調波構造を用いたモノラル音声からの残響除去に要する学習データの削減

Training data reduction for Blind dereverberation of single channel speech signal based on harmonic structure

渋谷　涼
Ryo Shibuya
法政大学情報科学部ディジタルメディア学科
E-mail: 08k1123@stu.hosei.ac.jp

Abstract

In this paper, reverberation that occurs the remote utterance recording in the room removing method is proposed. When reverberation has contained sound, it has made clarity and a performance of speech recognition dawn. Proposed method uses harmonic structure that is key properties of sound. So it can process with a single channel speech signal. Moreover, dereverberation filter created from short time input sound has had error and cause stationary noise. This noise rejected using spectral subtraction method (SS). Reverberation was reduced from input sound that reducing 5240 words (about 1 hour — conventional technique has used and made high performance) to 500 words (about 8 minute). A recognition rate and a spectrum compared the sound of reverberant with dereverberated .

1　はじめに

技術の進歩に伴い, 音声を集音し活用する場面が増えている. 例えば, 携帯電話やビデオカメラはもちろん, カーナビゲーションシステムやロボット, ゲームなどである. 音声を用いると, 発話するだけでよいため比較的容易であり, 誰でも気軽に利用できる, 手の離せない場面で用いることができるという利点がある中でも, スマートフォンや Kinect の台頭により, 集音した音声を音声認識し, 認識結果を利用するアプリケーションが増加している. 遠隔発話時の音声の集音・録音の多くの場合では, 残響や背景雑音が含まれてしまい, それらは音声の明瞭性を損ない, 音声認識システムの性能を低下させるため, アプリケーションの誤作動の原因となってしまう.

残響除去に関しては, 1 つの目的音に対して, 複数のマイクロホンを用いて観測した信号を用い, 音量差・遅延差などから音源方向を特定し, 空間特性によって付加された成分を取り除く方法が活発に研究されている [1]. しかし, 複数機材を用いての集音はコストがかかる. また, 軽量・小型化に反し利便性に欠ける. したがって, 手軽に行える録音時に発生する残響・雑音に対し, より一般的かつ容易に行える除去法が必要となる.

本研究では,[2] の手法をもとに音の基本的性質となる調波構造を用いることで残響除去残響除去フィルタを作成し, 残響除去に要する入力音声の短縮を行う. [2] のフィルタは目的音声や室内伝達関数など事前情報を用いないブラインド処理により単一入力のみで処理可能である. ここで, 入力音声を短くして作成したフィルタでは, 処理後に定常雑音が発生してしまう. 発生雑音による音声認識率の低下を防ぐため, 雑音除去を行った. 学習データとなる入力音声を短くすることで, この手法の適応範囲が増えると期待される.

Supervisor: Prof. Katunobu Itou

2　従来の残響除去

2.1　残響と残響除去

室内での音の伝わり方を表す関数で, 音に空間的な広がりを与えるものに室内伝達関数というものがある. これは, 室内の広さや壁・床などの素材によってそれぞれ変化し, 空間ごとに特有の値を持つ. これを用いると, 室内で発せられた音声 S は, 室内伝達関数 H によって響きの影響を受け, 録音地点での観測信号 X となる. また, 室内伝達関数 H は発話地点からマイクロホンへ音を真っ直ぐ飛ばす D と, 一度壁などでの跳ね返りを経過してマイクロホンに到達させる R に分けることが出来る. よって, 発話地点からマイクロホンへの到達の仕方は, 次のように表すことが出来る.

$$\begin{aligned} X &= HS \\ &= (D+R)S = DS + RS \end{aligned}$$

残響には, 音質の低下, 音声の明瞭性の損失, 音声認識システムの性能低下 (「法政大学」と発話した音声に (東館体育館の) 残響を付加させたところ, 認識結果は「思えへっ. 方法!」となる) を招くという問題点がある. そこで, これらの問題を解消するため, 残響の除去・軽減が必要となる. 多くの場合, 上式中の直接音 DS を残響除去後の目的音声として, 残響成分 RS を除去し, 残響を除去している. 従来の残響除去法は主に空間情報を用いる方法・複数音源を用いる方法・モノラル音声を用いる方法の 3 つに分けることが出来る.

