社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（17）
2004／12／21

音声認識率改善のための波形減算とスペクトル減算の併用による反射音除去法

大田　健紘†　　柳田　益造†

†同志社大学工学部　〒610-0394 京都府京田辺市多々羅都谷 1-3
E-mail: †dtd0736@mail4.doshisha.ac.jp　†myanagid@mail.doshisha.ac.jp

あらまし　本稿では，反射音を含む音声から反射音を除去し音声認識率を向上させる手法について報告する．提案法では，反射音は遅延時間と減衰率という２つのパラメータで記述できると仮定している．これら２つのパラメータは，複数マイクロフォンで受音した信号の自己相関関数を用いて推定している．そして，遅延波形の減算には，反射音を含む音声の音声・無音区間や摩擦音・撥音の識別を行い，本来無音であったと推定できる区間に対してはスペクトル減算を用いることで，反射音成分を十分に除去している．提案法は，部屋の特性や目的音声に対する事前知識を必要とすることなく処理を実行できる．提案法を用いることで，ライン入力で認識率 100%の音声が認識率約 80%に低下する無雑音環境において，反射音を含む音声の認識率が約 8%向上することを確認した．
キーワード　反射音，遅延時間，減衰率，自己相関関数，スペクトル減算

Removing Reflected Waves Using Delayed Wave Subtraction Combined with Spectral Subtraction

Kenko Ohta†　　Masuzo Yanagida†

†Faculty of Engineering, Doshisha University　1-3, Tatara-Miyakodani, Kyo-Tanabe, Kyoto, 610-0394, Japan
E-mail: †dtd0736@mail4.doshisha.ac.jp　†myanagid@mail.doshisha.ac.jp

Abstract　Proposed is a method of removing reflected waves from a mixed wave consisting of a source signal and reflected waves. The method is a kind of waveform subtraction referring to auto-correlation functions of multi channel speech signals assuming that a reflected wave has two parameters; path amplitude and delay time. The method estimates these parameters based on auto-correlation functions of signals received by microphones. The estimated delayed wave derived from the received wave is subtracted from the received wave using an estimated delay only for vocalic segments and spectral subtraction is applied to non-speech segments. The proposed method can be realized without prior knowledge about room characteristics or the target speech. Speech recognition rate for the signals picked up with 3 microphones in a reverberant environment is improved about 8% employing the proposed method.
Keyword　Reflected Wave, Delay Time, Amplitude, Auto-correlation Function, Spectral Subtraction

1. はじめに

現在音声認識システムは高精度化し，理想的な環境において接話マイクを用いて認識を行えば，実用化レベルに達している．しかしながら，リビングルームや会議室などの実環境においては周囲からの雑音や，壁からの反射音の影響により大幅に音声認識率が低下する．このことから，音声認識システムの実用化には，雑音や壁からの反射音などに対する頑健さが求められる．頑健な音声認識システムを構築するためのアプローチは，大きく分けて２つある．一つはマイクロフォンで受音した音声から雑音や反射音を除去する手法，そしてもう一つは音声認識のための音響モデルを雑音や反射音を含むデータを用いて再学習し，雑音や反射音に適応させる手法である．本稿では前者に関して，壁からの反射音を除去する一手法を提案する．

反射音除去法に関しては現在までにさまざまな研究が行われている[1-14].

古くは室内のインパルス応答を測定し，その逆特性を畳み込むことにより反射音の除去を行っていた[1]．しかし，この手法ではスピーカからマイクロフォンまでの伝達特性は不変なければならず，室温の変化や人の動きなどの要因により伝達特性が変化するので再測定しなければならない．そのため実環境への応用は難しいと考えられている．

そこで現在では，上記の問題点を解決するために，伝達特性の測定を必要としない方法が考えられている．

田沢らの提案する手法[2]は，単一マイクロフォンで受音した音声信号のケプストラムを計算し，直接音に対する反射音の時間遅れと振幅比を推定している．この手法は反射音の比較的少ない状況ではうまくいくが，反射音が増加するにつれて計算量が膨大になり，また反射音の除去も困難になるという問題点が挙げられている．

