音声のブラインド残響除去：最新の研究動向

Recent Advances in Blind Speech Dereverberation

木下慶介 Keisuke KINOSHITA
中谷智広 Tomohiro NAKATANI

吉岡拓也 Takuya YOSHIOKA

アブストラクト 室内で収録された音声信号には、音源からマイクロホンへ直接到達する直接音に加えて、壁や床などに反射して遅れてマイクロホンへ到達する残響が含まれている。この残響は収録した音声信号の明瞭性を低下させるとともに、コンピュータによる自動音声認識をはじめとする多くの音響信号処理アプリケーションの性能低下を招く原因となる。このため、収録音からの残響の除去は、古くから、実環境音響信号処理の実現に向けた重要課題とされてきた。その中でも特に重要度の高い課題である、収録条件が与えられていない任意の条件で収録された音声に含まれる残響の除去 (=ブラインド残響除去) の問題は、解決の困難なものとされており、多くの提案がなされてきた。本稿では、音声のブラインド残響除去の近年の幾つかの研究成果を挙げ、それらの特徴をまとめる。

キーワード 音声, ブラインド残響除去, 最新動向

1. はじめに

ハンズフリー電話やテレビ会議システム、補聴器、コミュニケーションロボット等の様々な音声コミュニケーションシステムにおいて、マイクロホンで収録される音声には、しばしば残響や雑音等の音響的なひずみが含まれる。音声信号に重畳するこれらの音響的なひずみは、音声の明瞭性を低下させ、自動音声認識の性能を大きく劣化させる要因である。実環境において音声コミュニケーションシステムの性能・精度の向上を図るには、収録音声からこれらの音響的なひずみの影響を低減することが重要である。

音声信号に重畳する雑音の影響を低減する技術（雑音除去技術）については、古くから研究されており、効果的な方法が提案されている。例えば、スペクトル減算法は、パワースペクトル領域での音声成分と雑音成分の加法性を仮定し、観測信号のパワースペクトルから雑音パワースペクトルの推定値を減算することで、雑音による音響的なひずみの影響を低減する⁽¹⁾。この方法を用いることで、音声の明瞭性や、自動音声認識の性能を改善できることが知られている。ここで用いられる雑音成分

図1 残響音声の生成過程, M はマイクロホンの個数

の推定方法も多岐に渡り、非音声区間のパワースペクトルの平均値からそれを推定するものや、近年では、あらかじめ学習した音声の統計モデルと観測信号を比較することで雑音成分を逐次推定する手法⁽²⁾等が提案されている。

一方、収録音声に含まれる残響の影響を抑圧（残響除去）することは長らく困難な問題とされてきた。本稿では、近年提案された残響除去方法の幾つかを挙げ、各々の手法の特徴を概説する。特に、ここでは残響除去方法をシングルチャネル処理、マルチチャネル処理に分類しまとめた。マルチチャネル処理に分類されているものの中には、シングルチャネル処理でも効果的に動作することが実験的に確認されているものも含まれるが、それらの処理の原理や定式化はマルチチャネル処理に基づいているため、同処理カテゴリーに分類した。また、本稿で挙げるものよりも前に提案された残響除去法については、文献(3)(4)などを参照されたい。

木下慶介 正員 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
E-mail kinoshita.k@lab.ntt.co.jp
吉岡拓也 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
E-mail yoshioka.takuya@lab.ntt.co.jp
中谷智広 正員 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
E-mail nakatani.tomohiro@lab.ntt.co.jp
Keisuke Kinoshita, Member (NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation), Takuya Yoshioka, Nonmember (NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation), Tomohiro Nakatani, Member (NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation).

電子情報通信学会 基礎・境界ソサイエティ
Fundamentals Review Vol.4 No.4 pp.301-310 2011 年 4 月
©電子情報通信学会 2011

図 2 音声生成と残響音声生成

2. 信号モデル

始めに、本稿で扱う残響の信号モデルを示す（図 1）． m 番目のマイクロホンで観測される残響を含む音声（残響音声）の時間領域信号 $x_m(n)$ は、クリーン音声 $s(n)$ と音源とマイクロホン間のインパルス応答 $h_m(n)$ の畳込みで表される．

$$x_m(n) = h_m(n) * s(n) = \sum_{i=0}^{K-1} s(n-i)h_m(i) \quad (1)$$

ここで K はインパルス応答 $h_m(n)$ を有限長と仮定した場合の応答長、 n は離散時間のインデックスを表す．行列表現を用いると、式 (1) は以下のように表される．

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}(n)$$

ここで

$$\mathbf{h}_m = [h_m(0), h_m(1), \dots, h_m(K-1)]^T$$

$$\mathbf{s}(n) = [s(n), s(n-1), \dots, s(n-(K-1))]^T$$

ブラインド残響除去とは、 $(x_m(n))_{1 \leq m \leq M, 1 \leq n \leq N}$ が観測されたときに、 $h_m(n)$ や $s(n)$ にかかる明示的な事前情報を用いずに、 $x_m(n)$ に含まれる $h_m(n)$ の影響を除去・低減する処理を指す．

3. 音声のブラインド残響除去の難しさ

音声のブラインド残響除去の難しさを示すため、図 2 に音声 $s(n)$ の生成過程も含めた残響音声 $x_m(n)$ の生成過程を示した．この図からも分かるとおり、音声 $s(n)$ は、白色の駆動信号 $u(n)$ と声道のインパルス応答 $\alpha(n)$ との畳込みにより生成され、残響音声 $x_m(n)$ は音声 $s(n)$ と室内インパルス応答 $h_m(n)$ の畳込みにより生成される．結果として、観測信号 $x_m(n)$ の生成過程は以下のように表されることが分かる．

$$x_m(n) = u(n) * \alpha(n) * h_m(n) \quad (2)$$

$x_m(n)$ だけを観測し、そこから音声 $s(n)$ を回復するには、観測信号の中に含まれる二つの未知のインパルス応答 $\alpha(n)$ と $h_m(n)$ を区別し、 $h_m(n)$ のみの影響を除去する必要がある．観測信号 $x_m(n)$ のみからこの二つの伝達関数を区別する方法は自明でないため、音声のブラインド残響除去は未解決問題として残され