2.1.1　空間情報を用いる方法

あらかじめ測定録音時の空間伝達関数 (インパルス応答) がわかっている場合の残響除去は, 比較的たやすい. インパルス応答の影響を打ち消す働きをする逆フィルタを求め, 畳み込むことによって残響を除去する. この方法は容易い反面, 反響しやすい公共の音楽ホールなどでインパルス応答を求めることが難しい, わざわざ測定をしなければならないなどの問題点がある.

2.1.2　複数音源を用いる方法

同時刻発話に対して複数のマイクロホンを用いて録音した音声に対しての手法も様々行われている. 複数音源に対しては, マイクロホンアレー法がよく使用されている. 方法としては, 用意したマイクロホンを発話者の前方の異なる位置に配置し, すべてで発話音声を録音する. すると, マイクロホンの配置位置の違いによって, 音の伝達の仕方に変化がつくため, そこから音の空間特性を得られるというものである. また, モノラル音声を用いた処理を複数マイクロホンに拡張し, 性能の向上を図られることもある.

2.1.3　モノラル音声を用いる方法

信号自体から音声の部分を予測し, 空間伝達関数を求める方法が一般的である. モノラル音声からの残響除去法の研究は従来研究も多いが, 未だ課題の残るものが多い. 残響が付随することによって音声の波形構造自体が変化してしまうため, 残響部分とそれ以外の部分を分離することが難しい. しかしながら, 多くの音声はモノラルなことが多いため, 研究が進めば, 応用範

囲の広い分野である．本研究で利用する調波構造を用いた方法もこの分類である．

3　調波構造を利用した残響除去

音は周期成分と非周期成分で構成されている．

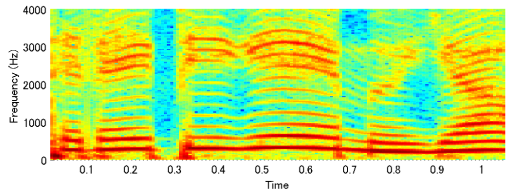

図 1. 発話音声（「テレビゲームや」）のスペクトログラム

このスペクトログラムには, 音素ごとに異なる複数の横縞模様が見られる．この横縞部分が周期成分である．基本周波数と呼ばれる周波数を元に 2 倍,3 倍,...,k 倍まで一定の周波数間隔を持った複数の周波数が含まれ, 調波構造と呼ばれる．その音固有の調波構造を知ることが出来たならば, 人為的に楽器音や声などに似せた音を作ることが出来, 比較的たやすく直接音を近似推定することが可能となる．

また, 除去目的となる残響は音声が減衰と時間遅延を伴いながら音声に残存する成分であるため, 調波構造である．しかしながら, 時間遅延を伴っていることから, 同時刻での音声成分と残響成分の調波構造（周波数）は異なるものと考えられる．そして, 減衰を伴っているため, 残響成分のパワーは同時刻の音声成分のパワーよりも小さくなる．

今手法では, 入力音声をそのまま学習データとして用いて残響除去フィルタを推定する．まず, 入力音声 X より調波構造を求め, これを直接音 DS の推定 $\hat{X}$ とする．次に, 入力音声 X と直接音の推定 $\hat{X}$ より室内伝達関数の逆フィルタの近似, 残響除去フィルタ $\hat{W}$ を求める．最後に, 作成した残響除去フィルタによって入力音声をフィルタ処理することによって残響除去を施すというものである．この手法は, 録音時の情報を必要とせず, モノラル音声のみで処理が可能であるということから, より多くの音声に対し一般的であり, 応用が利く方法といえるだろう．