雑音除去の手法としてしばしば用いられているが，反射音

除去の手法としてはあまり検討されていない，スペクトルサブトラクション（SS 法）を反射音除去に応用する研究も行われている．SS 法を反射音除去に応用した研究は，馬場ら[3]や中島ら[4]によって行われている．馬場らの手法では，単一マイクロフォンによる除去処理が行われている．この手法は，部屋の伝達関数を何度も測定する必要はないが，反射音の振幅を推定するために 1 度は部屋の伝達関数を測定する必要がある．中島らの手法は，反射音の減衰をあるモデルに当てはめ，残響付加音声のスペクトルから直接音の成分と反射音の成分を推定し反射音の成分をスペクトル上で減算している．この手法は，反射音の減衰特性を減衰時間に依存するモデルで近似しているため，残響時間が既知でなければならない．

さらに異なるアプローチとして MTF(Modulation Transfer Function)に基づいた手法も提案されている[5]．この手法は主にパワーエンベロープの回復を目的としている．

単一マイクロフォンで受音された信号の調波構造を用いて，逆フィルタを学習する手法も提案されている[6]．この手法の場合，効果的な反射音除去を行うためには正確な F0 抽出が必要となる．さらには，逆フィルタ学習のために，膨大な量の学習を必要とする．

以上にあげた手法は，入力信号に対して信号処理を施し，反射音を除去しクリーンな音声信号に近づけ音声認識率の改善を図るものである．一方で，入力信号に処理を施すのではなく，音声認識のための音響モデルに処理を施し，反射音の影響を考慮した音響モデルに作り変えるという手法も考えられている[7-9]．しかし，このような適応処理では，ある一定以上に残響時間が長くなると極端に認識率の改善効果が低下することが報告されている[6]．また，話者とマイクロフォンの相対的な位置は常に変わるので，状況に応じて複数のモデルを用意する必要がある．

これらのことから，実環境での利用を考えると，入力信号を処理し反射音に関する情報を推定し波形レベルもしくはスペクトルレベルで反射音を除去する手法が有利であると考えられる．そこで我々は，入力信号を音声の特徴によって分割し，その区間ごとに波形レベルでの減算とスペクトルレベルでの減算を切り替えて用いる手法を提案した[10]．しかし，この手法では音声区間の検出などで用いる閾値の設定はすべて手動で，一定の値を割り当てていたという問題点があった．しかし実際は，たとえ同一話者による発話であっても，発話内容が異なれば音韻の継続長や振幅が異なるということがあるため，一定の値ではなく各発話に最適な値を自動的に割り当てていく方が望ましい．

本稿では閾値の自動設定を試み，前回提案手法からの認識率改善効果を見る．またデータ量を増やしたことにより新たに明らかになった問題点についても言及し，今後の課題について検討している．

2. 提案法の概要

2.1 反射音除去の原理

本稿では，静かな音場環境を想定しておりノイズ源は考えず，壁や床，天井からの反射音のみがあるものとする．第 i 番目のマイクロフォンに受音された信号 $r_i(t)$は，音源からの出力信号 $s(t)$と，スピーカからマイクロフォンまでの経路に起因するインパルス応答が畳み込まれたものとなっている．第 i 番目のマイクロフォンに受音された信号 $r_i(t)$は，以下の式で表される．

$$r_i(t) = s(t) * h_i(t) \tag{1}$$

ただし，*は畳み込み演算を表している．そして，

$$h_i(t) = \sum_{j=0}^{J} h_{ij}(t) \tag{2}$$

であり，$h_{ij}(t)$は直接音(j=0)を含むスピーカから第 i 番目のマイクロフォンまでの第 j 番目の経路に起因するインパルス応答を表している．壁の反射の周波数特性は平坦であるとすると，第 i 番目のマイクロフォンに受音される信号の第 j 番目の経路に起因する反射音は，直接音と比較して α_{ij} (-1<α_{ij}<1)倍になっており，l_{ij} の遅延時間を持っている．第 i 番目のマイクロフォンに受音される信号の第 j 番目の経路に起因する反射音は，α_{ij} と l_{ij} を用いて $\alpha_{ij}s(t-l_{ij})$ と表されるので，複数の反射をしている音は反射回数の少ないものと比較して大幅に減衰している．音声認識率の低下に関係する反射音は，壁や床，天井での反射回数の少ないものであると考える．上記の仮定を元にして直接音 $\alpha_{i0}s(t-l_{i0})$ を推定する式は

$$\alpha_{i0}s(t-l_{i0}) = r_i(t) - \sum_{j=1}^{J} \alpha_{ij}s(t-l_{ij}) \tag{3}$$

