尤度最大化基準を用いたエコー推定に基づく車室内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 (富士通テン) , 滝口哲也，有木康雄 (神戸大)

1 はじめに

車内向け音声認識システムでは，車室内で音楽がスピーカから出力されている状況で音声認識を行うと，マイクに音響エコーが混入し，音声認識の妨げとなる．そこで，音響エコーをキャンセルし，SN を改善することによって認識率を確保する音響エコーキャンセラの開発を進めている．

車室内の音響エコーはマルチスピーカから出力され，単一マイクで観測される．このとき，音響エコーの基となる参照信号は通常 2ch 以上である．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．しかし，内装による車室内の音の反射は複雑化しているため，1ch 参照信号による推定ではキャンセル結果が十分に収束しないと考えられる．

そこで，[2] において，車室内マルチスピーカ環境で 2ch の参照信号を独立に用いて音響エコーを推定しキャンセルする方法を検討し，音響エコーをシミュレーションした信号による実験で，音響エコーをキャンセルできる目処付けを行った．本稿では，[2] で提案した音響エコーキャンセラにおいて，実環境で収録した信号による実験を行い，その結果音響エコーをキャンセルし，車室内における音声認識率を改善できることを示す．

2 車室内音響エコーキャンセラのモデル

マルチスピーカから出力され，単一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．

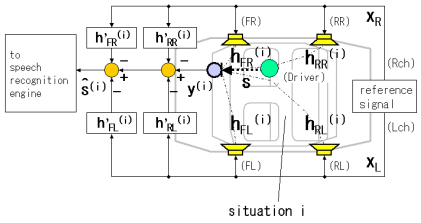

Fig. 1 車室内音響エコーキャンセラの構成

車内環境 i において，マイクの観測信号 $y^{(i)}$ は

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

と書ける．s はドライバの音声である．$N^{(i)}$ 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，$h_{FL}^{(i)}$，$h_{FR}^{(i)}$，$h_{RL}^{(i)}$，$h_{RR}^{(i)}$ はそれぞれ車内環境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，推定すべき音響エコー $N'^{(i)}$ は

$$N'^{(i)} = \sum x_L(h'^{(i)}_{FL} + h'^{(i)}_{RL}) + \sum x_R(h'^{(i)}_{FR} + h'^{(i)}_{RR}) \tag{3}$$

と書ける．よって，ドライバのクリーンな音声 $\hat{s}^{(i)}$ は，

$$\hat{s}^{(i)} = y^{(i)} - N'^{(i)} \tag{4}$$

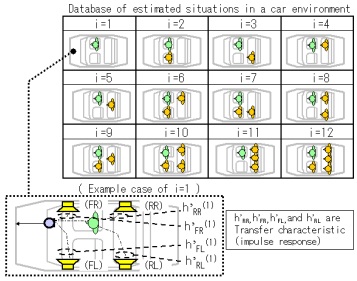

Fig. 2 12 通りの伝達特性

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．

[1] や適応フィルタなど 1ch に信号をまとめる推定では，音響エコー推定結果が十分に収束しない．そこで観測信号 $y^{(i)}$ に対し，音響エコー推定結果 $N'^{(i)}$ を最適化する伝達特性 $h'_{FL}{}^{(i)}$，$h'_{FR}{}^{(i)}$，$h'_{RL}{}^{(i)}$，$h'_{RR}{}^{(i)}$ を選択することを考える．

3 尤度最大化基準を用いた車室内音響エコーキャンセラ

車環境で想定できる複数の伝達特性を元に音響エコーを作成し，観測信号からキャンセルする．実環境と想定環境の伝達特性が一致する場合，キャンセル結果にはクリーン音声のみ存在していると考えられるが，実環境と想定環境の伝達特性にミスマッチがある場合，キャンセル結果にはクリーン音声とエコー推定誤差が存在する．キャンセル後の信号に対して，音響モデルを用いて尤度計算を行い，尤度最大化基準により最適な結果を選択する．

3.1 伝達特性

伝達特性とは，車室内のいろいろな状況を想定し，それぞれの状況下でインパルス応答を測定 [3] した結果とする．

今回は人員配置がそれぞれ異なる 12 通りの車内状況を想定した (Fig. 2)．状況 i において伝達特性を計算した結果を $h'_{FL}{}^{(i)}$，$h'_{FR}{}^{(i)}$，$h'_{RL}{}^{(i)}$，$h'_{RR}{}^{(i)}(i = 1, 2, \ldots, 12)$ とする．

