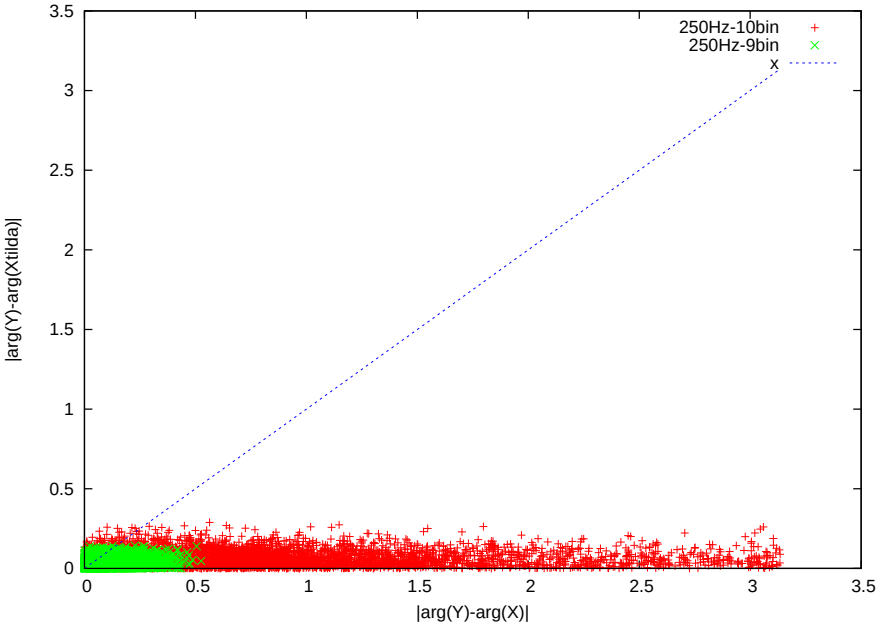

図 18: 位相揃え平均前後における正弦波スペクトル位相との位相差の変化 (正弦波周波数 250Hz、9 ビンと 10 ビンの結果)

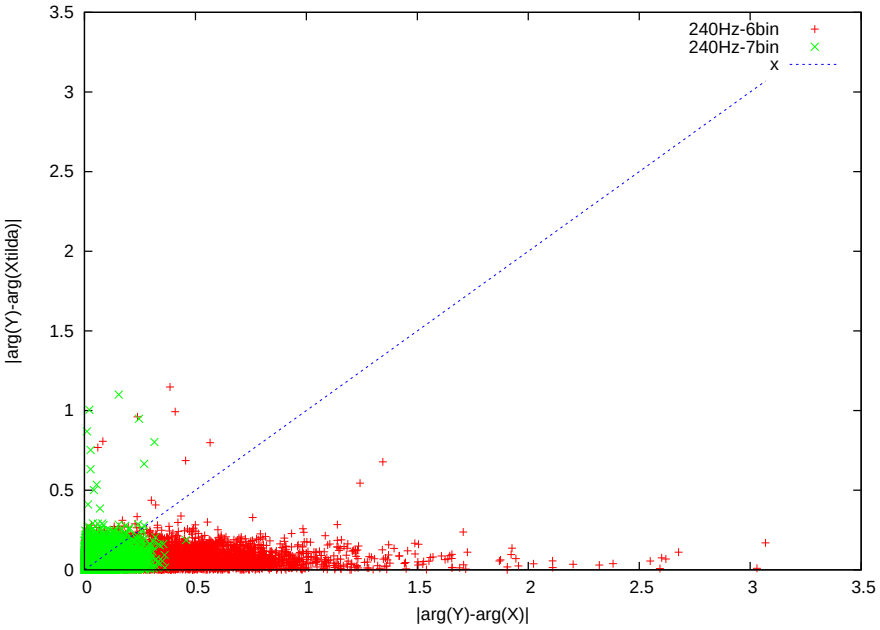

図 19: 位相揃え平均前後における正弦波スペクトル位相との位相差の変化 (正弦波周波数 240Hz、6 ビンと 7 ビンの結果)

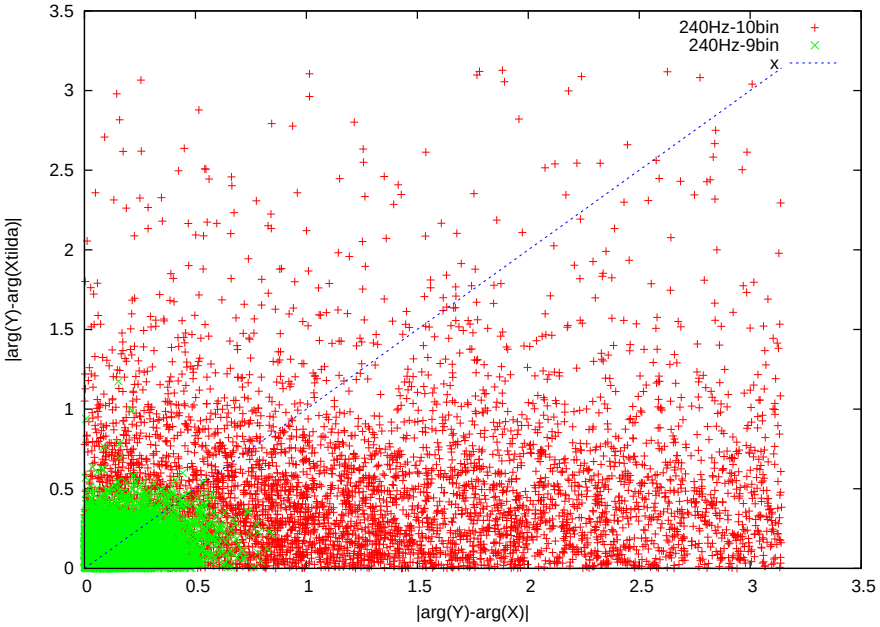

図 20: 位相揃え平均前後における正弦波スペクトル位相との位相差の変化 (正弦波周波数 240Hz、9 ビンと 10 ビンの結果)

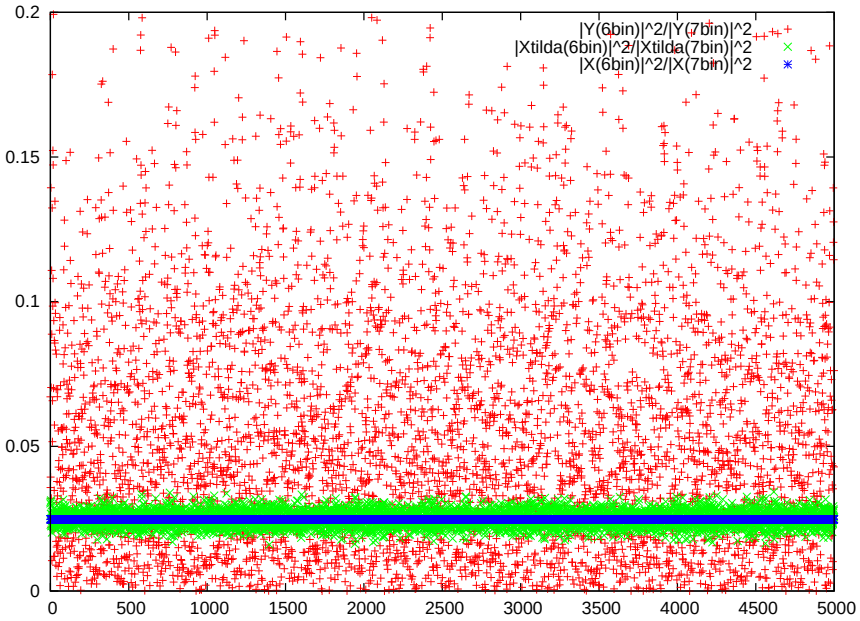

図 21: 位相揃え平均前後における振幅比の変化 (正弦波周波数 250Hz、6 ビンと 7 ビンの比)

