音響尤度を用いた マルチスピーカ音響エコーキャンセラの検討*

○古賀健太郎, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

カーナビの音声操作など，音楽が鳴っている環境での音声認識では，音楽雑音がマイクで観測されるため観測信号の SN 比が悪くなり，音声認識率が低下する．そこで，観測信号の SN 比を改善する音響エコーキャンセラが必要になる．

車内の場合，音楽雑音のチャンネル数は 2ch 以上で，マルチスピーカから出力される．先行研究 [1] では，2ch の参照信号を 1ch にしてエコー推定を行っているが，複数のエコーパスをまとめて扱うエコー推定ではキャンセル結果が十分に収束しないことがある．そこで，[2]，[3] において，マルチスピーカからマイクまでの各エコーパスを独立に推定するような，音響尤度に基づくマルチスピーカ音響エコーキャンセラの検討を行い，学習同定法に基づく音響エコーキャンセラよりも観測信号の SN 比を改善できることを示した．

しかし，音響尤度に基づくマルチスピーカ音響エコーキャンセラは，推定する環境が多くなればなるほど計算に時間がかかる欠点がある．本稿では，[3] において，推定する環境を減らした場合のエコーキャンセル性能について調査を行う．

2 マルチスピーカ音響エコーキャンセラのモデル

マルチスピーカ（スピーカ数：4）からの音楽雑音が 1ch マイクで観測される音響エコーキャンセラのモデルを Fig. 1 に示す．

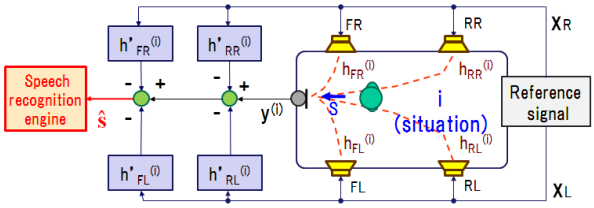

Fig. 1 マルチスピーカ音響エコーキャンセラの構成

環境 i において，マイクの時間領域における観測信号 $y^{(i)}$ は

$$y^{(i)} = s + N^{(i)} \tag{1}$$

と書ける．s は話者の音声である．$N^{(i)}$ は音響エコーで，

$$N^{(i)} = \sum x_L(h_{FL}^{(i)} + h_{RL}^{(i)}) + \sum x_R(h_{FR}^{(i)} + h_{RR}^{(i)}) \tag{2}$$

と書ける．x_L，x_R は 2ch の参照信号，$h_{FL}^{(i)}$，$h_{FR}^{(i)}$，$h_{RL}^{(i)}$，$h_{RR}^{(i)}$ は環境 i における各スピーカ (FL，FR，RL，RR) からマイクまでの伝達特性である．このとき，推定した音響エコーを $N'^{(i)}$ とすると，話者のクリーン音声 $\hat{s}^{(i)}$ は，

$$\hat{s}^{(i)} = y^{(i)} - N'^{(i)} \tag{3}$$

となる．$\hat{s}^{(i)}$ において，目的とする音声 s を残すため，推定誤差を最小にするように $N'^{(i)}$ は推定されるべきである．

3 尤度最大化に基づくエコー推定を用いたマルチスピーカ音響エコーキャンセラ

変化する環境 i に対し，推定したい数だけの環境 i $(i = 1, 2, \ldots, N)$ でインパルス応答を測定し [4]，各スピーカ-マイク間のエコーパスに対応した固定フィルタを $h'_{FL}{}^{(i)}$，$h'_{FR}{}^{(i)}$，$h'_{RL}{}^{(i)}$，$h'_{RR}{}^{(i)}$ $(i = 1, 2, \ldots, N)$ とする．

参照信号 x_L，x_R に，実環境 i $(i = 1, 2, \ldots, N)$ にて測定した固定フィルタをそれぞれ畳み込み，観測信号 $y^{(i)}$ からキャンセルして，クリーン音声候補 $\hat{s}^{(1)}, \hat{s}^{(2)}, \ldots, \hat{s}^{(N)}$ を算出する．

クリーン音声候補の中から，尤度最大のクリーン音声候補を選択し，クリーン音声とする．$\hat{s}^{(1)}, \hat{s}^{(2)}, \ldots, \hat{s}^{(N)}$ から，音声の MFCC 特徴量 $\hat{S}^{(1)}_{MFCC}, \hat{S}^{(2)}_{MFCC}, \ldots, \hat{S}^{(N)}_{MFCC}$ を計算する．MFCC は音声データに対し FFT を行い，パワー成分の対数を取った値を離散コサイン変換したものである．