てきた．近年考案されている残響除去技術の多くは、これら伝達関数に係る情報に何らかの統計的なモデルを仮定し、その仮定を基に区別を行なっている．

4. シングルチャネル処理

本章では、シングルチャネル処理、すなわち $M = 1$ の状況下で残響除去を達成する手法についてまとめる．一般的に、シングルチャネル処理の場合、室内インパルス応答の因果的逆フィルタは存在せず⁽⁵⁾、また非因果的な逆フィルタのブラインド推定はいまだ難しいため、現在までに提案されたものはモデル化精度の粗いものが多い．ただし、一般的には、その分、それらの処理は外乱等に頑健であるという特長を有している．

4.1 統計的室内インパルス応答モデルに基づく方法

Lebart らは、自動音声認識における残響の影響を低減するため、残響音声に含まれる後部残響の影響を低減する方法を提案した⁽⁶⁾．図 3 に、処理ブロック図を示した．この方法では、式 (3) に示す Polack の統計的室内インパルス応答モデルを用いる．

$$h_1(t) = \begin{cases} a(t)e^{-\Delta t}, & \text{for } t > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$$\Delta = \frac{3\ln(10)}{\text{RT}_{60}}$$

ここで、 $a(t)$ は、分散 σ_a^2 平均 0 の定常白色雑音、 RT_{60} は残響時間である．この統計的モデルでは、室内インパルス応答のエネルギーは指数減衰することが仮定されている．式 (3) を用いると、以下のように、フレーム l の後部残響の短時間パワースペクトル $|Z_1(l, f)|^2$ を、 N_{shift} フレーム前の観測信号のパワースペクトルから求めることができる⁽⁶⁾．

$$|Z_1(l, f)|^2 = e^{-2\Delta T_{\text{shift}}} |X_1(l - N_{\text{shift}}, f)|^2 \quad (4)$$

ここで、 $N_{\text{shift}} = \lfloor T_{\text{shift}} f_s / \lambda \rfloor$ であり、 T_{shift} は、クリーン音声信号が定常とみなせる区間を T_s とした場合、 $T_s \leq T_{\text{shift}} \ll \text{RT}_{60}$ となるように設定され、通常は 50ms 程度に設定される． f_s はサンプリング周波数、 λ は短時間フーリエ変換 (STFT) に用いるフレームシフト量 (サンプル) である．上記のように得られた後部残響のパワースペクトルを、観測信号から減算し、

図3 統計的室内インパルス応答モデルに基づく方法の処理ブロック図

最終的な信号を得る^(注1):

$$|\hat{S}(l, f)|^2 = \begin{cases} |X_1(l, f)|^2 - |Z_1(l, f)|^2, \\ \text{(if } |X_1(l, f)|^2 - |Z_1(l, f)|^2 \geq 0) \\ 0, \text{ otherwise} \end{cases} \quad (5)$$

上記 Lebart らの方法は、全ての周波数帯域で同一の残響時間を用いている。また、直接音対間接音比が 0dB 以上の場合（音源とマイクロホンが近接している場合）、式 (3) の残響モデルの精度が悪くなり、残響除去後の音声に含まれるひずみが大きくなるという問題があった。それらの問題を解決するため、直接音対間接音比と、周波数ごとの残響時間を設定できるように式 (3) に改良を加えた研究例もある⁽⁷⁾。

いずれの方法を用いた場合でも、実際には、残響時間 RT_{60} や直接音対間接音比などを、何らかの方法^(注2)を用いて、観測信号からブラインドで求める必要がある。Löllmann らは、最よう推定された残響時間を用いて上記残響除去処理を行う方法を提案しており、残響時間推定処理を含めた残響除去処理の性能の評価を行っている⁽⁸⁾。実験の結果、バブル雑音 (SN 比:5dB) のある残響環境 (RT_{60} :0.9 秒) 下で、 ± 0.2 秒程度の推定誤差内で残響時間を推定可能であり、この程度の推定誤差であれば、残響時間を既知とした場合と同程度の性能が得られることが示されている。また、文献 (8) では残響除去に加え、雑音推定・除去も同時になされており、客観評価、主観評価の上で良好な残響雑音除去が達成されていることが報告されている。

Habets らは、残響除去処理とエコーキャンセラを組み合わせ、エコーキャンセラにより同定されたエコー経路から近端側の部屋の残響時間と直接音対間接音比を求め、近端話者の音声に重畳する残響除去に利用している⁽⁹⁾。

一方、吉岡らは、フレーム l の後部残響の短時間パワースペクトルを求めるために、式 (4) の $e^{-2\Delta T_{\text{shift}}}$ （予測利得と呼ぶ）を用いずに、予測利得そのものを観測信号から直接推定する方法を提案している⁽¹⁰⁾。音声に残響が重畳することで音声信号のスパース性が失われることに着目し、スパース性を最大化する

予測利得を求める。この方法は、残響時間や直接音対間接音比の推定を行うことなく、予測利得を求めることができるという利点を持つ。また、本方法で求めた予測利得から、周波数ごとに異なる残響時間を推定することも可能となる。

以上までに挙げた統計的残響モデルを用いる方法は推定すべきパラメータが少ないため、少ない計算量で頑健に動作する一方、残響のモデル化精度が粗く、高精度な残響除去を達成するには至っていない。

4.2 時間包絡フィルタリングに基づく方法

図4に、時間包絡フィルタリングに基づく方法（時間包絡フィルタリング法）の一般的なブロック図を示す。始めに、観測信号をフィルタバンク分析で各周波数帯域の信号に分割し、各帯域のパワー時間包絡を抽出する。次に、クリーン音声のパワー包絡を回復するためにフィルタ処理等を行う。最後に、パワー包絡を抽出した際に分離した情報（微細構造や位相成分）と回復したパワー包絡を合成し、時間領域に変換した後、最終的な回復音を得る。