3.1　調波構造推定

調波構造を利用して, 入力音声から残響以外の音声すなわち直接音を推定する．まず, 文章単位の入力音声を短い時間フレーム (42ms 幅の時間フレームを 1ms ずつシフトしていく) に分割し, 各時間フレームごとに基本周波数である F0 や倍音の位相・振幅を求め, 正弦波合成法により調波構造を作成．各時間フレームごとにできた調波構造を時間軸上の 1 つの波形に戻すことによって, 時間の経過により周波数の変動がみられる調波構造を推定することができる．また, 今手法では, 周期的であり調波構造のある部分を音声としているため, 調波構造ではない s, k, t などの非周期的な子音部分は音声として考えないものとする．よって, 全周波数のパワーと調波構造のパワー比を用いて有声判定を行い, 子音部分を取り除いた．

3.2　残響除去フィルタ推定

先に調波構造によって推定した直接音を用いて, 残響除去フィルタを推定した．空間伝達関数から残響の影響を打ち消す, すなわち, 直接音成分のみを抜き出す逆フィルタ W は以下のようになる．この逆フィルタ W は残響付きの入力音声 X と推定直接音 X' により求められる．

$$\begin{aligned} W &= D/H, \\ WX &= D/H * HS = DS \\ W &= X'/X \end{aligned}$$

各文個々に含まれていない周波数を補うため, 異なる文章において求めた $W(\omega)$ の平均をしている．

3.3　従来手法での問題点

この手法では, 入力音声から直接音を調波構造フィルタ処理によって推定し, その調波構造と入力音声の調波構造との差異を取ることで, 空間伝達関数の逆フィルタを推定している．ここで, 様々な音声に対して適応可能なフィルタを推定するためには, 周波数全範囲に対する周波数毎の変化特性が必要となる．発話音声において, 調波構造は時間的に変化していくため, 長時間 (4250 語：約 1 時間以上) 用いて残響除去フィルタを学習したならば, ほぼすべての周波数に対しての残響特性を得ることができる．よって, 室内の伝達特性が一定であり, 十分に長い観測信号が得られる場合には, 高性能な残響除去が可能となる．

0.3213s 残響付加発話音声約 30 分から実際に残響除去フィルタを推定した．残響として用いたインパルス応答の逆の特性を求め, それを求めたい正解として比較すると周波数応答の大まかな形は似ているものの, 特性にばらつきが出てしまった．

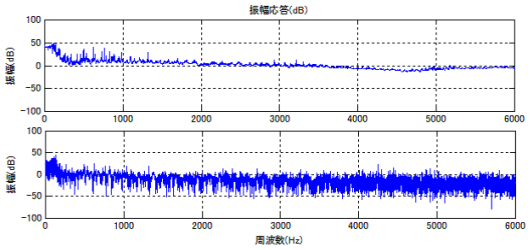

図 2. 用いた残響の逆特性 (上) と推定した残響除去フィルタ (下)

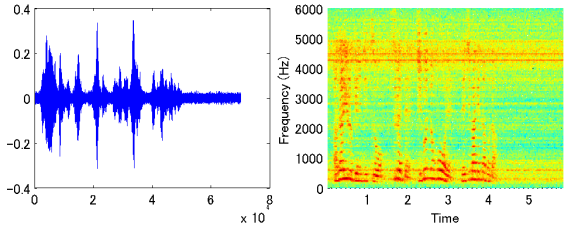

図 3. 残響除去後の波形とスペクトログラム

このフィルタを用いて残響除去してみると, 定常的な雑音が発生してしまう．このように, 調波構造を用いて行う当手法は短時間発話音声を用いて学習し, 推定された残響除去フィルタからも音声信号の時間構造, 及びスペクトル構造が効果的に回復可能であり, 残響は除去可能だが, 周波数ごとの残響特性の不足により, 出力結果に非周期成分に由来する推定誤差が含まれ, ランダムな定常雑音が発生してしまうという問題点があげられている．