で表せる．ここで J はすべての反射音の数を表している．$s(t-l_{ij})$ は元の信号なので未知であるため，受音した信号によって近似することにする．式(3)を書き換えると次のようになる．

$$\alpha_{i0}s(t-l_{i0}) \cong r_i(t) - \sum_{j=1}^{J} \alpha_{ij}r(t-l_{ij}) \tag{4}$$

離散的に書くと次のようになる．

$$\alpha_{i0}s_{k-l_{i0}} \cong r_{ik} - \sum_{j=1}^{J} \alpha_{ij}r_{i\,k-l_{ij}} \tag{5}$$

これを逐次計算に書き直す．

$$\alpha_{i0}r_{k-l_{i0}}^{(j)} \cong r_{ik}^{(j-1)} - \alpha_{ij}r_{i\,k-l_{ij}}^{(j-1)} \quad j = 1 \cdots J \tag{6}$$

反射音の除去は式(6)にしたがって行われる．

2.2 自己相関関数に基づいた時間遅れの推定方法

実環境において，話者から離れたところに置かれた複数のマイクロフォンで音声を受音すると，壁や床，天井からの反射音も同時に受音される．そして，元の音声と反射音は高い相関を持っているため，各マイクロフォンについて自己相関関数を計算すると，音源からマイクロフォンへの経路上の反射点までの距離に応じた時間遅れで，他のマイクロフォンの同じ時間遅れにおける自己相関関数の値よりも大きくなる．しかし，反射音の影響を受けていないとしても，元の音声信号そのものに相関があるため，自己相関関数は平坦ではなく凸凹がある．これらのことから，反射音の影響を受けた音声は，単純に自己相関関数を見ても，ある時間遅れでの自己相関関数のピークが，元の音声のものであるか，反射音の影響によるものであるのか区別がつかない．

提案法では，複数のマイクロフォンを用いることにより上記の問題を解決している．図 1 に J=3 の場合，すなわち 3 本，このマイクロフォンを使って問題を解決する方法を示す．

第 1 番目のマイクロフォンで受音した r_1 に含まれる，比較的パワーの強い反射音を取り除くことを考える．そのため，まず第 2 番目のマイクロフォンと第 3 番目のマイクロフォンの自己相関関数の平均値 $\overline{R}_1$ を計算する．ただし，添え字は平

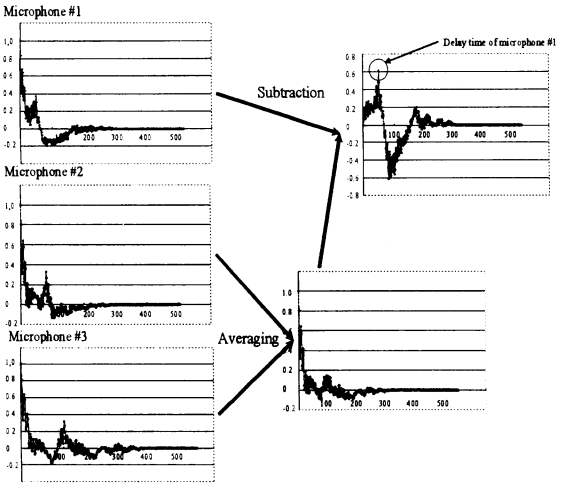


図1　頑健な時間遅れの推定法

均自己相関関数の計算に用いなかったマイクロフォン番号を表している．ここで$\bar{R}_1$は，元の音声信号の自己相関関数R_Sの近似になっているとみなしている．その理由は，異なる位置に配置した複数のマイクロフォンで受音した信号の自己相関関数を平均化することにより，空間の伝達特性を平均化することができ，反射音による影響が少なくなると考えることができるためである．$\bar{R}_1$は後の処理で，第j番目の経路の反射音の減衰率α_{1j}の推定でも用いる．

次に，r_1の自己相関関数R_1と$\bar{R}_1$の差を計算する．反射による時間遅れは，マイクロフォンと壁の相対的な位置によって決定される．そのため，$R_1-\bar{R}_1$が大きくなる時間遅れを，第j番目の経路の反射音と，直接音のマイクロフォンへの到達時間差であると考えてよい．以上のことから第1番目のマイクロフォンに受音される第j番目の反射音の時間遅れl_{1j}は，$R_1-\bar{R}_1$の正の最大値を見つけることによって推定できる．