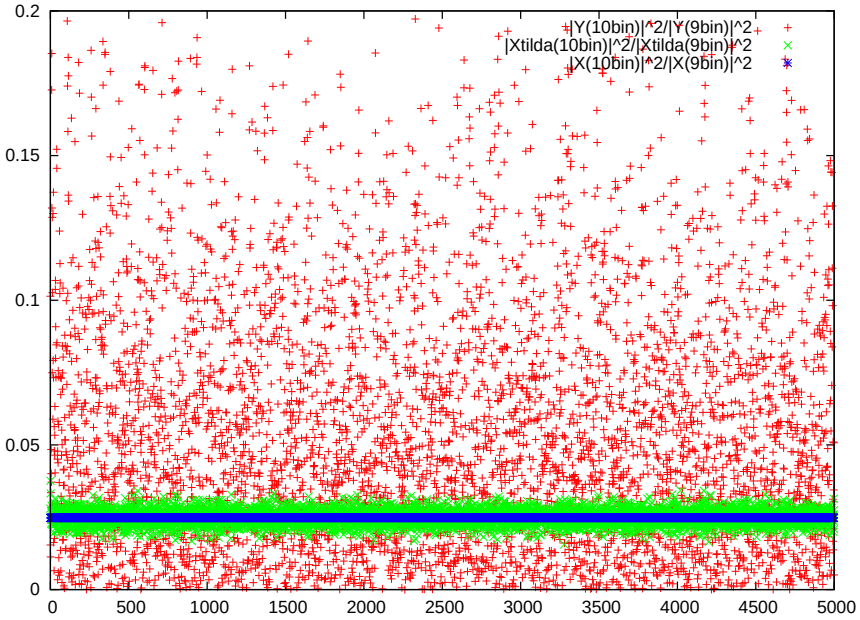

図 22: 位相揃え平均前後における振幅比の変化 (正弦波周波数 250Hz、9 ビンと 10 ビンの比)

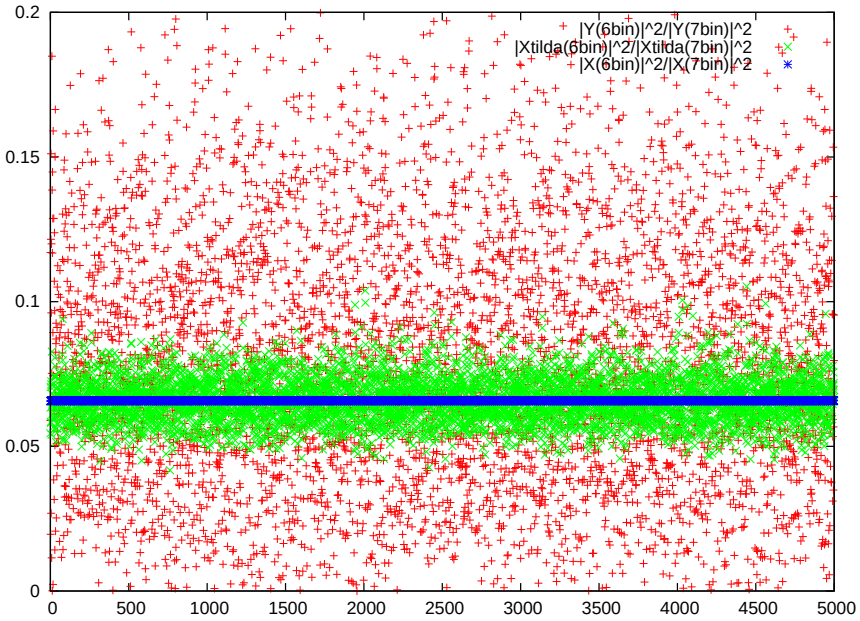

図 23: 位相揃え平均前後における振幅比の変化 (正弦波周波数 240Hz、6 ビンと 7 ビンの比)

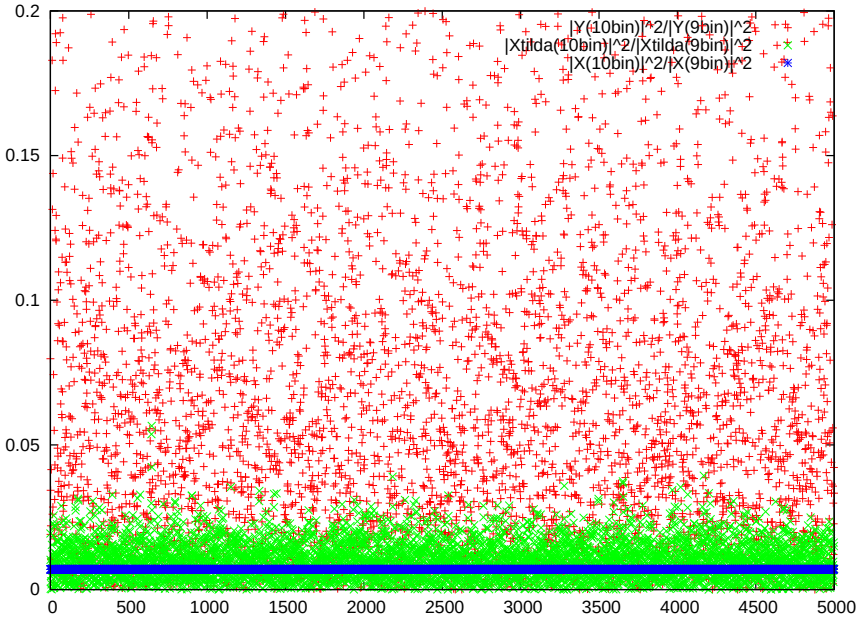

図 24: 位相揃え平均前後における振幅比の変化 (正弦波周波数 240Hz、9 ビンと 10 ビンの比)

7　調波周波数抽出実験

提案方法がどの程度調波周波数の抽出を行なえるのか、様々な入力信号に対して実験を行なった。

7.1　実験条件

入力信号には正弦波（250Hz,240Hz)、三角波 charp 信号 (基本周波数は一秒で 200Hz から 300Hz に変化)、男女話者一名の音声ファイル 1 つずつを用いた。正弦波、三角波 charp 信号それぞれの音声ファイルの先頭 128ms を無音とし、その後 1s それぞれの信号が存在するファイルを作成した。これは、提案方法で用いる雑音スペクトルの分散 $\sigma_V^2(t,k)$ の推定をこの先頭フレームで行なうためである。また、事前に音声ファイルの先頭 128ms に音声情報が入っていないのを確認した。

また、これらの入力信号に自作プログラムで作成した白色雑音を 0dB で重畳し、これも同様に調波周波数抽出実験を行なった。

その他の実験条件は以下のとおりである。

表 6: 実験条件

入力信号	正弦波 (250Hz,240Hz)、 三角波の charp 信号 (1 秒で 200Hz から 300Hz への変化)、 ATR 研究所日本語音声データベースセット A より 男女話者各一名一発話 (「あいて」)
使用雑音	白色雑音
SNR	0,∞dB
サンプリング周波数	16kHz
分析窓	Minimum 3-term 窓
フレーム長	512 点
フレーム周期	64 点
位相揃え平均のフレーム数	8 フレーム

7.2　結果と考察

図 25 は 250Hz の正弦波の推定周波数、図 26 は白色雑音を 0dB で重畳したときの推定基本周波数である。どちらも周波数の推定が精度良く行なわれていることが確認できる。

図 27 は 240Hz の正弦波の推定周波数、図 28 は白色雑音を 0dB で重畳したときの推定周波数である。正弦波の周波数がビンの中心周波数とずれている場合でも周波数推定が行なわれていることが確認できる。