4　短時間音声による残響除去

従来手法は調波構造と逆フィルタの原理を用いるためわかりやすいこと, モノラル信号のみにより処理が可能という点で将来性が大きいといえるだろう．一方, 録音時の発話者の空間的な移動（部屋から廊下に出るなど）がない音声を対象としているため, より多くの音声に適応可能とするため, 必要学習データ数の短縮は必須であり, 一番の課題であるといえるだろう．そこで, 従来研究より少しでも短い入力音声からフィルタの推定および残響除去を行い, 発生雑音に対処することで残響除去に必要な入力音の短縮を行う．

4.1　評価

4.1.1　音声

評価に使用する発話音声は, 音素バランス文 (日本語に含まれる音節や音素をできるだけ多く含んでいる文) の発話と, 新聞読み上げ文章を用いて行う．音素バランス文は, 日本語音声に含まれる 124 音節・27 種類の音素をすべて含み, 前後の音の組合せとして出来るだけ多くの音節の組み合わせを含んだ設計となっている．それらサブセット A18〜A50,B〜I：各 50 文,J：53 文の計 503 文,3199 語と, 新聞読み上げ文章から 1921 語を使用した．

主にフィルタ推定・評価用に用いる残響音は人為的に用意し

た．まず, 文単位の発話音声に, 個別に残響を付加させていく．これは, 事前に測定済みの残響成分であるインパルス応答を畳みこみ, 人為的に行った．残響音としてこのような音声を用いる理由としては, 今回, 残響特性となるインパルス応答を特性の異なる複数用いるため, それぞれに対して容易に残響音を得られるためである．また, 実際の残響付加前後の音声情報および残響・雑音の特性がそれぞれ個別にわかるため, 推定値の精度の評価が行えるという狙いがある．

これらの音声を学習に用いる際は,5240 語 (従来手法),2000 語,1000 語,500 語の複数の学習データ長を用いた．

4.1.2 残響

実験に用いた残響の特性は TSP 応答を用いて測定した [4]．用いたスイープ音は 1 秒間に 10Hz から 8kHz まで変化するスイープ +5 秒間の無音が 50 回連続する信号であり,50 回同期加算することによって雑音の影響を軽減している．計測には, 会議室と小体育館の 2 環境でマイクロホン, 一般家庭でもゲームなどにより使用頻度の高い kinect を用いた．また, 会議室の計測では, 音源から録音位置までを 1m,2m,3m と複数行った．

残響時間は, 音のエネルギーが発生した時点から-60dB(RT60 という) になるまでの時間と定義されている．よって, シュレーダー積分 ([4]) により残響曲線を求め, その曲線の中で残響特性を最も表している部分の減衰時間を割り出し, そこから-60dB まで減衰するまでの時間を計算した．

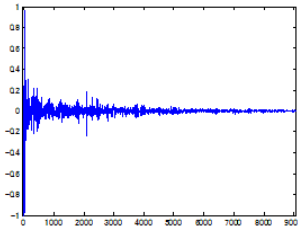

図 4. 観測されたインパルス応答の波形

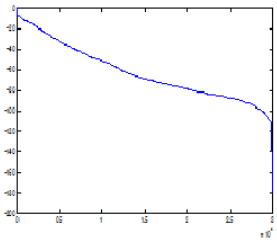

図 5. 残響曲線

このように残響曲線から直接-60dB まで減衰する時間を決めない理由としては, インパルス応答計測時の雑音のレベルが大きい, すなわちインパルス応答の S/N 比が大きくなってしまうと, シュレーダー積分による残響曲線が-60dB まで減衰せず, このような現象がよく起きるためである．計測した各残響から求めた残響時間は以下のとおりである．

表 1. 測定した各残響時間

		kinect	Mic
小体育館		0.3213	0.6210
会議室	1m	0.0853	0.1767
	2m	0.1488	0.1790
	3m	0.1702	0.2077

この中から, 評価には比較的長い残響として 0.7169s(小体育館・マイクロホン), 短時間残響として 0.0853s(会議室 1m・kinect), その中間として 0.3213 s (小体育館・kinect) の 3 つの残響を用いた．