2.3　減衰率α_{ij}の推定方法

提案法では，時間遅れと減衰率という2つのパラメータを推定する必要がある．本節では減衰率の推定方法について説明する．

減衰率の推定には，前節で推定した時間遅れと，平均自己相関関数を用いる．$\bar{R}_i$は，前節で述べたようにR_Sの近似になっているとみなしている．そのため，$\bar{R}_i$を減衰率を推定するための基準に用いて，$\hat{R}_i$の減少を見ながらα_{ij}を調整する．図2に減衰率を推定するためのアルゴリズムを示す．

まず，減衰率α_{ij}の初期値を0に設定する．そして，前節で推定したl_{ij}とα_{ij}を式(6)に代入し，ある経路の反射音の除去を行う．その結果，推定信号$\alpha_{i0}s^{(j)}{}_{ik-l_{i0}}$が計算される．この推定信号の自己相関関数$\hat{R}_i^{(j)}$を計算する．

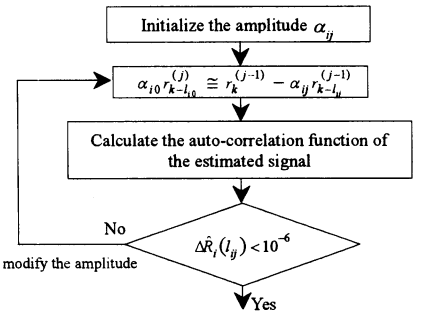


図2　減衰率の推定フロー

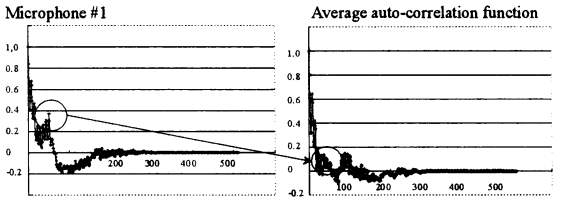


図3　減衰率の推定の自己相関関数上での処理

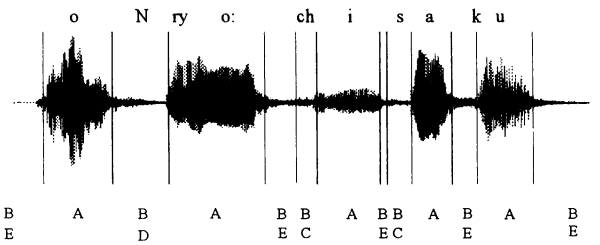


図4　信号の区間分割

次に$\Delta R_i^{(j-1)}(l_{ij}) = \hat{R}_i^{(j)}(l_{ij}) - \bar{R}_i^{(j-1)}(l_{ij})$を計算する．$\bar{R}_i$を$R_S$とみなしているため，$\hat{R}_i^{(j)}(l_{ij})$を十分に$\bar{R}_i^{(j-1)}(l_{ij})$に近づけることによって，ある経路の反射音を除去することができる．そのため，$\Delta R_i^{(j-1)}(l_{ij}) < 10^{-6}$となるように最急降下法を用いて$\alpha_{ij}$の最適値を推定する．図3に減衰率を最適推定する際の自己相関関数上での操作を示す．

2.4　信号の区間分割

従来提案していた手法[11]においては，すべての区間で反射音を除去する際に波形上での減算を行っていた．しかし，この方法では信号のパワーの弱い子音区間や，鼻音区間の音が消えてしまうという問題があった．そのため，十分な認識率の改善を得ることができなかった．

そこで提案法では，問題のあった子音の区間や鼻音の区間を検出し，その区間に関しては，スペクトル上で特徴を強調することを行った．例えば，摩擦音の区間では高域を強調し，鼻音では低域を強調している．さらに，反射音を十分に除去するために，音声区間であるか，無音区間であるかの判定も

行い，本来は無音であったと判定された区間に関してはスペクトル減算を行い十分に無音に近づける．そして，音声と判定された区間に関しては従来どおり波形上での減算を行う．

区間分割の方法について説明する．まず，音声区間検出の際に用いた閾値を用いて，受音信号を振幅の大きい区間(A)と，小さい区間(B)に分割する．そして，振幅の小さい区間(B)について周波数スペクトルの最大値を見つける．最大値が4kHz 以上(ただしサンプリング周波数は 16kHz)の場合は摩擦音(C)であると判定する．摩擦音でないと判定された区間については，まず F0 の抽出を行う．人の通常の音声の場合，F0 は男性で 100~200Hz，女性で 200~300Hz である．そのため，F0 が 400Hz 以下にありかつ，周波数スペクトルの最大値とほぼ一致している場合，その区間の音は鼻音(D)であると判定する．(C)，(D)どちらにも当てはまらない(B)の区間は無音区間(E)であると判定する．