図 29 はの三角波の charp 信号の推定調波周波数、図 30 は白色雑音を 0dB で重畳したときの推定調波周波数である。調波周波数がビンとビンの間の周波数にあるときの推定が上手く行なわれていないが、それ以外では調波周波数の推定が上手く行なわれていることが確認できる。雑音が重畳された場合、基本周波数はとれても高調波の推定が上手く行なわれていないことも分かった。

図 31 は女性話者の発話音声の推定基本周波数、図 32 は白色雑音を 0dB で重畳したときの推定基本周波数である。雑音が重畳されているときのほうが音声の周波数の特徴をとらえている結果となった。緑色の結果は、SPTK(音声信号処理ツールキット)[8] による結果である。雑音が乗っていないときは SPTK のほうが推定精度が高いが、雑音が重畳されているときは SPTK では推定出来なかった基本周波数の後半部分が推定できていることが分かった。

図 33 は女性話者の発話音声の推定調波周波数、図 32 は白色雑音を 0dB で重畳したときの推定調波周波数である。緑色の結果は、SPTK(音声信号処理ツールキット)[8] による結果である。どちらの場合も、高調波の特徴をとらえていることが分かる。

図 35 は男性話者の発話音声の推定基本周波数、図 36 は白色雑音を 0dB で重畳したときの推定基本周波数である。緑色の結果は、SPTK(音声信号処理ツールキット)[8] による結果である。SPTK の結果と比べ、雑音の有無に関わらず提案方法では基本周波数よりも高い調波周波数を基本周波数として推定してしまっている。これは、男性話者の基本周波数が 100Hz 前後と低いため、二番目に高い周波数も基本周波数の候補に入ってしまったために誤推定が起こったと考えられる。

図 37 は男性話者の発話音声の推定調波周波数、図 38 は白色雑音を 0dB で重畳したときの推定調波周波数である。緑色の結果は、SPTK(音声信号処理ツールキット)[8] による結果である。基本周波数の推定では上手く推定が行なえたとは言えないが、調波周波数全体を見ると特徴を捉えていることが分かった。また、雑音を重畳すると推定精度が下がってしまうことも分かった。