4.1.3 評価方法

残響を付加させた学習データ長 5240 語 (約 1 時間：従来手法),2000 語,1000 語,500 語 (目標) それぞれを用いて推定された残響除去フィルタの評価は, 実際に聞き比べるとともに, スペクトログラムや波形, 音声認識によって行う．

—スペクトログラム・波形—

残響除去前後のスペクトログラムや波形を表示し, 比較する．音声自体を直接視覚的に見ることができるため, 一番効果がよくわかる方法だろう．また, 今研究では人工的に残響を付加しているため, 残響付加前の音声がきちんと復元できているかの確認も行う．

—音声認識—

各学習データ長によって推定したフィルタによって残響付加音 100 単語の残響除去前後の認識率を比較する．認識には Julius を用い, 音素バランス文サブセット A〜J までの 503 文中の単語 3304 語を含む辞書を用いた．

4.2 残響除去部

調波構造を推定する際, 入力音声から推定した F0 を元に倍音成分を求めているため,F0 の推定精度は重要である．しかしながら, 残響が付随するとある地点の波形にその前の波形が重畳され, 音声波形自体が変化してしまうため, 入力音声から直接的に基本周波数の F0 を推定することは難しい．そこで,F0 を推定する前処理として,[3] に挙げられる同周波数に持続する信号の抑制を行い, 残響が F0 推定に与える影響を軽減した後 F0 を推定した．この処理は, 原音 S が減衰と時間遅延を伴ったものが残響 RS であるという残響の性質を利用したものである．例えば, 残響の影響を受けた観測信号 x のある時間 t の音声 $x(t)$ は,$x(t)$ での直接音 DS に $x(t)$ 以前の信号が弱まったものが加わってできた信号ということになる．重畳されている成分の強さを r, どの程度前の成分が含まれているかの長さを m で与え, 観測信号より差し引いて残響部分を軽減している．

この前処理をするにあたって, 変数である r と m の値に関し, 従来手法では $r = 1/(m-1)$ 以下の正定数とし,r, m について, “ 最適フィルターパラメーターは残響時間によるが, 私たちは全状態に適応できるパラメータを選んだ ” とされており, 明確に提示されていない．そのため, 今研究で用いる変数 r,m の値を決定するため, 検討実験を行った．r,m の値を変化させて前処理を行った音声波形より F0 を推定し, 正解 F0 との一致率を求めた．一致率は,[5] に挙げられている正答率の評価尺度を用いて計算した．

$$\text{一致率} = \frac{N_{F0_{\text{正解}}}}{N_{F0_{\text{前処理}}}(E)} * 100$$

$N_{F0_{\text{正解}}}$ ： 正解 $F0$ の個数

$N_{F0_{\text{前処理}}}(E)$ ： $|F0_{\text{正解}} - F0_{\text{前処理}}|/F0_{\text{正解}} \leq E$ を満たす個数

ここで, 各 N は,$F0 < 0$ の範囲のみで算出し, 正解 F0 と前処理後 F0 の誤差 E=0.05(5 %) とした．これを,$r <= 1/(m-1)$ と定義があるため,

・$r = (1/(m-1+a))$ とした場合の,a の値：0〜1000 まで 100 刻み

・m の値：50〜500 の 50 刻み +500〜16000 まで 500 刻み

でとり, 全組み合わせ行った．正解 F0 にはクリーン音声から求めた F0 の値を用い,RT=0.7169s,0.0853s,0.3213 s の残響付加音に対して行った．$m = 500$,$a = 200$ 付近において, 今回用いた 3 つの一致率が高くなる傾向が見られたため, 残響時間が異なるのものであっても, この値を用いることである程度の F0 の推定精度の向上が見込まれる．よって今回前処理を行う際は, 一律 $m = 500$,$a = 200$ の値を用いることにした．RT=0.7169s では, 前処理を行う前の F0 の一致率 35 %であったものが,57 %まで向上した．同様に,RT=0.3213 s では 38 → 46 %,RT=0.0853s では 46 → 53 %となった．