時間包絡フィルタリング法の多くは、残響による音声のパワー包絡の変化と音声明瞭度の低下の關係に着目し⁽¹¹⁾研究が進められている。

広林らは、パワー包絡の逆フィルタ処理による残響除去を提案している⁽¹²⁾。始めに、音声信号 $s(t)$ をその包絡 $e_s(t)$ と白色のキャリア信号 $z_1(t)$ とを用いて以下のように表す。

$$s(t) \triangleq e_s(t)z_1(t) \quad (6)$$

また、室内インパルス応答は式 (3) のモデルを用いる。これらのモデルを用いると、観測信号 $x_1(t)$ のパワー包絡の期待値 $\langle (x_1(t))^2 \rangle$ は、以下のように、室内インパルス応答のパワー包絡 $E_h(t)$ と音声信号のパワー包絡 $E_s(t)$ の畳み込みで表すことができる。 $\langle \cdot \rangle$ はアンサンブル平均を表す。

$$\begin{aligned} \langle (x_1(t))^2 \rangle &= \langle (h_1(t) * s(t))^2 \rangle \\ &\simeq \langle (e^{-\Delta t})^2 * (e_s(t))^2 \rangle \end{aligned} \quad (7)$$

$$= E_h(t) * E_s(t) \quad (8)$$

つまり、観測信号のパワー包絡に室内インパルス応答のパワー包

(注1): 実際の処理では、直接的な引き算を用いずに、スペクトル抑圧係数を求め、それを観測信号 $X(l, f)$ に乗算し、処理を行う。ここでは、簡単のため、単純な減算処理を記載した。

(注2): 文献 (6) では、入力発話末尾の減衰特性から残響時間を算出している。

図 4 時間包絡フィルタリングに基づく処理のブロック図

図 5 特徴量領域での残響除去処理のブロック図

絡の逆特性を畳み込めば，原音声のパワー包絡を回復でき，明瞭性が向上することが期待される⁽¹¹⁾．式 (3) のモデルのパワー包絡 $E_h(t)$ の逆特性は z 領域で以下のように求めることができる⁽¹²⁾，(注3)：

$$\frac{1}{E_h}(z) = \frac{1}{\sigma_a^2} (1 - e^{\frac{-2\Delta}{f_s}} z^{-1}) \quad (9)$$

鵜木らは，この逆フィルタに用いる室内インパルス応答のパラメータ σ_a と残響時間 RT_{60} を観測信号から求める方法を提案し，本処理を音声認識の前処理に用いている⁽¹³⁾．回復したパワー包絡から，音声認識でしばしば用いられる特徴量である MFCC (Mel-Frequency Cepstral Coefficient) 係数を直接計算し，音声認識実験を行った結果，残響時間 1 秒程度の環境で，ケプストラム平均値正規化 (CMS)⁽¹⁴⁾ のみを行った場合に対して，5% 程度認識率を改善できることが示された．

一方，亀岡らは非負行列分解 (NMF)⁽¹⁵⁾ と同様のアプローチを用いて，観測信号のパワー包絡を音声のパワー包絡と室内インパルス応答のパワー包絡に分解することで，音声のパワー包絡を抽出する方法を提案している⁽¹⁶⁾．この方法は，室内インパルス応答の包絡の指数減衰を仮定せずに，その包絡を直接モデル化しており，柔軟な推定を可能としている．

4.3 特徴量領域での残響除去

自動音声認識と残響除去処理の親和性を高めるため，MFCC を抽出する前段階で用いられる，対数メルパワースペクトルの領域で残響除去処理を行う方法が提案されている⁽¹⁷⁾，(18)．図 5 に，その一例⁽¹⁷⁾ の処理ブロック図を示した．

始めに，対数メルパワースペクトル領域での観測信号のモデルを示す．観測信号の対数メルパワースペクトルを $X_{\text{mel}}(l, f')$ とすると， $X_{\text{mel}}(l, f')$ は以下のようにモデル化される．

$$X_{\text{mel}}(l, f') = \log \sum_{i=0}^{K_{\text{mel}}-1} e^{S_{\text{mel}}(l-i, f') + H_{\text{mel}}(i, f')} \quad (10)$$

ここで， $S_{\text{mel}}(l, f')$ はクリーン音声の対数メルパワースペクトル， $H_{\text{mel}}(l, f')$ は室内インパルス応答の対数メルパワースペクトルに対応する．また， f' はメル周波数帯域のインデックス， K_{mel} は対数メルパワースペクトル領域でのインパルス応答長であり， $K_{\text{mel}} \simeq K/\lambda$ である． K は，時間領域でのインパルス応答長， λ は STFT に用いるフレームシフト量である．

現在までに提案されている方法の多くは，事前に学習された $S_{\text{mel}}(l, f')$ と $H_{\text{mel}}(l, f')$ の統計モデルを用いて，観測の残響音声からクリーン音声を推定している (17) では， $e^{H_{\text{mel}}(l, f')}$ の各 i, f' 成分は独立のガウス分布に従うという仮定の基に $H_{\text{mel}}(l, f')$ のモデル化を行い，事前に測定した幾つかのインパルス応答からそのモデルパラメータを決定している．また， $S_{\text{mel}}(l, f')$ の統計モデルとしては，隠れマルコフモデル (HMM) を用いており，そのパラメータも事前学習により決定されている．これら $H_{\text{mel}}(l, f')$ と $S_{\text{mel}}(l, f')$ の統計モデルと式 (10) を用いモデルの合成 (図 5 中，モデル合成) を行い，合成後のゆわ度が最も高くなる $S_{\text{mel}}(l, f')$ を残響除去信号の推定値としている．式 (10) にあるように，現在の観測信号 $X_{\text{mel}}(l, f')$ は，現在のクリーン音声 $S_{\text{mel}}(l, f')$ だけでなく，それよりも前のクリーン音声 $S_{\text{mel}}(l-1, f'), \dots, S_{\text{mel}}(l-K_{\text{mel}}, f')$ にも依存している．そのため，現フレームで推定されたクリーン音声 $\hat{S}_{\text{mel}}(l, f')$ は次フレーム ($l+1$) 以降の計算のために，図 5 のようにモデル合成部にフィードバックされる．実験では，本方法を用いることで，連続数字認識の精度を大きく改善することが示されている．