このような処理を行うことによって，図4に示すように受音信号を分割する．このように受音信号の区間を分け，区間に応じて処理方法を切り替えることによって，認識率が改善するとともに音質の向上も図ることができる．

2.5 反射音の減算方法

提案法では，反射音の除去に波形上での減算とスペクトル上での減算を併用している．波形上での減算とスペクトル上での減算の使い分けは，2.4 で説明した区間分割に基づいている．波形上での減算は，式(6)で示したとおり，推定した時間遅れと減衰率に基づいて逐次的に減算していく．本節では，スペクトル上での減算について説明する．

SS 法の原理は次のようになっている．

$$\hat{X}(j\omega) = \left|Y(j\omega)\right|^2 - \left|\hat{N}(j\omega)\right|^2 \Big|^{\frac{1}{2}} e^{j\arg(Y(j\omega))} \qquad (5)$$

ただし，$\hat{X}(j\omega)$ は元の音声信号の推定したスペクトル，$Y(j\omega)$ は雑音付加音声のスペクトル，そして $\hat{N}(j\omega)$ は推定した雑音のスペクトルを表している．

しかし，式(5)を用いて減算処理を行う場合に，雑音のスペクトルの推定誤差により，ある周波数では雑音スペクトルが減算対象の信号のスペクトルを上回ってしまい，スペクトルが負の値をとる可能性がある．そのため，負の値になる場合にはスペクトルの補正処理を行う．補正処理の例としては，

$$\hat{X}(j\omega) = \begin{cases} \left|Y(j\omega)\right|^2 - \left|\hat{N}(j\omega)\right|^2 \Big|^{\frac{1}{2}} e^{j\arg(Y(j\omega))}, & \left|Y(j\omega)\right| > \left|\hat{N}(j\omega)\right| \\ 0, & \text{otherwise} \end{cases} \qquad (6)$$

や

$$\hat{X}(j\omega) = \begin{cases} \left|Y(j\omega)\right|^2 - \left|\hat{N}(j\omega)\right|^2 \Big|^{\frac{1}{2}} e^{j\arg(Y(j\omega))}, & \left|Y(j\omega)\right| > \left|\hat{N}(j\omega)\right| \\ \left|\hat{N}(j\omega)\right|^2 - \left|Y(j\omega)\right|^2 \Big|^{\frac{1}{2}} e^{j\arg(Y(j\omega))}, & \text{otherwise} \end{cases} \qquad (7)$$

のようなものがある．しかし，これらの補正処理を施しても，明瞭度は改善するものの，信号が歪んでしまい音声認識率の大幅な改善には至らない．この理由として考えられるのが，SS 法自体が雑音信号の定常性を仮定しており，信号の分散の

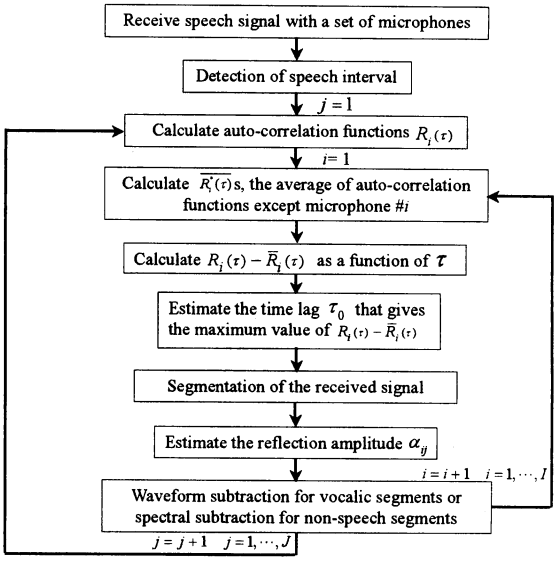


図5 反射音除去の処理手順
(ただし I はマイクロフォンの数，J は除去する反射音の数)