4.3 定常雑音除去

4.3.1 雑音除去法検討

学習データとなる入力音声を短くした際, また, 今回は残響除去時に発生してしまう雑音を除去する．雑音除去に関しては, 従来, 多くの研究がなされている．ここでは, スペクトルサブトラクション (SS) 法 [6], ウィーナフィルタ (WF) 法, ランニングスペクトルアナリシス (RSA) 法の 3 種類の雑音除去法を実験し, 今回の残響除去法と相性がよく, 今回の雑音を効果的に除去する手法の検討を行った．

各手法ごとに雑音の除去程度は異なった．残響除去のみでは 35 %であった認識率が,SS 法適応後 59 %,WF 法 29 %,RSA 法 53 %となった．本研究では, 一番認識率がよく, 処理時間が

短いという理由から残響除去には,SS 法を用いて評価を行っていくこととした.

4.4 雑音除去適応

残響除去→雑音除去の流れに伴う雑音付きの入力音声の変化は図 6〜図 9 のようになる.

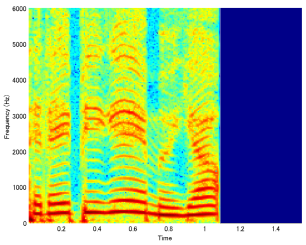

図 6. クリーン音声

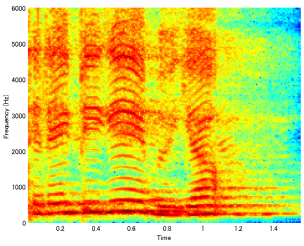

図 7. 残響付き

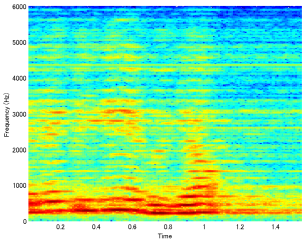

図 8. 残響除去処理後

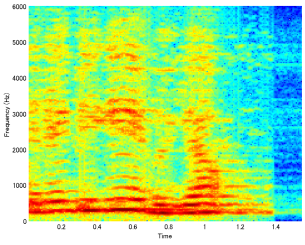

図 9. SS 処理後

全体的に残響成分が軽減できていることがみられる. 特に, 音声の末尾に付随していた残響が短くなっている.

また,Julius を用いての音声認識結果は以下のようになった.

表 2. 残響除去後の単語認識率 (%,SS 法適応前→適応後)

		付加残響時間		
		0.6210s	0.3213s	0.0853s
残響音		19 → 22	40 → 45	73 → 76
学習データ量	5240 語	16 → 24	35 → 59	43 → 64
	2000 語	15 → 24	24 → 56	22 → 64
	1000 語	10 → 24	21 → 53	17 → 59
	500 語	7 → 18	17 → 52	17 → 54

残響除去のみではすべての場合で除去前よりも認識率が下がってしまった. これは, 先の残響除去部分でもあげたとおり, 作成した残響除去フィルタの現状により, 定常雑音が発生しやすくなっており, その定常雑音が原因となっていると考えられる.

SS 法適応あるなしでの認識率で比較すると, すべての場合で認識結果の向上が見られた. 特に, 学習データを少なくしていくにつれて,SS 法適応後の認識率の向上率が大きく,5240 語に近い値となっていることより, 学習データを削減した際に発生してしまう定常雑音の除去が効果的に行えていることがわかる. これにより, 残響除去後に雑音除去 (SS 法) を組み合わせることは, 音声の劣化を防ぐために有効でり, 学習データすなわち入力音声の短時間化にも有効であると考えられる. SS 法を行っても誤認識になってしまった音声には, 元の音声成分が消えてしまったり, ミュージカルノイズの発生がみられた.