MFCC 特徴 o の尤度 $P(o)$ は式 (4) の通り，W 個の重みつき正規分布の和として求められる．正規分布の w 番目の平均は μ_w，分散は σ_w である．また，λ_w は，$\sum_1^W \lambda_w = 1$ となる重み係数である．

$$P(o) = \sum_{w=1}^{W} \lambda_w N(o; \mu_w, \sigma_w) \tag{4}$$

各話者の音響モデルを $\psi = \{\lambda, \mu, \sigma\}$ としたとき

$$\hat{i} = \arg\max_i P(\hat{S}^{(i)}_{MFCC} | \psi) \tag{5}$$

となる $\hat{i}$ を計算し，このときの $\hat{s}^{(\hat{i})}$ が尤度最大のクリーン音声候補である．全体の構成図を Fig. 2 に示す．

4 実験条件

物の配置が異なる 8 通り環境 (Fig. 3) での学習同定法と提案手法の SN 比改善効果を，シミュレーション実験で示す．

観測信号は，話者数が 5，話者ごとの文章数が 20，周波数 16kHz である．音楽雑音は参照信号にインパルス応答を畳み込んだシミュレーション信号である．アルゴリズムのうち，学習同定法の適応フィルタのタップ長は 1200，また，2ch 参照信号を足し合わせて 1ch にした参照信号を適応フィルタの入力としている．提案手法の固定フィルタのタップ長は 1200，GMM 学習に用いた話者数は 1(特定話者)，文章数は 20，混合数は 32，MFCC の次元数は 16，特徴抽出時のフレーム幅は 32，シフト幅は 8(ms) としている．

5 推定環境を減らす検討

実験結果を Fig. 4 に示す．ベースライン 17.8(dB) に対し，学習同定法によるキャンセラを用いた場合が 15.8(dB)，尤度最大化基準に基づくキャンセラを用いた場合が 35.3(dB) で，提案手法の方が SN 比が改善されている．

---

*Multi-loudspeaker acoustic canceller using acoustic likelihood, by KOGA Kentaro, TAKIGUCHI Tetsuya, ARIKI Yasuo (Kobe University)

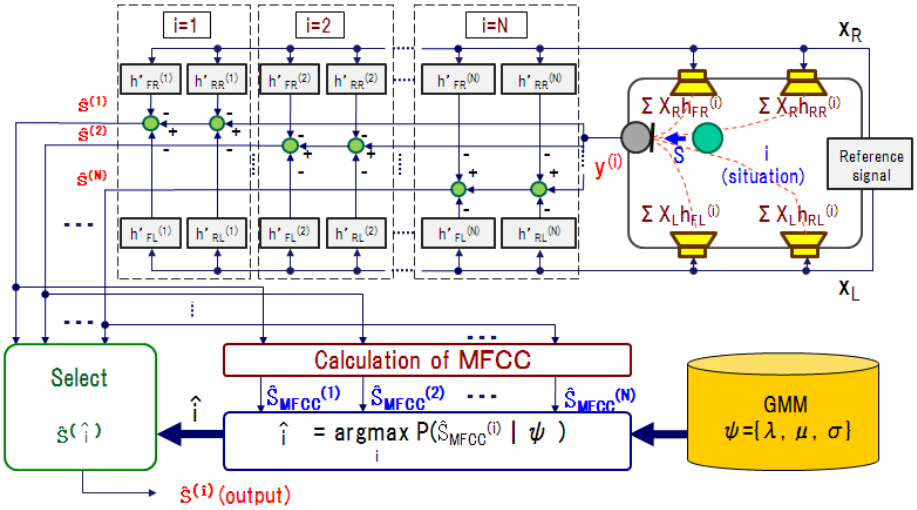

Fig. 2　尤度最大化基準を用いた車室内音響エコーキャンセラの構成図

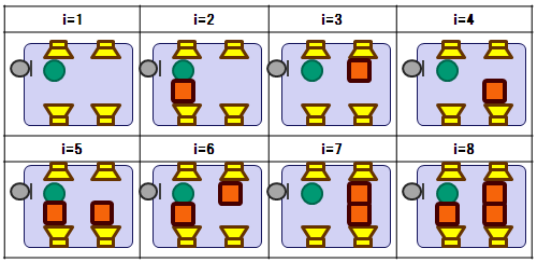

Fig. 3　物の配置が異なる環境 (8 通り)