変動を無視していることがあげられる．そのため，提案法の SS 法では信号の分散を考慮に入れている．具体的には，式(8)に示す．

$$\hat{X}(j\omega) = \begin{cases} 0, & \left|\hat{\sigma}(j\omega)\right| < \left|Y(j\omega)\right|^2 - \left|\bar{\hat{N}}(j\omega)\right|^2 \\ \left|Y(j\omega)\right|^2 - (\left|\bar{\hat{N}}(j\omega)\right|^2 + \hat{\sigma}(j\omega))\Big|^{\frac{1}{2}} e^{j\arg(Y(j\omega))}, & \hat{\sigma}(j\omega) < \left|Y(j\omega)\right|^2 - \left|\bar{\hat{N}}(j\omega)\right|^2 \\ \left|\alpha * \left|Y(j\omega)\right|^2\right|^{\frac{1}{2}} e^{j\arg(Y(j\omega))}, & \text{otherwise} \end{cases} \qquad (8)$$

ただし，$\bar{\hat{N}}(j\omega)$ は推定した雑音のスペクトルの数フレームについて周波数ごとに平均したもの，そして $\hat{\sigma}(j\omega)$ は推定した雑音スペクトルの数フレームの周波数ごとの標準偏差を表している．提案法で用いているスペクトル減算の方法が必ずしもよいわけではないが，減算方法を変更することにより反射音の除去の精度が変わってくる．そのため，スペクトル減算の方法については，今後の課題の 1 つである．

2.6 処理の流れ

提案手法の処理の流れについて説明する．全体の処理の流れを図示したものを図 5 に示す．

まず，複数のマイクロフォンで受音した信号それぞれに対して二重閾値法を用いて音声区間を検出する．閾値の決め方は，音声が入力される前の雑音のみとみなせる区間と音声のみとみなせる区間それぞれから，信号の平均と分散を計算し，各発話に対して適切な値を自動的に設定している．ただし現在，提案法ではリアルタイム性を考慮に入れていない．そして，切り出された信号に対して，それぞれの自己相関関数を計算する．この自己相関関数を元に，2.2，2.3 で説明した手

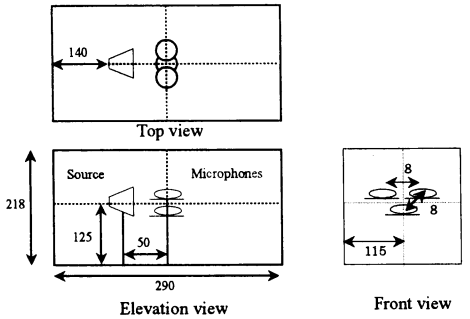
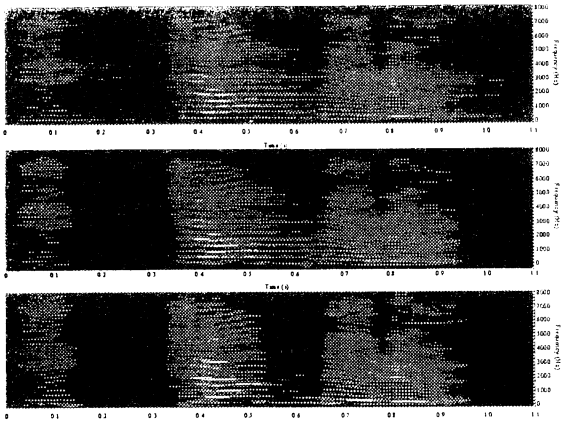


図 6　リビングルームシミュレータ内のスピーカとマイクロフォンの配置図



i　　q　ch　　a　　N　n　e　r　u

図 7　スペクトログラムの比較
(上段：反射音除去前　中段：反射音除去後　下段：元音声)

段によって時間遅れと減衰率を推定する．次に，信号の特徴によって区間の分割を行い各区間における反射音の除去方法を決定する．そして最後に反射音の除去を波形レベル，スペクトルレベルそれぞれで実行する．以上の処理を(反射音除去処理に用いるマイクロフォン数)×(除去する反射音の数)だけ繰り返す．

3.　認識実験

3.1　実験環境

提案法の有効性を検討するために実験を行った．実験環境は，リビングルームシミュレータという反射の多い部屋を用いている．残響時間は，白色雑音を用いて 440ms である．リビングルームシミュレータの中にスピーカを 1 個とマイクロフォンを 3 本設置した．スピーカとマイクロフォンの配置図を図 6 に示す．