(注3)： λ_s をサンプリング周期とすれば，

$$\begin{aligned} E_h(z) &= \sigma_a^2 + \sigma_a^2 e^{-2\Delta T_s} z^{-1} + \sigma_a^2 - e^{-4\Delta T_s} z^{-2} + \dots \\ &= \frac{\sigma_a^2}{1 - e^{-2\Delta T_s} z^{-1}} \end{aligned}$$

図 6 伝達関数推定を介した逆フィルタ設計に基づく処理の処理ブロック図

ている。

5. マルチチャネル処理

本章では、マルチチャネル残響除去処理、すなわち $1 \sim M (M \geq 2)$ 番目のマイクロホンで観測される全ての信号に基づき行われる処理についてまとめる。特に、指向性を制御し、残響を抑圧するビームフォーマと、マイクと音源間の室内伝達関数の影響を直接除去することを目的とした逆フィルタリング処理について述べる。逆フィルタの設計方法は、室内伝達関数をブラインド推定した後に逆フィルタを構成する方法と、観測信号から逆フィルタを直接推定する方法に分け、それぞれ概説する。一般的に、マルチチャネル処理を用いれば、室内インパルス応答の有限長・因果的な逆フィルタを構成することが可能となる。モデル化の精度が高くなるため、モデルにそぐわない外乱等に対してぜい弱になるという欠点を持つことがあるものの、実効的にはシングルチャネル処理よりも精度の高い残響除去を行うことが可能な場合が多い。

5.1 ビームフォーマ

ビームフォーマは、目的音の到来方向を推定し、その方向から到来する信号を強調したり、その方向以外から到来する信号を抑圧することで、目的音の強調を行う処理である。

残響環境で現在最も広く使われている音声強調処理は、遅延和ビームフォーマである⁽¹⁹⁾。遅延和ビームフォーマは、大規模なマイクロホンアレー（100 素子程度）の使用が可能な場合は、高い残響抑圧効果をもたらす。始めに、目的音声の到来方向を求め、直接音が同時刻に並ぶよう各マイク信号間の遅延を調整し、最後に全てのマイク信号を足し合わせることで相対的に直接音を強調する。遅延和ビームフォーマは、計算量が少なく、実装も容易である。しかし、より良い残響抑圧性能を得るためにはマイクロホンアレーの素子数を増やす必要があるものの、素子数を倍にしても抑圧量の増加は 3dB 程度^(注4)となるため、必然的に大規模な装置が必要となってしまう、導入は必ずしも容易ではない。例えば文献(20)では、63 個のマイク素子を利用し、一辺 1 m の二次元マイクロホンアレーを用いた処理を行っ

5.2 伝達関数推定を介した逆フィルタ設計に基づく処理

ここでは、室内インパルス応答を観測信号から推定し、推定した室内伝達関数を基に multiple-input/output inverse theorem (MINT) 法⁽⁵⁾等を用いて逆フィルタを構成する方法について概説する。図 6 に、処理ブロック図を示した。

MINT 型の逆フィルタ $\mathbf{g}$ とは、以下の式を満たす逆フィルタである。

$$\mathbf{b} = \mathbf{H}\mathbf{g} \quad (11)$$

ここで

$$\begin{aligned} \mathbf{b} &= [1, 0, \dots, 0]^T \\ \mathbf{g}_m &= [g_m(0), g_m(1), \dots, g_m(L-1)]^T \\ \mathbf{g} &= [\mathbf{g}_1^T, \mathbf{g}_2^T, \dots, \mathbf{g}_M^T]^T \\ &\quad \leftarrow \quad L \quad \rightarrow \\ \mathbf{H}_m &= \begin{pmatrix} \mathbf{h}_m & & 0 \\ & \mathbf{h}_m & \\ & & \ddots \\ 0 & & & \mathbf{h}_m \end{pmatrix} \quad \begin{matrix} \uparrow \\ L+K-1 \\ \downarrow \end{matrix} \\ \mathbf{H} &= [\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_M] \end{aligned}$$

$\mathbf{h}_m$ は、音源と m 番目のマイクロホン間の室内伝達関数のベクトルを表す。室内伝達関数がマイクロホン間で共通の零点を持たず、条件 $L \geq (K-1)/(M-1)$ が満たされれば、MINT 型逆フィルタ $\mathbf{g}$ の解は存在し、以下のように求めることができる。

$$\mathbf{g} = \mathbf{H}^+ \mathbf{b} \quad (12)$$

$\mathbf{A}^+$ は、行列 $\mathbf{A}$ の Moore-Penrose 擬似逆行列を表す。室内インパルス応答 $\mathbf{H}$ が既知の場合、上式で求めた逆フィルタ $\mathbf{g}$ を用いることで、観測信号から残響の影響を完全に取り除くことができる。

ブラインド残響除去の枠組みの中で MINT 型逆フィルタを求めるためには、室内伝達関数行列 $\mathbf{H}$ を観測信号から推定しなければならない。室内伝達関数 $\mathbf{H}$ の推定方法としては、例えば、観測信号の相関行列の零空間を用いて伝達関数を推定する

(注4)：残響成分を、各マイクロホン間で相関のない雑音とみなした場合の抑圧量の改善量

図 7 せん度最大化に基づく方法の処理ブロック図

Subspace 法が提案されている^{(21),(22)}. この方法は、通信チャネルのマルチパス問題を解決するために提案されたものであり⁽²¹⁾, その後に音声の信号処理にも適用され精緻な伝達関数推定が可能であることが示された⁽²²⁾. しかし、相関行列の零空間を求めるための一般化固有値分解 (GED) を行うためには、室内インパルス応答の応答長 K を正確に知る必要があり、応答長の推定に誤差が含まれる場合、残響抑圧性能は大幅に劣化してしまう⁽²³⁾. また、観測信号に少量の雑音が含まれる場合も (SN 比:20dB 程度)、大幅に性能が低下するため、実環境への適用は困難である.