音声成分が消えてしまった部分については, 周波数がある程度一定状態で持続している場合に多く, 残響として推定されてしまったために, 残響除去時に除去されてしまったと考えられる. また, ミュージカルノイズ発生については, 直接音では音声成分があまりなかった部分で, 残響によって成分が多く見られるようになった時間帯や周波数で多く発生していた. これは, 残響除去フィルタの推定誤差により, 残響として軽減される成分とされない成分がまばらに存在してしまったことと,SS 法の際の雑音成分の差し引きすぎによるものであると思われる. このように, 雑音除去を行っても誤認識となってしまったり, 音質の低下がみられるものの主な原因としては, やはり, 調波構造を用いて推定した残響除去フィルタの推定精度があまり良くないということがあげられるため, 残響除去フィルタをしっかりと再現することができれば, これらの事例は減少させることができる. また, やはり, 学習データが短い場合ではフィルタの周波数特性の分散が大きくなってしまい, 入力音声の時間平均を取ったとしても収束せず, 定常雑音の発生量が多い. 試しに 500 語以下の学習データ数で残響除去フィルタの推定・除去,SS 法を行ったところ, データ数が少なくなるにつれて, 元の音声部分が除去されてしまった.

5 まとめ

残響付きの入力音声から調波構造をもとに残響除去フィルタを推定し残響除去を施すとともに,SS 法適応による雑音除去を組み合わせて行った. 適応後の音声では, いずれの場合もスペクトル表示によって残響成分の軽減を確認することができ, 認識結果も適応前より良い結果を得られた. SS 法を適応することによって定常雑音を除去することはできたが, 差し引きすぎてしまう場合があると考えられ, 音声自体も削られてしまうものもあった. これは,SS 法の適応の際の課題でもあるが, 残響除去フィルタの推定精度があまく, 残響除去結果には SN 比の小さい定常雑音が発生しやすいことが主な原因だと思われる. 調波構造の推定の際, 総じて高調波成分, 特に 2000〜4000HZ 付近の成分がうまく推定できず, その帯域に多く見られた. 今回は, 調波構造を求める際の F0 推定時前処理段階において, フィルタパラメータ r と m を実験により算出した. その結果, 今回用いた残響すべてに対して比較的よい結果となる値を 1 つ決めた. ここで, 大まかな残響の長さによって残響除去フィルタ推定時の各種パラメータを残響ごとに変化させることで, より正確な調波構造が求められるのではないかと期待される. しかし, 大まかな残響の長さをどのように決めるかということが必要となる. これらを考慮に入れ, 全体的な残響除去性能の向上にむけ, 今後も残響除去フィルタの推定精度向上を行っていきたい.

学習データ数については, 認識結果より雑音除去を組み合わせることによって 500 語まで減らすことができると考えられる. 現段階では雑音除去を組み合わせても認識率は残響付き音声より多少良い程度だが, 残響除去フィルタの再現度を上げることで効果的な残響除去が可能となると期待できる.

参考文献

[1] 木下慶介, 吉岡拓也, 中谷智広, “ 音声信号のブラインド残響除去：最新の研究動向 ”, 電子情報通信学会技術研究報告, vol. 110, no. 56, SP2010-5, pp. 25-30, 2010 年 5 月

[2] 中谷智広, 三好正人, 木下慶介, “ 調波構造に基づくモノラル音声信号のブラインド残響除去 ”, 電子情報通信学会論文誌, 2005/3 Vol. J88-D-ll No.3.

[3] Tomohiro Nakatani, Masato Miyoshi, “ Blind dereverberation of single channel speech signal based on harmonic structure ”, Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03). 2003 IEEE International Conference on ,I - 92-5 vol.1

[4] M. R. Schroeder, ”A new method of measuring reverberation time”, J. Acoust. Soc. Am.,vol. 37,pp 409 − 412,1965

[5] 鵜木祐史, 細呂木谷敏弘, 石本祐一 “ 残響環境下におけるロバストで正確な F0 推定法の比較評価 ”, 電子情報通信学会技術研究報告, vol. 107, no. 435, SP2007-168, pp. 7-12, 2008 年 1 月

[6] Steven F. Boll, ”Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Transactions on Signal Processing, 27(2),pp 113-120, 1979