3.2　実験条件

実験に用いた音声データは，防音室で接話マイクを用いて収録したもので，そのままでの音声認識率は 100%である．音声データの内訳は，男性 4 名，女性 1 名の計 5 名にそれぞれ 51 発話を発声してもらい，計 255 発話である．音声データの発声内容はテレビ操作コマンドで，例えば，「テレビ ON」や「チャンネル 1」などである．そして，表 1 に音声データの収録の条件および，音声認識に用いる辞書のサイズと文法ルール

表 1　音声収録条件と認識辞書，文法ルール数

Sampling rate	16ksamples/sec
Quantization	16bits

Vocabulary size	99
Grammar rules	13

表 2　各マイクロフォンの音声認識率の比較(%)

	Microphone#1			Microphone#2			Microphone#3		
	baseline	previous	proposed	baseline	previous	proposed	baseline	previous	proposed
M1	88	90	90	88	90	86	82	88	90
M2	61	76	76	67	82	88	61	78	82
M3	80		75	80		80	76		80
M4	90		92	90		96	86		92
F1	82	84	88	82	80	78	63	78	84

表 3　各発話について 3 本のマイクロフォンの多数決をとった認識率の比較(%) (ただし*有意水準 5%，**有意水準 1%)

	baseline	previous	proposed
M1	84	90	90
M2	67	84*	88**
M3	82		82
M4	90		96
F1	76	84	86

数を示す．音声認識デコーダは「Julian」[15]を用いている．

3.3　実験結果

本節では，音質の改善を評価するために，受音信号 $r_i(t)$と推定信号$\hat{s}^{(J)}(t)$のスペクトログラムの比較を行う．さらには，認識実験を行い，音声認識率改善に対する効果も評価する．

3.3.1　スペクトログラムの比較

図 7 にスペクトログラムの比較を示す．上段の図がリビングルームシミュレータ内であるマイクロフォンが受音した信号のものであり，中段の図が反射音除去処理後の信号のものである．そして下段の図が防音室で接話マイクを用いて受音した信号のものである．

上段の図は，壁や床，天井からの反射音の影響を受け，本来は無音の区間であるところにスペクトルが伸びてきていることが確認できる．しかし，中段の図を見るとスペクトルが延びていた区間が無音に近づいていることがわかる．また，これまでに提案してきた手法[11]では，反射音除去処理を全音声区間に渡って行っていたため，信号のパワーの弱い子音の部分や鼻音の部分が消えてしまうという問題があったが，中段のスペクトログラムからわかるように，鼻音の区間のスペクトルが十分に保存されていることが確認できる．

反射音除去処理によって，若干ではあるが音質が改善していることが確認できる．

3.3.2　音声認識率の比較

提案法が音質の向上だけでなく，音声認識率の改善にも有効であることを示す．

表 2 に音声認識率を比較した結果を示す．音声認識率は反

射音除去処理前(baseline)とこれまでに提案してきた手法(previous)[10], そして今回提案した手法(propose)の3つについて比較を行った. 表2を見るとわかるように, M1の2本目, M3の1本目, F1の2本目で若干認識率が低下しているところがあるが全体的に認識率が改善している. 平均すると約8%認識率が改善している.

表3には3本のマイクロフォンについて多数決をとり, 音声認識率を再計算した結果を示す. この場合は, 認識率が低下するものはなく, M3で変化がなく残りはすべて改善している. そして平均約8.6%認識率が改善している. 各話者について符号検定を行ったところ, M2の話者に関しては有意水準1%で有意差有りとなった.

4. 考察

5人の話者すべてのデータを用いて符号検定を行ったところ, 有意水準1%で有意差ありという結果になった. このことから提案手法が, 反射音を除去し音声認識率の改善に有効であるということが確認できた.

我々が以前に提案していた手法は, 閾値は手動で設定していたのだが, 今回は各発話について動的に設定している. 表2, 表3の結果から若干の認識率の低下は見られるが全体としては前回の認識率のレベルを維持していることが確認できる. そして, 実験的に閾値を決定するという手間を省くことができることを考えると, 動的に設定できる方が有効と言える.

新たに2人分, 102発話のデータを追加したが, 十分な改善が得られているとは言えない. この原因として考えられることは, マイクロフォンやケーブルの特性が他のものと比較して大幅に異なっていたことである. つまり, そのマイクロフォンについては現在用いている方法により設定した閾値では, 発話区間の検出に失敗しており, 反射音除去処理が適切に行われなかったのである. そのため, マイクロフォンの特性や, 今後雑音を考慮に入れていくことを考えた場合に, 頑健な区間検出法が必要になる.