5.3 直接的な逆フィルタ推定に基づく処理

5.3.1 せん度最大化に基づく方法

Gillespie らは、残響音声とクリーン音声の短時間フレームごとの線形予測残差の系列のせん度を比較した場合、クリーン音声のそれに対して残響音声のそれが低くなることを実験的に示し、線形予測残差のせん度を最大化する線形フィルタを推定し、観測音に適用することで残響除去を達成する方法を提案した⁽²⁴⁾. 図 7 に、処理ブロック図を示す. 処理により最大化したい線形予測残差のせん度 $J(t)$ は、線形予測残差に線形フィルタを畳み込んだ信号を $\tilde{x}(t)$ とすると以下の式で定義される.

$$J(t) = \frac{E\{\tilde{x}^4(t)\}}{E\{\tilde{x}^2(t)\}^2} - 3 \quad (13)$$

上記関数は解析的に最大化することができないため、勾配法を用いて最大化されている. 文献⁽²⁴⁾では、定量的な評価はなされていないが、インパルス応答の等化により残響除去の効果を見て取ることができる. 文献⁽²⁵⁾では、この方法を収録音に始めに適用し初期反射音を低減し、その後、統計的室内インパルス応答モデルに基づく方法を用いて後部残響を抑圧するような 2 段階構えの手法も提案されている. また、文献⁽²⁶⁾では、線形予測残差と残響除去フィルタを交互に繰り返し推定することで、残響除去性能が改善されることが示されている.

せん度に代表される、高次 (三次以上) 統計量を用いることで、非最小位相伝達関数の推定や等化が原理的には可能となるため、シングルチャネルでも逆フィルタ処理に基づく残響除去が可能となる. 一方、高次統計量を精度良く推定するには多くの観測データが必要となるため、時々刻々と変化する室内音響特性への追従性が課題となる.

5.3.2 MINT 型逆フィルタに基づく方法

古家らは、MINT 型逆フィルタを観測信号の二次統計量からブラインドで直接計算する方法を提案している⁽²⁷⁾. 図 8 に処理ブロック図を示す. この方法では始めに、観測信号 $x_1(n)$ の平均振幅スペクトルを用いて音源信号 $s(n)$ に起因する相関 (= 式 (2) の $\alpha(n)$ に起因する相関) を観測信号から取り除く白色化フィルタを計算する^(注5). この白色化フィルタを用いて、観測信号 $x_m(n)$ ($m = 1, 2, \dots, M$) から $s(n)$ に起因する相関を消去することで、擬似的に式 (1) 中の音源信号 $s(n)$ を白色信号とみなすことができるとしている. 次に、観測信号 $x_m(n)$ を白色化した信号 $\tilde{x}_m(n)$ を用いて、相関行列を算出する.

$$\begin{aligned} \mathbf{R} &= E\{\tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}(n)^T\} \\ \tilde{\mathbf{x}}(n) &= [\tilde{x}_1(n)^T, \tilde{x}_2(n)^T, \dots, \tilde{x}_M(n)^T]^T \\ \tilde{\mathbf{x}}_m(n) &= [\tilde{x}_m(n), \tilde{x}_m(n-1), \dots, \tilde{x}_m(n-L+1)]^T \end{aligned}$$

ここで求めた相関関数は、白色化の効果により、以下のようにみなすことができると考える.

$$\mathbf{R} = E\{\tilde{\mathbf{x}}(n)\tilde{\mathbf{x}}(n)^T\} \simeq \mathbf{H}^T \mathbf{H} \quad (14)$$

MINT 型逆フィルタの定義である式 (11) の両辺に $\mathbf{H}^T$ をかけると、以下の式を得ることができ、

$$\mathbf{H}^T \mathbf{b} = \mathbf{H}^T \mathbf{H} \mathbf{g}$$

この式に式 (14) を代入すると、以下のように MINT 型逆フィルタを観測信号の二次統計量のみから算出できることが分かる.

$$\mathbf{g} \simeq h_1(0) \mathbf{R}^{-1} \mathbf{b} \quad (15)$$

ここで 1 番目のマイクロホンと音源間のインパルス応答 h_1 の 0 タップ目 $h_1(0)$ は未知であるが、逆フィルタのスケールは残響除去効果とは直接関係しないため、任意の値 (例えば、 $h_1(0) = 1$) に設定すればよい.

この手法を用いて効果的な残響除去を達成する $\mathbf{g}$ を求めるためには、音源 $s(n)$ に起因する相関を精度良く除去する白色化フィルタ $W(f)$ が必要であり、この白色化フィルタを高い精度で推定するためにはある程度長い観測信号 (例えば、文献 (27) では 20 秒程度の観測信号を要している) が必要である. また、式 (15) 中 $\mathbf{b}$ の設計のためには、 M 個のマイクロホンの中で音

(注5): 効果的な白色化を達成するためには、標準化周波数 12kHz の場合、 $W(f)$ には 500 タップ程度必要.