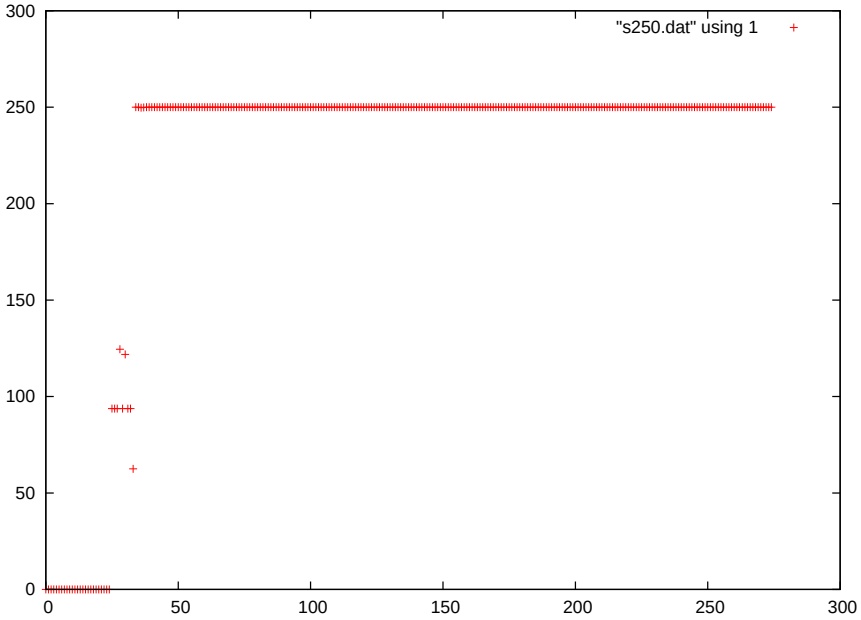

図 25: 250Hz の正弦波の推定周波数

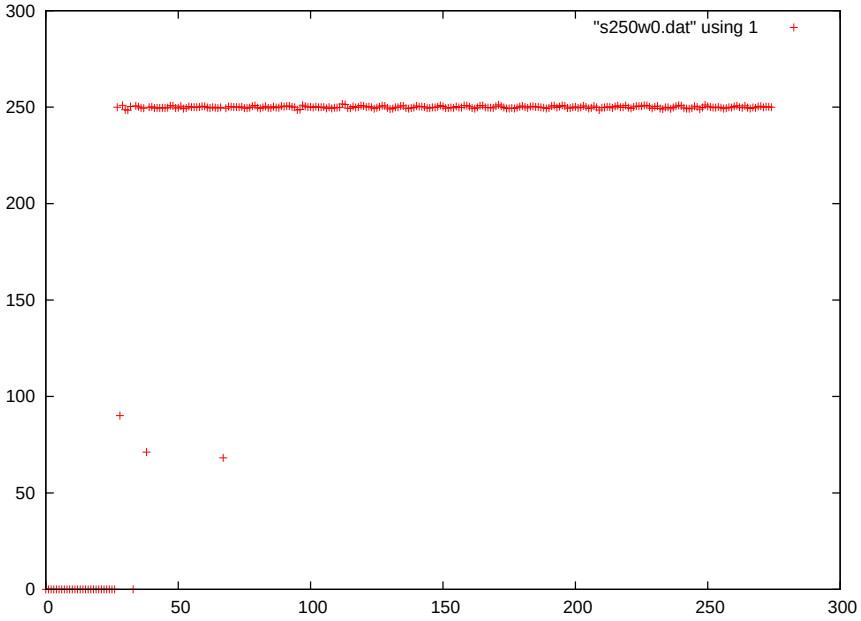

図 26: 250Hz の正弦波 (白色雑音 0dB 重畳) の推定周波数

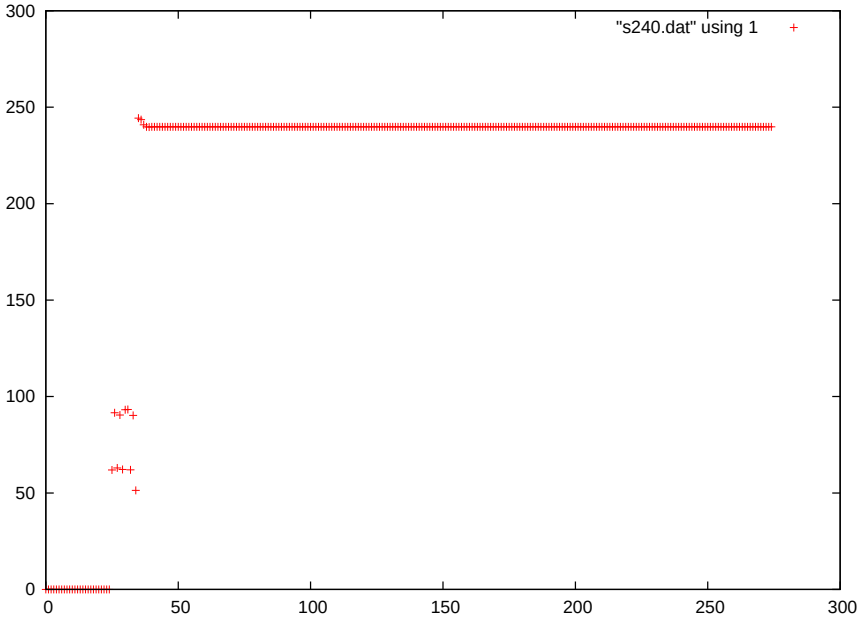

図 27: 240Hz の正弦波の推定周波数

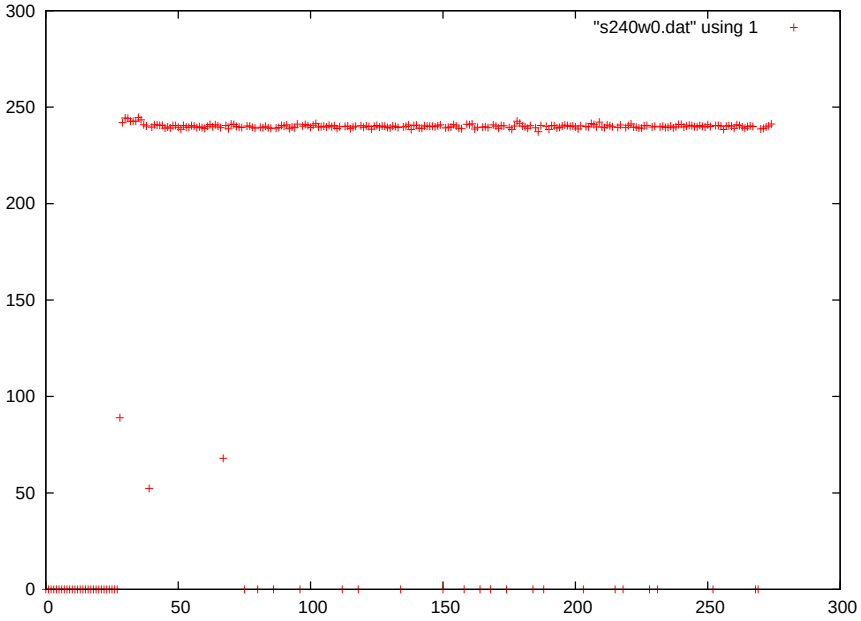

図 28: 240Hz の正弦波 (白色雑音 0dB 重畳) の推定周波数

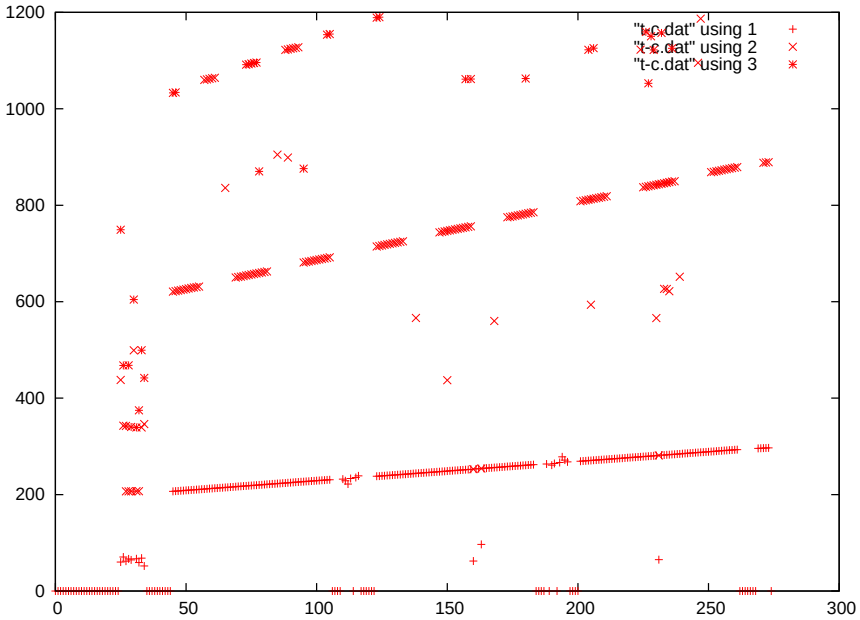

図 29: 三角波 charp 信号の推定周波数

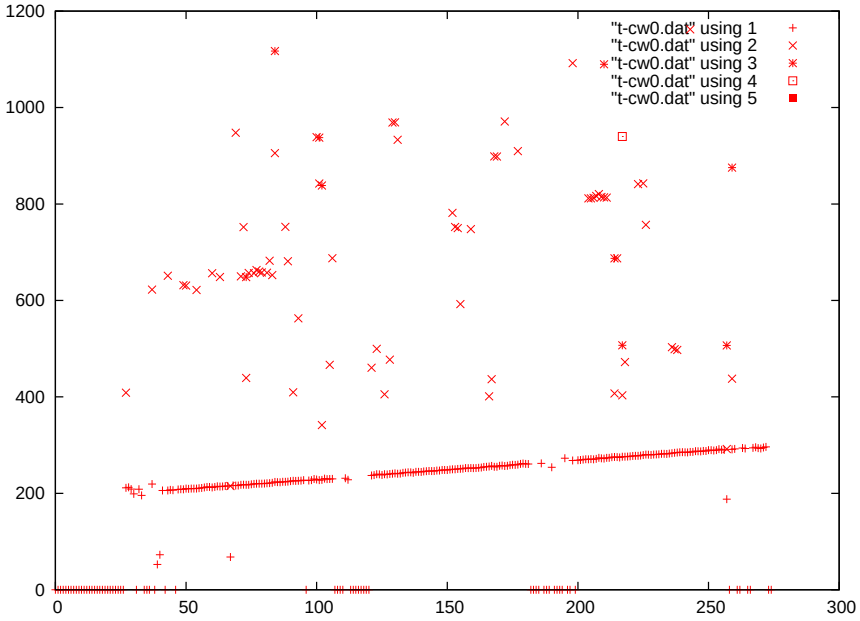

図 30: 三角波 charp 信号 (白色雑音 0dB 重畳) の推定周波数

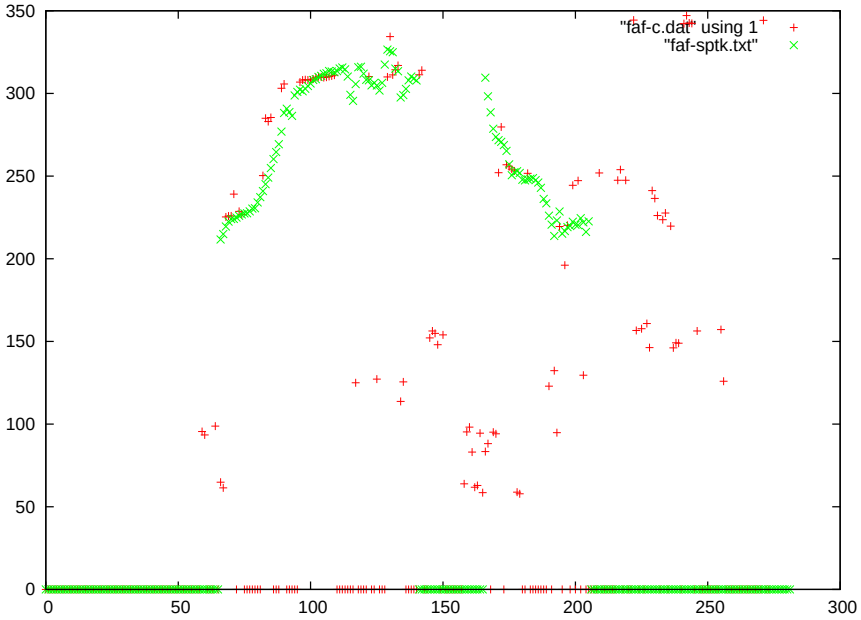

図 31: 女性話者の音声の推定基本周波数

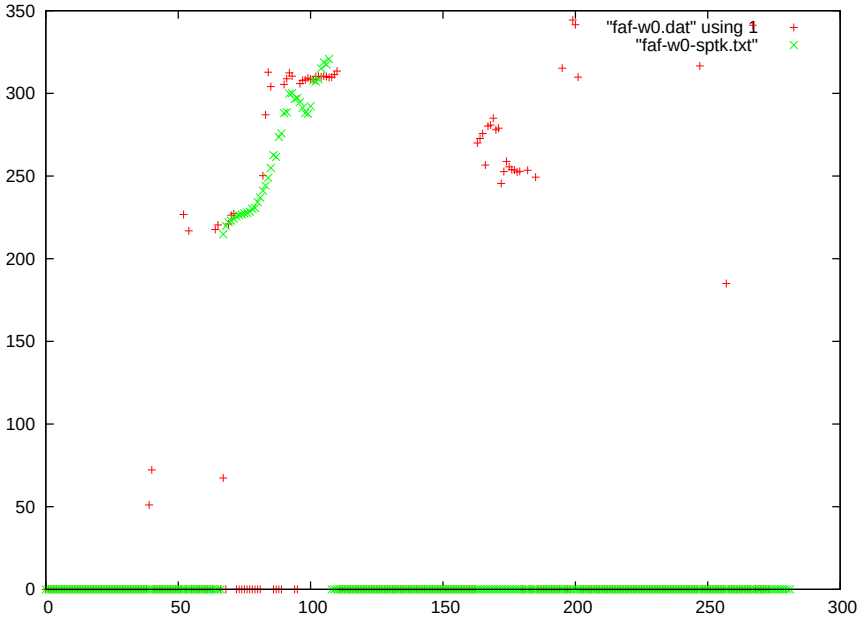

図 32: 女性話者の音声 (白色雑音 0dB 重畳) の推定基本周波数

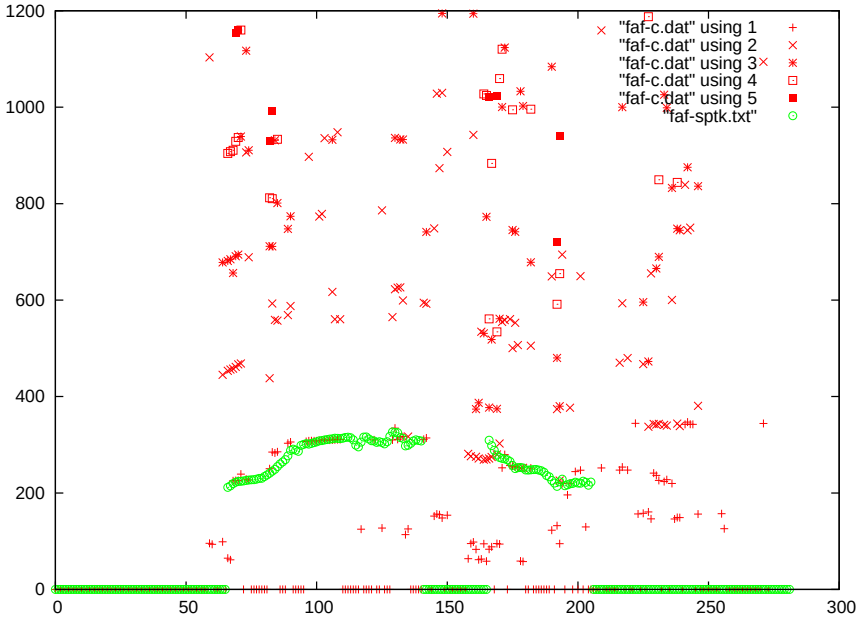

図 33: 女性話者の音声の推定周波数

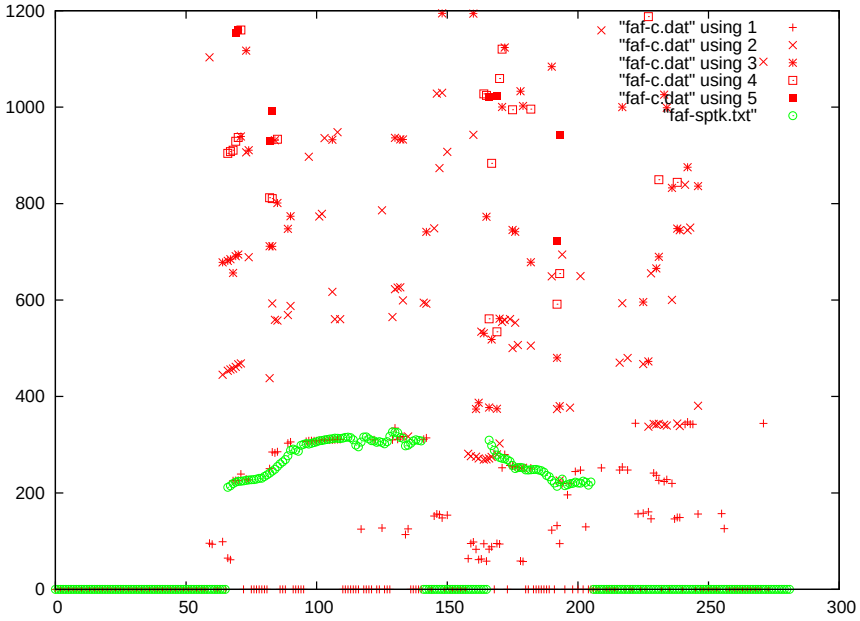

図 34: 女性話者の音声 (白色雑音 0dB 重畳) の推定周波数