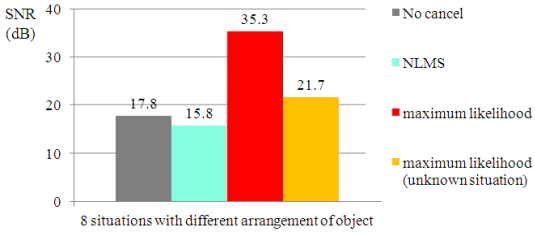

Fig. 4　実験結果

次に提案手法に対して，環境 o で観測した信号 $y^{(o)}$ に対して，同じ環境 o を推定した固定フィルタ $h'^{(o)}$ を用いなかった場合 (未知環境の場合) の SN 比を，Fig. 4 に併せて示す．同じ環境 o を推定した固定フィルタ $h'^{(o)}$ を用いている場合と比べて，21.7(dB) と，SN 比改善効果が小さい．

このとき，使用しなかった環境 o の代わりにどの環境 $\hat{o}$ が選択されているのかを調査した結果を Fig. 5 に示す．

o \ $\hat{o}$	1	2	3	4	5	6	7	8
1		**95%**			5%			
2	**85%**				5%	10%		
3		5%		30%	55%	10%		
4			25		**75%**			
5		10%	5%	**70%**		10%		5%
6	10%	5%						**85%**
7		5%	20%	5%		55%		15%
8	5%					**95%**		

Fig. 5　環境 o の代わりに選択される環境 $\hat{o}$ の調査結果

$o = 1$ の場合は 95% の確率で $h'^{(2)}$ が選択され，$o = 2$ の場合は 90% の確率で $h'^{(1)}$ が選択される．このことより，$h'^{(1)}$ と $h'^{(2)}$ は似た環境であると考えられる．同様に，$o = 4$ の場合は 75% の確率で $h'^{(5)}$ が選択され，$o = 5$ の場合は 70% の確率で $h'^{(4)}$ が選択される．$o = 6$ の場合は 85% の確率で $h'^{(8)}$ が選択されており，$o = 8$ の場合は 95% の確率で $h'^{(6)}$ が選択される．$h'^{(4)}$ と $h'^{(5)}$，$h'^{(6)}$ と $h'^{(8)}$ もそれぞれ似た環境であると考えられる．

以上より，$h'^{(1)}$ と $h'^{(2)}$，$h'^{(4)}$ と $h'^{(5)}$，$h'^{(6)}$ と $h'^{(8)}$ の推定結果はお互いをある程度補完できると考えられる．

Fig. 5 の結果より，$h'^{(1)}$ と $h'^{(2)}$ の片方，$h'^{(4)}$ と $h'^{(5)}$ の片方，$h'^{(6)}$ と $h'^{(8)}$ の片方を用いない，物の配置が異なる環境 5 通りを推定する場合を考える．

このような環境 5 通りの組み合わせは，$o = \{(1,3,4,6,7),(1,3,4,7,8),(1,3,5,6,7),(1,3,5,7,8),(2,3,4,6,7),(2,3,4,7,8),(2,3,5,6,7),(2,3,5,7,8)\}$ の合計 8 パターンである．観測信号に対する環境 5 通りの推定による音楽雑音キャンセル効果を SN 比で比較し，8 パターンの平均を取る．

実験結果を Fig. 6 に示す．環境 8 通りの場合と比較すると SN 比改善の度合いは下がるが，環境 5 通りの場合でも平均 30.8(dB) の改善が出来ることを実験において示すことが出来た．

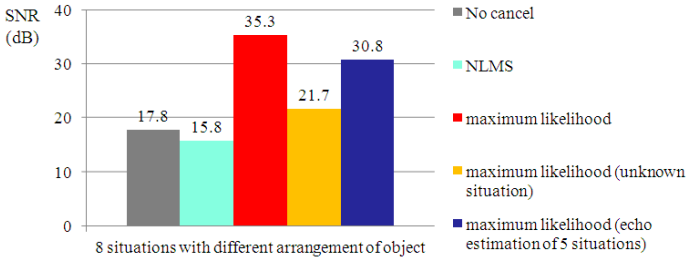

Fig. 6　実験結果 (推定環境を減らした場合含)

6　おわりに

本稿では，尤度最大化に基づく音響エコーキャンセラによって，推定環境を減らした場合のキャンセル性能についてシミュレーション実験を行った．今後は，データを増やしてより精密な実験を検討する．

参考文献

[1] S. Miyabe, T. Takatani, Y. Mori, H. Saruwatari, K. Shikano, and Y. Tatekura, “Double-Talk Free Spoken Dialogue Interface Combining Sound Field Control with Semi-Blind Source Separation”, ICASSP 2006.

[2] 古賀，2008 春季研究発表会，3-P-6

[3] 古賀，2011 秋季研究発表会，1-P-3

[4] 佐藤，日本音響学会誌 58 巻 10 号，pp.669-676，2002.