図 8 MINT 型逆フィルタに基づく方法の処理ブロック図

源に一番近いマイクロホンを指定する必要があるが、式 (15) では 1 番目のマイクロホンが音源に一番近いとしている^(注6)。文献 (27) では、 g を用いた逆フィルタリング処理の後に残留した残響を除去するため、スペクトル減算を用いた後処理も提案されており、この後処理も組み合わせた上で、精度の良い残響除去が達成できることが示されている。逆フィルタ計算にサブバンド処理を導入し、性能を改善する試みもなされている⁽²⁸⁾。

5.3.3 マルチチャネル線形予測を用いた方法

線形予測は、予測係数を介して逆フィルタを構築する手法であり、逆システム推定には効果的な手段である。特に、マルチチャネル線形予測は、シングルチャネルの線形予測とは異なり、対象システムの伝達関数が非最小位相系の場合でも逆フィルタをブラインド推定することができるという利点を有している。本節の議論には、木下らによって提案された、通常のマルチチャネル線形予測を一般化したマルチチャネルマルチステップ線形予測⁽²⁹⁾を用いる。図 9 に処理の一例を示す。

マルチステップ線形予測は、観測信号に含まれる後部残響成分の抑圧に特化した手法である。始めに、本処理が抑圧を行う残響成分を明確に規定するため、マルチステップ線形予測内に導入される遅延（若しくはステップサイズ） D を用いて、式 (1) を以下のように書き換える。

$$x_m(n) = \sum_{k=0}^{D-1} h_m(k)s(n-k) + \sum_{k=D}^{K-1} h_m(k)s(n-k) \triangleq d_m(n) + r_m(n) \quad (16)$$

$d_m(n)$ は直接音に初期反射音加わったもので、 $r_m(n)$ は後部残響である。ここで、室内伝達関数が互いに共通する零点を持たなければ、式 (16) は、マルチチャネルマルチステップ線形予測を用いて、十分に大きな線形予測次数 L に対して以下のように表現することが可能であることが示されている^(注7)。

$$x_m(n) = \sum_{i=1}^M \sum_{k=0}^{L-1} w_{m,i}(k)x_i(n-D-k) + d_m(n) \quad (17)$$

(注6): 文献 (27) では、音源に一番近いマイクの情報は既知としているが、ブラインド推定を行うか、他チャネルの信号を遅延させることで人工的に操作することもできる。

(注7): $D = 1$ の場合、式は通常のマルチチャネル線形予測と等価。

$w_{m,i}$ は予測係数である。式 (17) から分かるように、理想的な予測係数 $w_{m,i}$ が求まれば、マルチチャネルマルチステップ線形予測では、 m 番目のマイクロホンで観測された現在の信号 $x_m(n)$ は、現在よりも D サンプル前までの過去の信号の重み付き和（第一項）と、後部残響を含まない信号 $d_m(n)$ （第二項）の加算で表される。そして、推定した予測係数 $w_{m,i}(k)$ を用いた逆フィルタリング処理は、以下を行う。

$$\hat{s}_1(n) = x_1(n) - \sum_{i=1}^M \sum_{k=0}^{L-1} w_{1,i}(k)x_i(n-D-k) \quad (18)$$

式 (17) と上式を比べれば、 $\hat{s}_1$ が $d_1(n)$ の推定値になっていることが分かる。ここでは簡単のため、1 番目のマイクロホンで観測された信号 $x_1(n)$ に含まれる後部残響を除去するための逆フィルタリング処理を例示した。また、ここで上記の線形逆フィルタリング処理よりも予測係数推定誤差等に頑健な処理を行うことのできるスペクトル減算を用いる方法も提案されている⁽²⁹⁾。この処理を行うには、始めに観測信号 $x_1(n)$ と、予測された後部残響成分 $\sum_{i=1}^M \sum_{k=0}^{L-1} w_{1,i}(k)x_i(n-D-k)$ （式 (17) の第一項）のそれぞれを短時間フレームに分割し、STFT を用いてフレーム l でのそれぞれのパワースペクトル、 $|X_1(l, f)|^2$ 、 $|Z_1(l, f)|^2$ を求める。その後、式 (5) のように観測信号から残響成分を減算し、残響除去を達成する。

いずれの処理を行う場合も、予測係数 $w_{m,i}$ をより正確に推定することが重要であり、予測係数 $w_{m,i}$ の推定精度が高くなれば、 $d_1(n)$ により近い信号を回復することができる。文献 (29) では、従来の線形予測と同様に、予測残差の平均二乗エネルギーを最小にするように予測係数を求めている：

$$\mathbf{w}_1 = E\{\mathbf{x}(n-D)\mathbf{x}(n-D)^T\} + E\{\mathbf{x}(n-D)\mathbf{x}_1(n)\} \quad (19)$$

$$\mathbf{x}(n) = [\mathbf{x}_1(n)^T, \mathbf{x}_2(n)^T, \dots, \mathbf{x}_M(n)^T]^T$$

$$\mathbf{x}_m(n) = [x_m(n), x_m(n-1), \dots, x_m(n-L+1)]^T$$

$$\mathbf{w}_m = [\mathbf{w}_{m,1}^T, \mathbf{w}_{m,2}^T, \dots, \mathbf{w}_{m,M}^T]^T$$

$$\mathbf{w}_{m,i} = [w_{m,i}(0), w_{m,i}(1), \dots, w_{m,i}(L-1)]^T$$

ここで、音源信号 $s(n)$ が白色信号であれば、上式で求めた $\mathbf{w}_m$ は、 $x_m(n)$ に含まれる後部残響を（期待値の意味で）完全に除去

図9 線形予測を用いた方法の処理ブロック図

することのできる予測係数になる。また、音源信号 $s(n)$ について以下の条件が成立する場合、線形予測内の遅延 D を $D > T_s$ と設定すれば、 $x_m(n)$ に含まれる後部残響を高精度に推定・除去することのできる予測係数になる⁽³⁰⁾。

$$E\{s(n)s(n'')\} = 0 \quad \text{if } |n - n''| > T_s$$

文献(29)では、2~3秒程度の学習データを用いて式(19)のように予測係数を求め、スペクトル減算を用いて残響除去を行うことで、大語彙自動音声認識の性能改善が可能であることが示されている。また、NISTのデータ(RT-06s)を用いた試験でも類似の改善が報告されている⁽³¹⁾。