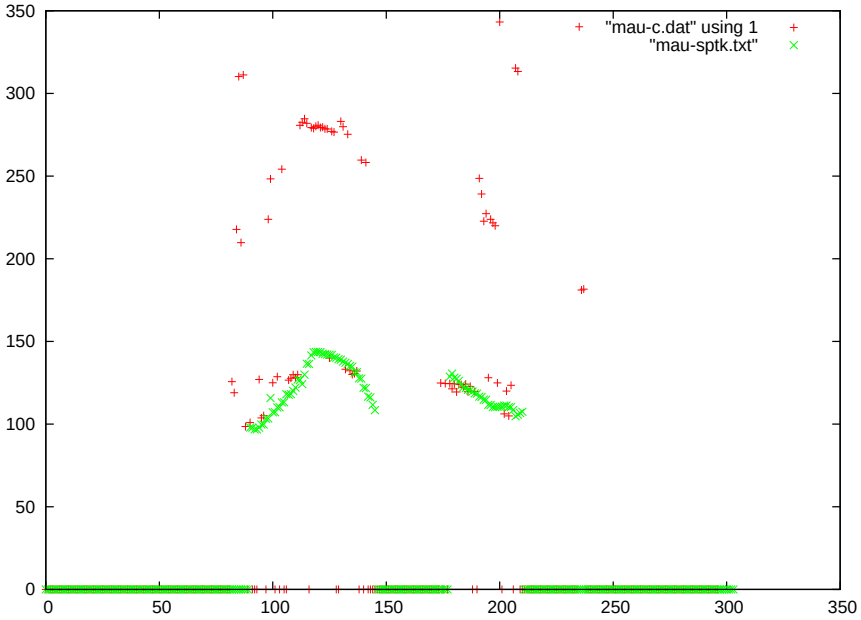

図 35: 男性話者の音声の推定基本周波数

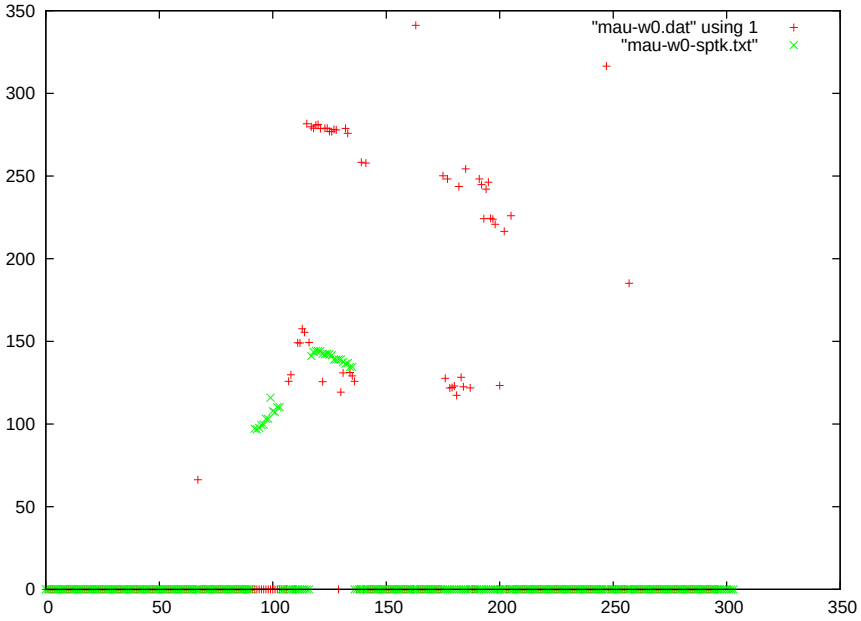

図 36: 男性話者の音声 (白色雑音 0dB 重畳) の推定基本周波数

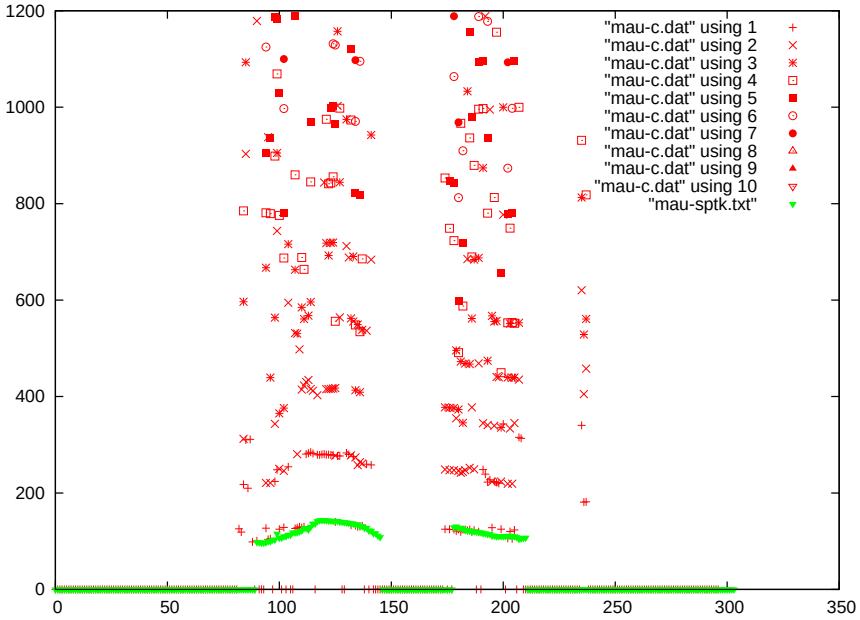

図 37: 男性話者の音声の推定周波数

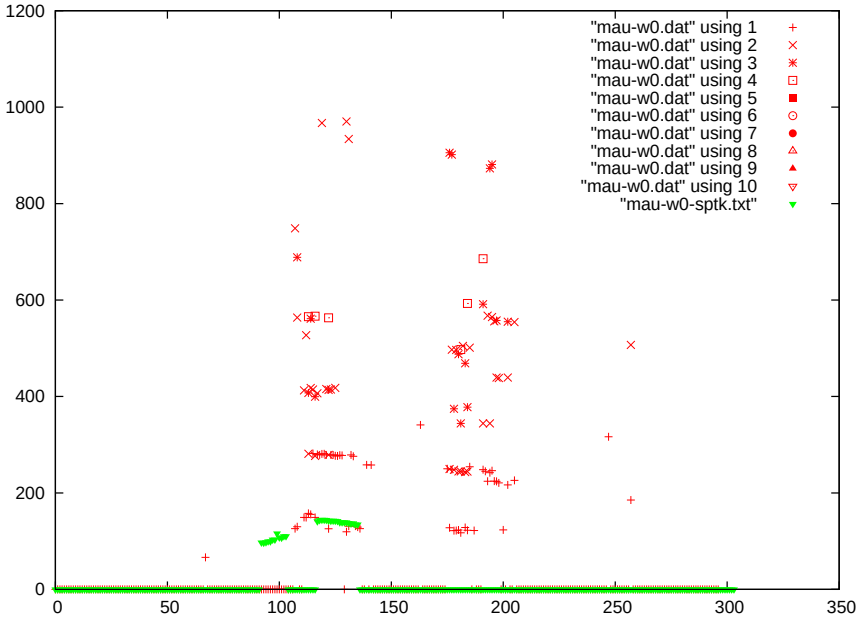

図 38: 男性話者の音声 (白色雑音 0dB 重畳) の推定周波数

8　評価実験

提案手法の雑音低減の有効性を調べるために、先行研究である Wiener フィルタ [2] と SD 法 [4] との性能比核実験を行う。

8.1　実験条件

評価用信号として、音声は ATR 研究所日本語音声データベースセット A の音声を用いた。雑音は、NTT アドバンステクノロジ社の環境雑音データベースより実環境雑音 4 種と、自作プログラムで生成した白色雑音を用いた。サンプリング周波数はすべて 16kHz で統一した。SD 法による推定に必要な基本周波数の推定は SPTK(音声信号処理ツールキット)[8] を、その他の周波数は基本周波数の整数倍を用いた。また、日本語音声の有声音の存在確率 ζ は事前実験で $\zeta = 0.668601$ とした。

表 7: 実験条件

音声	ATR 研究所日本語音声データベースセット A 計 281 単語 男女各三名 (faf,ffs,fym,mau,mht,mtk)
使用雑音	NTT-AT 社環境騒音データベース [9] 実環境雑音 4 種 (空港雑音、ロビー雑音、オフィス雑音、レストラン雑音) 白色雑音