さらに, 認識率の改善が十分に得られていない理由としては, 反射面での周波数特性を平坦であると仮定していることにあると考えられる. 文献[4]によると周波数ごとの残響時間を変化させた実験を行っているが, 若干の改善しか得られなかったという報告がなされている. しかし, 今回の我々の実験の場合にも同様のことが言えるかを検証する必要がある.

もう一つ考えられる理由としては, 反射音の減算の方法が十分でないことである. 現在の方法では, 主に信号のパワーの弱い部分についてのみスペクトル減算を適用しており, 無音区間に関しては十分な反射音の除去ができているのだが, 信号レベルでの減算を行っている音声区間に関してはまだまだ引き残りがあることがわかった. このことから, 音声信号全区間にスペクトル減算を行った場合についても考慮していく必要がある.

5. まとめ

本稿では, 複数マイクロフォンの自己相関関数に基づいた波形減算とスペクトル減算を用いた反射音除去法を提案した. 認識実験の結果, ライン入力で認識率100%の音声が認識率約80%に低下する無雑音環境において, 約8%の音声認識率の改善が得られた. また音声区間検出のための閾値を動的に設定することで提案法の汎用性が向上した.

また, データを追加したことによって, これまで明らかになっていなかった問題点も明らかになり, 今後の改善につながる結果が得られた.

今後の課題としては, 以下のものが考えられる.

- 閾値の設定方法を含めた頑健な区間検出法の検討
- 反射面での周波数特性の考慮
- 適切なスペクトル減算の方法
- 実際の部屋のインパルス応答と推定したものとの比較

これらの課題を解決することにより, さらに認識率が改善するものと考えられる.

謝辞

本研究の一部は, 文部科学省知的クラスタ創成事業ならびに同志社大学学術フロンティア開発事業の援助を受けた.

文献

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

[2] 田沢徹, 大西昇, 杉江昇, "単一マイクロフォンを用いた未知環境におけるエコーキャンセラー", 計測自動制御学会論文集, Vol.30, No.4, pp.460-466, Apr.1994.

[3] 馬場朗, 松本大典, 李晃伸, 猿渡洋, 鹿野清宏, "家庭環境におけるスペクトルサブトラクションによる残響抑圧を利用した音声認識", 音響講論, no.1-1-9, pp.17-18, Sept.2004.

[4] 中島弘史,鶴秀生,東山三樹夫,"残響時間の周波数特性を考慮した残響時間低減処理,"音響講論, no.1-Q-4, pp.567-568, Mar.1998.

[5] 古川正和,鵜木祐史,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,"信学技報,SP2002-15,pp. 49-54,Apr.2002.

[6] Nakatani, T., and Miyoshi, M., "Blind dereverberation of single channel speech signal based on harmonic structure," Proc. ICASSP-2003, vol. 1, pp. 92-95, Apr.2003.

[7] Takiguchi, T., Nakamura, S., Huo, Q., and Shikano, K., "Model adaptation based on HMM decomposition for reverberant speech recognition," Proc. ICASSP-97, vol. 2,pp. 827-830,Apr.1997.

[8] 馬場朗, 李晃伸, 猿渡洋, 鹿野清宏, "残響適応音響モデルを用いた音声認識", 音響講論, no.1-9-14, pp.27-28, Sept.2002.

[9] 山本仁, 西本卓也, 嵯峨山茂樹, "フレームごとのモデル合成による残響下音声認識" 信学技報,SP2003-134,pp.127-132, Dec.2003.

[10] 大田健紘, 柳田益造, "自己相関関数に基づいた遅延時間推定による反射音除去法", 音響講論, no.2-4-20, pp.311-312, Sept.2004.

[11] 大田健紘,柳田益造,"実環境下における音声認識率向上のための残響除去技術の検討," 情報講論(2),pp.47-48,Mar.2004.

[12] 松本大典, 馬場朗, 李晃伸, 猿渡洋, 鹿野清宏, "ロボットへの音声入力を目指した遠隔発話の認識", 音響講論, no.1-1-5, pp.9-10, Sept.2004.

[13] 戸井真智, 鵜木裕史, 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", 音響講論, no.2-3-19, pp.643-644, Sept.2004.

[14] 鵜木裕史, 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", 音響講論, no.3-P-7, pp.609-610, Mar.2004.

[15] http://julius.sourceforge.jp/