より精度の良い予測係数 $w_{m,i}$ を得るため、予測係数推定処理に確率的定式化を導入する試みがなされている^{(32),(33)} (注8)。確率的定式化を行うことで、所望信号(=残響除去信号)の統計的性質をより正確に捉える確率的音源モデルを導入することができる。音源信号の短時間フレーム信号 $\mathbf{s}(n) = [s(n), s(n-1), \dots, s(n-N_{\text{frame}}+1)]^T$ の確率モデルは以下の確率密度関数 $p_s(\mathbf{s}(n))$ として与えられ、具体的には平均0、共分散 $\mathfrak{R}_n$ の N_{frame} 次元の多変量ガウス過程で表現される。

$$p_s(\mathbf{s}(n)) = \mathcal{N}(\mathbf{s}(n); 0, \mathfrak{R}_n)$$

上記モデルを用いることで、通常の線形予測と異なり、所望信号を時変有色信号としてモデル化することができる。ここで、共分散 $\mathfrak{R}_n$ は異なるフレームにおいて異なる値を取るものと仮定している。上記所望信号の確率モデルを予測係数推定に導入するために、以下の対数尤度関数を用いる。 $\mathfrak{R}_n$ は事前に与えることができないため、尤度関数の中では推定すべきパラメータ R_n として扱う。

$$\mathcal{L}(\theta) = \sum_n \log(p_s(\mathbf{d}_1(n); \theta)) \quad (20)$$

ここで、

$$\theta = (\mathbf{w}_1, R), \quad R = [R_1, R_2, \dots]$$

$$\mathbf{d}_1(n) = [d_1(n), d_1(n-1), \dots, d_1(n-N_{\text{frame}}+1)]^T$$

なお、 $\mathbf{d}_1(n)$ は式(17)のように、観測信号 $x_m(n)$ と予測係数

$w_{m,i}$ から計算される。 R_n の推定に対する拘束の与え方としては、音源が時変の自己回帰過程に従うとする方法⁽²⁶⁾や、音源の事前知識として自己相関コードブックを用いる方法⁽³⁴⁾等が提案されている。上記のように残響除去後の信号をより正確に表すモデルを導入し予測係数を算出することで、式(19)よりも精度の良い予測係数、逆フィルタを得ることのできるということが報告されている。

また、上記の計算をより効率的に行うために、線形予測をSTFT領域の処理に拡張し、各サブバンドごとに線形予測係数を求める方法も提案されている⁽³⁵⁾。STFT領域で処理を行うことで、より効率的な計算が可能となるだけでなく、音源分離や雑音除去との統合も容易になる。音源分離と残響除去の統合⁽³⁶⁾、雑音除去と残響除去の統合⁽³⁷⁾が既に達成されており、実音場での高精度な残響除去が可能であることが示されている。

図10に、(a) クリーン音声、(b) 残響の重畳した音声、(c) 残響除去音声のスペクトログラムを示した。ここでは、マイクロホン数は2とし、文献(35)で提案されている方法で残響除去を行った。例えば、(a)のクリーン音声では調波構造や明確な音声終端を見て取ることができるが、(b)では残響によりそれらの特徴が見えにくくなっている。それに対し、(c)では、残響除去を行うことでそれらの特徴が効果的に回復されていることが分かる。

6. まとめ

実環境でマイクロホン収音された音声には、しばしば残響による音響的なひずみが含まれる。この残響は収音した音声信号の明瞭性を低下させるとともに、コンピュータによる自動音声認識をはじめとする多くの音響信号処理アプリケーションの性能低下を招く原因となることが知られている。実環境においてこれら音響信号処理アプリケーションの性能・精度の向上を図るには、収音音声から残響による音響ひずみの影響を取り除く技術(ブラインド残響除去技術)の確立が重要である。

本稿では、近年提案された音声のブラインド残響除去処理の幾つかを挙げ、それらの特徴をまとめた。音声のブラインド残響除去は従来困難な問題であったが、近年の成果により、ある程度静的な実音響環境での高精度な残響除去は可能となった。今後、実環境への残響除去の適用を進めるには、シングルチャネル処理でのより精度の高い残響除去技術の提案、少量データ(1

(注8): 幾つかの文献では $D=1$ としているが、 $D=1$ のときと $D>1$ の場合で同様の音源モデルを用いることが可能であることが実験的に分かっている。

図 10 スペクトログラム : (a) クリーン音声 , (b) 残響の重畳した音声 , (c) 残響除去音声 (文献 (35) の手法を用いて処理)