SNR	-10,-5,0,5,10,15dB
サンプリング周波数	16kHz
分析窓	ハミング窓（Wiener フィルタ、SD 法)、Minimum 3-term 窓 (提案法)
フレーム長	512 点
フレーム周期	256 点（Wiener フィルタ、SD 法)、64 点 (提案法)
位相揃え平均のフレーム数	2 フレーム (Wiener フィルタ、SD 法)、8 フレーム (提案法)

8.2　評価方法

雑音低減の指標には、セグメンタル SNR 改善値と対数スペクトル歪みの二つの尺度を用いた。

8.2.1 セグメンタル SNR 改善値

セグメンタル SNR[3](以下、SegSNR) とは、音声ファイル全体を短時間フレームに分けて SNR を計算し、その dB 値の平均をとったもので、以下のように定義できる。

$$SegSNR = \left[10 \log_{10} \frac{\sum_{n=tM-N/2}^{tM+N/2} x^2(n)}{\sum_{n=tM-N/2}^{tM+N/2} [x(n) - z(n)]^2} \right] \tag{78}$$

それぞれ M はフレーム周期、N はフレーム長、T は窓掛けによって生成したセグメントの数である。

フィルタをかけた後のセグメンタル SNR からかける前のセグメンタル SNR を引いたものをセグメンタル SNR 改善値とする。このセグメンタル SNR 改善値を用いて、どれだけ雑音が除去されたかを評価する。数値が高ければ高いほど雑音が低減されたことになる。今回の実験では $M = 240$(15.0ms),$N = 480$(30.0ms) で評価を行った。

8.2.2 対数スペクトル歪み

対数スペクトル歪み [3](以下、LSD) は、推定音声の対数パワースペクトルと音声信号の対数パワースペクトルとの差を用いた評価法で、以下の式で定義される。

$$LSD = \frac{1}{T} \sum_{t=0}^{T-1} \left[\frac{2}{N} \sum_{k=1}^{N/2} (f(X_{tk}) - f(\hat{X}_{tk}))^2 \right]^{\frac{1}{2}} \tag{79}$$

$$f(X_{tk}) = \max[20 \log_{10} |X_{tk}|, \delta] \tag{80}$$

$$\delta = \max[20 \log_{10} |X_{tk}| - 50] \tag{81}$$

LSD は、値が小さいほど音声信号に近づくため、良い結果といえる。

8.3 結果と考察

図 39 が SegSNR 改善値の結果、図 40 が LSD の結果である。

どの場合においても Wiener フィルタと比べて SD 法・提案方法共に良い結果となった。

SD 法と提案方法で比較すると、0dB 以下の雑音の割合が強い観測信号やレストラン雑音のような非定常の雑音に対して高い性能を発揮していることが結果からわかる。そこで、推定周波数を確認したところ、SD 法で用いた SPTK による基本周波数推定は 0dB 以下の SNR の観測信号に対してほとんど 0 となっていた。周波数が推定できないフレー

ムは統計的推定法のみを用いることになる。そのため、SD 法は性能を発揮しきれず、提案法の方が良い結果となったと考えられる。

一方で、10dB などの元々雑音成分の少ない観測信号による推定結果は SD 法よりも悪い性能であった。これは、提案法の周波数推定の絞り込みが甘く、うまく推定ができていないことが原因と思われる。

バス雑音を重畳した観測信号からの推定では、全体的に SD 法の方が提案法よりも良い結果となった。これは、バス雑音が低周波数帯にパワーを持つ雑音であるために、本来の音声の基本周波数よりも低い周波数も音声の周波数であると推定されてしまったのが原因と考えられる。この周波数推定の精度向上は、今後の課題である。

数値評価での性能の改善は見られたが、いくつかの音声について実際に聞いて確認したところ、0dB 以下の低 SNR 環境下では音声として聞こえない場合もみられた。これらの音声が対象であっても言葉として聞こえるような音声推定を行なうことも、今後の課題である。

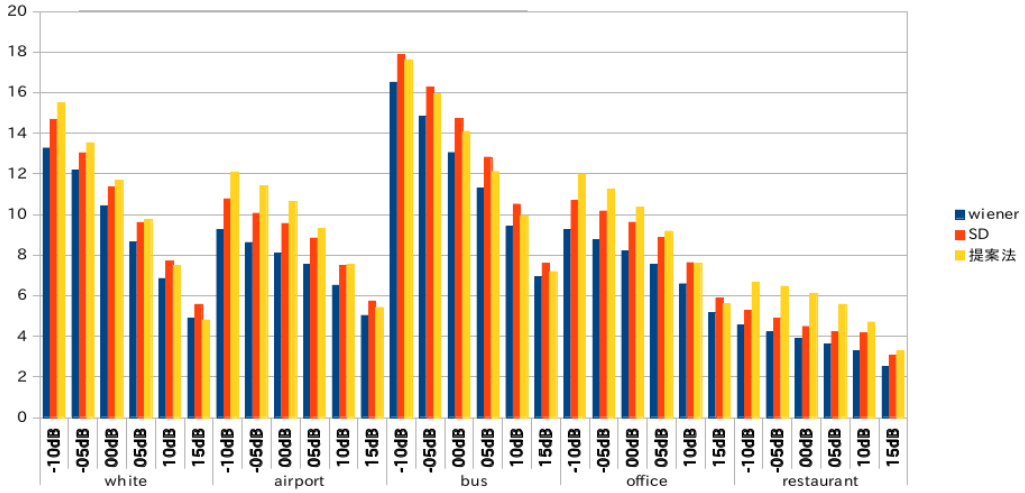

図 39: 実験結果:SegSNR

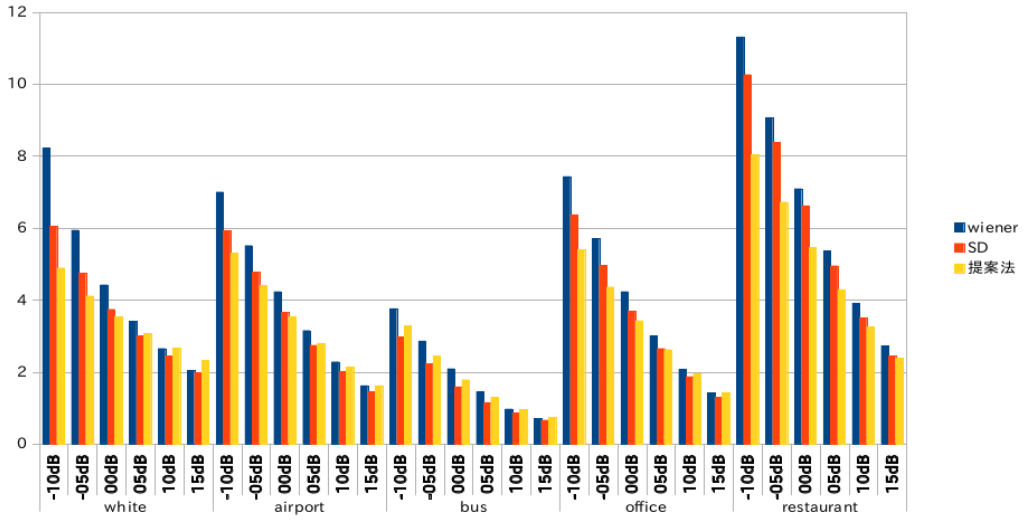

図 40: 実験結果:LSD

9　おわりに

本研究では観測信号スペクトルからの音声信号スペクトルの推定を目的とし、窓関数の特性を用いて観測信号から音声を再構成する方法と統計的モデルを組み合わせた音声スペクトル推定システムを提案した。そして、従来法である SD 法と SegSNR 改善値と LSD による比較実験を行った。その結果、雑音が音声の振幅よりも大きい環境下での音声スペクトル推定に対し高い性能が得られた。

今後の課題としては、窓関数の特性を用いた音声スペクトル推定システムの中で行われる周波数推定の精度向上が挙げられる。

10　謝辞

研究を進めるにあたり、指導やアドバイスをいただきました西野哲朗教授、吉田利信教授、高木一幸助教に心より感謝致します。

参考文献

[1] Boll. S: “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans., VOL.ASSP-27, NO.2, 1979.

[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Transactions on audio, sppech, and language processing, vol.ASSP-32, no.6, pp.1109-1121,1984.

[3] Jacob Benesty, M.M.Sondhi, Yiteng Huang(Eds.): “Springer Handbook of Speech Processing, ” Springer, 2008.

[4] R.Hendriks, R. Heusdens, and J.Jensen, “An mmse estimator for speech enhancement under a combined stochastic-deterministic speech model,” IEEE Transactions on audio, speech, and language processing, vol.15, no.2, pp.406-415, 2007.

[5] C.W.Therrien, “Discrete Random Signals and Statistical Signal Processing” Englewood Cliffs, NJ: Prentice-Hall,1992.

[6] A. Nuttall, “Some windows with very good sidelobe behavior,” IEEE Transactions on acoustics, speech, and signal processiong, vol. ASSP-29, no.1, pp.84-91, 1981.

[7] 吉田利信, “信号中に含まれる正弦成分抽出装置、正弦成分抽出方法及びプログラム” 特許出願予定 2014 年 2 月 5 日.

[8] SOURCEFORGE.NET “Speech Signal Processing Toolkit (SPTK)” [http://sptk.sourceforge.net/], アクセス (2014/1/30)

[9] NTT Advanced Technology Corporation : “Ambient Noise Database for Telephonometry 1996,” 1996.

[10] 古井貞熙, “新音響・音声工学, ” 近代科学社, 2006.