秒以下)での高精度な残響除去,他の音声強調処理との統合などが課題となると思われる。

文 献

- (1) S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Speech Audio Process., vol.27, no.2, pp.113–120, 1979.
- (2) M. Fujimoto and S. Nakamura, "A non-stationary noise suppression method based on particle filtering and polyak averaging," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.922–930, March 2006.
- (3) 三好正人, 中谷智広, 向井 良, 澤田 宏, 引地孝文, 荒木章子, 木下慶介, "ブラインド信号処理技術の研究動向," 信学技報, EA2004-21, pp.23–30, 2004.
- (4) E. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," PhD thesis, Technische Universiteit Eindhoven, 2007.
- (5) M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE Trans. Speech Audio Process., vol.36, no.2, pp.145–152, 1988.
- (6) K. Lebart and J.M. Boucher, "A new method based on spectral subtraction for speech dereverberation," Acta Acoust., vol.87, pp.359–366, 2001.
- (7) E.A.P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol.5, pp.173–176, 2005.
- (8) H.W. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," EURASIP J. advances in signal processing, 437807, 2009.
- (9) E.A.P. Habets, S. Gannot, I. Cohen, and P.C.W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," IEEE Trans. Audio, Speech, and Language Process., vol.16, no.8, pp.1433–1451, 2008.
- (10) 吉岡拓也, 亀岡弘和, 中谷智広, 奥乃 博, "少量データに頑健な残響抑圧のための MSPP 法," 日本音響学会秋季研究発表会, pp.651–654, 2009.
- (11) T. Houtgast and H. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am., vol.77, no.3, pp.1069–1077, 1985.
- (12) 広林茂樹, 野村博昭, 小池恒彦, 東山三樹夫, "パワーエンベロープ伝達関数の逆フィルタ処理による残響音声の回復," 信学論 (A), vol.1, J81-A, no.10, pp.1323–1330, Oct. 1998.
- (13) L. Xugang, M. Unoki, M. Furukawa, and M. Akagi, "Comparative evaluation of modulation-transfer-function-based blind restoration of sub-band power envelopes of speech as a front-end processor for automatic speech recognition systems," Acoust. Sci. Technol., vol.29, no.6, pp.351–361, 2008.
- (14) B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., vol.55, no.6, pp.1304–1312, 1974.
- (15) D.D. Lee and H.S. Seung, "Learning the parts of objects by nonnegative matrix factorization," Nature, vol.401, pp.788–791, 1999.
- (16) H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp.45–48, 2009.
- (17) A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," IEEE Trans. Audio, Speech, and Language Process., vol.18, no.7, pp.1676–1691, 2010.
- (18) A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," IEEE Trans. Audio, Speech, and Language Process., vol.18, no.7, pp.1692–1707, 2010.
- (19) J.L. Flanagan, "Computer-steered microphone arrays for sound transduction in large rooms," J. Acoust. Soc. Am., vol.78, no.11, pp.1508–1518, 1985.
- (20) J.L. Flanagan, A.D. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi, "Autodirective microphone systems," Acustica, vol.73, pp.58–71, 1991.
- (21) M.I. Gurelli and C.L. Nikias, "EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals," IEEE Trans. Signal Process., vol.43, no.1, pp.134–149, 1995.
- (22) S. Gannot and M. Moonen, "Subspace methods for multi microphone speech dereverberation," EURASIP J. Applied Signal Process., vol.2003, no.11, pp.1074–1090, 2003.
- (23) K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: Experimental validation," EURASIP J. Audio, Speech, and Music Process., 51831, 2007.
- (24) B.W. Gillespie, H.S. Malvar, and D.A.F. Florncio, "Speech dereverberation via maximum-kurtosis sub-band adaptive filtering," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol.1, pp.3701–3704, 2001.
- (25) M. Wu and D.L. Wang, "A one-microphone algorithm for reverberant speech enhancement," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol.1, pp.844–847, 2003.
- (26) T. Yoshioka, T. Hikichi, and M. Miyoshi, "Deverberation by using time-variant nature of speech production system," EURASIP J. Advances in Signal Processing, 65698, 2007.
- (27) K. Furuya and A. Kataoka, "Robust speech derever-

beration using multichannel blind deconvolution with spectral subtraction,” IEEE Trans. Audio, Speech, and Language Process., vol.15, no.5, pp.1579–1591, 2007.

- (28) O. Hazrati, K. Kokkinakis, and P.C. Loizou, “A blind subband-based dereverberation algorithm,” Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol.1, pp.4714–4717, 2010.
- (29) K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction,” IEEE Trans. Audio, Speech, and Language Processing, vol.17, no.4, pp.534–545, 2009.
- (30) M. Miyoshi, M. Delcroix, and K. Kinoshita, “Calculating inverse filters for speech dereverberation,” IEICE Trans. Fundamentals, vol.E91-A, no.6, pp.1303–1309, June 2008.
- (31) M. Wölfel, “Enhanced speech features by single-channel joint compensation of noise and reverberation,” IEEE Trans. Audio, Speech, and Language Process., vol.17, no.2, pp.312–323, 2009.
- (32) 中谷智広, 吉岡拓也, 木下慶介, 三好正人, “時変ガウス音源モデルと多チャネル自己回帰観測モデルに基づく最ゆう法による音響信号の残響除去,” 信学論 (A), vol.E92-A, no.5, pp.294–304, May 2009.
- (33) T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech dereverberation based on variance-normalized delayed linear prediction,” IEEE Trans. Audio, Speech, and Language Process., vol.18, no.7, pp.1717–1731, 2010.
- (34) T. Nakatani, B.-H. Juang, T. Hikichi, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, “Study on speech dereverberation with autocorrelation codebook,” Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol.1, pp.193–196, 2007.
- (35) T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation,” Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol.1, pp.85–88, 2008.
- (36) T. Yoshioka, T. Nakatani, and M. Miyoshi, “An integrated method for blind separation and dereverberation of convolutive audio mixtures,” Proc. European Signal Processing Conference (EUSIPCO), 2008.
- (37) T. Yoshioka, T. Nakatani, and M. Miyoshi, “Integrated speech enhancement method using noise suppression and dereverberation,” IEEE Trans. Audio, Speech, and Language Process., vol.17, no.2, pp.231–246, 2009.

(EA 研究会提案, 平成 22 年 12 月 22 日受付
平成 23 年 1 月 26 日最終受付)

木下慶介 (正員)

平 13 上智大・理工・電子電子卒。平 22 同大学院博士課程了。平 15 日本電信電話株式会社入社。同社 NTT コミュニケーション科学基礎研究所にて、残響除去の研究に従事。博士 (工学)。平 17 年度本会論文賞, 平 22 年度日本音響学技術開発賞各受賞。

吉岡拓也

平 16 京大・工・情報卒。平 22 同大学院博士課程了。平 18 日本電信電話株式会社入社。NTT コミュニケーション科学基礎研究所にて、残響除去、音源分離、雑音抑圧の研究に従事。博士 (情報学)。平 22 日本音響学会栗屋潔学術奨励賞受賞。IEEE, 日本音響学会各会員。

中谷智広 (正員)

平 3 京大大学院工学研究科修士課程了。平 13 同大学院博士号取得。平 3 日本電信電話株式会社入社。以来、実環境における頑健な音声認識の実現に向けて音源分離・残響除去の研究に従事。現在、同社コミュニケーション科学基礎研究所主幹研究員。平 20 から名大客員准教授を兼任。博士 (情報学)。平 17 年度本会論文賞受賞。