A　付録 A Overlap-add 法について

推定音声スペクトルを式 (3) で逆フーリエ変換して再構成を行なう際に Overlap-add 法が用いられる。

フレーム長 N、フレームシフト幅 L、フレーム番号 i、時刻 $\tau(0 \leq \tau \leq L)$ のときの推定音声は次のように得られる。

$$\hat{x}(iL+\tau) = \sum_{m=0}^{N/L-1} \hat{x}(i-m, mL+\tau)w(mL+\tau) \tag{82}$$

B　付録 B ウィナーフィルタの導出

推定音声スペクトル $\hat{X}(t,k)$ を観測信号スペクトル $Y(t,k)$ とフィルタ係数 $H(t,k)$ の積で求められるものとする。

$$\hat{X}(t,k) = H(t,k)Y(t,k) \tag{83}$$

このとき、$\hat{X}(t,k)$ と $X(t,k)$ の平均二乗誤差 $J[H(t,k)]$ は次のように表される。

$$J[H(t,k)] = E[|X(t,k) - H(t,k)Y(t,k)|^2] \tag{84}$$

$J[H(t,k)]$ が最小になるように $H(t,k)$ を決める。$J[H(t,k)]$ を $H^*(t,k)$ について微分を行うと、次のようになる。

$$\begin{aligned}
\frac{dJ[H(t,k)]}{dH^*(t,k)} =& \frac{d}{dH^*(t,k)} E[|X(t,k) - H(t,k)Y(t,k)|^2] && (85)\\
=& \frac{d}{dH^*(t,k)} E[(X^* - H^*(t,k)Y^*(t,k))(X(t,k) - H(t,k)Y(t,k))] && (86)\\
=& \frac{d}{dH^*(t,k)} E[X^*(X(t,k) - H(t,k)Y(t,k))] \\
& - \frac{d}{dH^*(t,k)} E[H^*(t,k)Y^*(t,k)(X(t,k) - H(t,k)Y(t,k))] && (87)\\
=& E[\frac{d}{dH^*(t,k)} H^*(t,k)Y^*(t,k)(X(t,k) - H(t,k)Y(t,k))] && (88)\\
=& - E[Y^*(t,k)(X(t,k) - H(t,k)Y(t,k))] && (89)
\end{aligned}$$

ここで、X^* は X の複素共役を表す。$\frac{dJ[H(t,k)]}{dH(t,k)}$ が 0 になるとき $J[H(t,k)]$ が最小になる。

$$-E[Y^*(t,k)(X(t,k)-H(t,k)Y(t,k))]=0 \tag{90}$$

$$E[Y^*(t,k)X(t,k)]-H(t,k)E[Y^*(t,k)Y(t,k)]=0 \tag{91}$$

$$E[(X^*(t,k)+V^*(t,k))]X(t,k)]-H(t,k)E[Y^*(t,k)Y(t,k)]=0 \tag{92}$$

$$\tag{93}$$

$X(t,k)$ と $V(t,k)$ は互いに無相関でそれぞれ平均 0 であるので、

$$E[(X^*(t,k)X(t,k)]+E[V^*(t,k)]E[X(t,k)]-H(t,k)E[Y^*(t,k)Y(t,k)]=0 \tag{94}$$

$$E[|X(t,k)|^2]-H(t,k)E[|Y(t,k)|^2]=0 \tag{95}$$

$$H(t,k)=\frac{E[|X(t,k)|^2]}{E[|Y(t,k)|^2]} \tag{96}$$

となる。

C 付録C Minimum 3-term窓のスペクトルの導出

Minimum 3-term窓は次のように定義される。

$$
\begin{aligned}
w(t) =& 1 + \frac{a_1}{a_0}\cos\left(\frac{2\pi}{N}t\right) + \frac{a_2}{a_0}\cos\left(\frac{4\pi}{N}t\right) \qquad (97)\\
& |t| < \frac{N}{2}\\
& a_1 = 0.4973406, a_2 = 0.0782793, a_0 = 1 - a_1 - a_2
\end{aligned}
$$

この短時間フーリエ変換は次のようになる。

$$
\begin{aligned}
W(\xi) &= \int_{-\frac{N}{2}}^{\frac{N}{2}} w(\tau) e^{j2\pi\xi\frac{\tau}{N}} d\tau \qquad (98)\\
&= \int_{-\frac{N}{2}}^{\frac{N}{2}} e^{j2\pi\xi\frac{\tau}{N}} \times \\
&\quad \left(1 + \frac{a_1}{a_0}\cos\left(\frac{2\pi}{N}\tau\right) + \frac{a_2}{a_0}\cos\left(\frac{4\pi}{N}\tau\right)\right) d\tau \qquad (99)\\
&= \frac{1}{2}\left[\frac{2}{j2\pi(\frac{1}{N}\xi)} e^{j2\pi\frac{\xi}{N}\tau}\right.\\
&\quad + \frac{a_1}{a_0}\frac{1}{j2\pi\frac{1}{N}(\xi-1)} e^{j2\pi\frac{1}{N}(\xi-1)\tau} + \frac{a_1}{a_0}\frac{1}{j2\pi\frac{1}{N}(\xi+1)} e^{j2\pi\frac{1}{N}(\xi+1)\tau}\\
&\quad \left. + \frac{a_2}{a_0}\frac{1}{j2\pi\frac{1}{N}(\xi-2)} e^{j2\pi\frac{1}{N}(\xi-2)\tau} + \frac{a_2}{a_0}\frac{1}{j2\pi\frac{1}{N}(\xi+2)} e^{j2\pi\frac{1}{N}(\xi+2)\tau}\right]_{-N/2}^{N/2} \qquad (100)\\
&= \frac{1}{2}\left[2\frac{\sin(\pi\xi)}{\pi\frac{1}{N}\xi} + \frac{a_1}{a_0}\frac{\sin(\pi(\xi-1))}{\pi\frac{1}{N}(\xi-1)} + \frac{a_1}{a_0}\frac{\sin(\pi(\xi+1))}{\pi\frac{1}{N}(\xi+1)}\right.\\
&\quad \left. + \frac{a_2}{a_0}\frac{\sin(\pi(\xi-2))}{\pi\frac{1}{N}(\xi-2)} + \frac{a_2}{a_0}\frac{\sin(\pi(\xi+2))}{\pi\frac{1}{N}(\xi+2)}\right] \qquad (101)\\
&= \frac{1}{2}\left[2\frac{\sin(\pi\xi)}{\pi\frac{1}{N}\xi} - \frac{a_1}{a_0}\frac{\sin(\pi\xi)}{\pi\frac{1}{N}(\xi-1)} - \frac{a_1}{a_0}\frac{\sin(\pi\xi)}{\pi\frac{1}{N}(\xi+1)}\right.\\
&\quad \left. + \frac{a_2}{a_0}\frac{\sin(\pi\xi)}{\pi\frac{1}{N}(\xi-2)} + \frac{a_2}{a_0}\frac{\sin(\pi\xi)}{\pi\frac{1}{N}(\xi+2)}\right] \qquad (102)\\
&= \frac{1}{2}\frac{\sin(\pi\xi)}{\pi\frac{1}{N}\xi}\left[2 - \frac{a_1}{a_0}\frac{\frac{1}{N}\xi}{\frac{1}{N}\xi - \frac{1}{N}} - \frac{a_1}{a_0}\frac{\frac{1}{N}\xi}{\frac{1}{N}\xi + \frac{1}{N}}\right.\\
&\quad \left. + \frac{a_2}{a_0}\frac{\frac{1}{N}\xi}{\frac{1}{N}\xi - \frac{2}{N}} + \frac{a_2}{a_0}\frac{\frac{1}{N}\xi}{\frac{1}{N}\xi + \frac{2}{N}}\right] \qquad (103)\\
&= \frac{N\sin(\pi\xi)}{\pi\xi}\left[1 - \frac{a_1}{a_0}\frac{\xi^2}{\xi^2 - 1^2} + \frac{a_2}{a_0}\frac{\xi^2}{\xi^2 - 2^2}\right] \qquad (104)
\end{aligned}
$$