3.2 音声尤度の計算

まず，音声の MFCC(Mel Frequency Cepstrum Coefficient) 特徴量 o を計算する．MFCC は音声データに対し FFT を行い，結果の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．求めた MFCC を用いて，各話者毎の GMM(Gaussian Mixture Model) を学習する．MFCC 特徴 o の音声尤度 $P(o)$ は式 5 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．

$$P(o) = \sum_{w=1}^{W} \lambda_w N(o; \mu_w, \sigma_w) \tag{5}$$

*A study on acoustic echo canceller based on echo estimation with maximum likelihood for multi-loudspeakers in a car environment, by KOGA Kentaro (Fujitsu TEN) , TAKIGUCHI Tetsuya, ARIKI Yasuo (Kobe University)

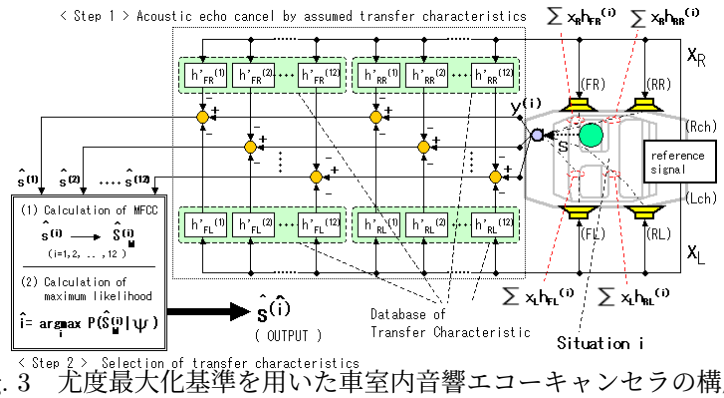

Fig. 3　尤度最大化基準を用いた車室内音響エコーキャンセラの構成図

3.3　尤度最大化基準を用いた車室内音響エコーキャンセラ

構成図を Fig. 3 に示す．ある車内環境 i にて観測されたマイクの観測信号を $y^{(i)}$ とする． $y^{(i)}$ に対し 12 通りの音響エコー $N'^{(1)}, N'^{(2)}, \ldots, N'^{(12)}$ を作成する．次に，それぞれの推定音響エコーを観測信号からキャンセルした結果 $\hat{s}^{(1)}, \hat{s}^{(2)}, \ldots, \hat{s}^{(12)}$ を計算する．

$\hat{s}^{(1)}, \hat{s}^{(2)}, \ldots, \hat{s}^{(12)}$ から，音声の MFCC 特徴量 $\hat{S}_M^{(1)}, \hat{S}_M^{(2)}, \ldots, \hat{S}_M^{(12)}$ を計算する．各話者のモデルを $\psi = \{\lambda, \mu, \sigma\}$ としたとき

$$\hat{i} = \arg\max_i P(\hat{S}_M^{(i)}|\psi) \qquad (6)$$

となる $\hat{i}$ を計算し，このときの $\hat{s}^{(\hat{i})}$ が音声尤度を最大とするキャンセル結果となる．

4　評価実験

実環境で収録した音楽重畳音声信号に対し，提案手法を適用し，学習同定法と比べて SN と音声認識率を改善できることを示す．

Fig. 2 に示す 12 通りの人員配置の状況の実環境で測定した音楽重畳音声信号を $y^{(i)}(i = 1, 2, \ldots, 12)$ とする．状況 o で観測した信号 $y^{(o)}$ に対しては，同じ人員配置 o で測定した伝達特性 $h'^{(o)}$ を用いてキャンセルした結果 $\hat{s}^{(o)}$ を尤度最大化基準で選択することが求められる．$y^{(o)}$ に対して $\hat{s}^{(o)}$ を選択できる割合を，正しい状況の選択率と定義する．

学習同定法によるキャンセラを用いた場合，尤度最大化基準に基づくキャンセラを用いた場合で正しい状況の選択率，SN，認識率を求めたものをそれぞれ Fig. 4，Fig. 5 に示す．Fig. 5 には，元の観測信号 $y^{(i)}$ の SN 平均と認識率，尤度最大化キャンセラによる理想値 (全ての $y^{(o)}(o = 1, 2, \ldots, 12)$ に対し，出力が $\hat{s}^{(o)}$，つまり正しい状況の選択率 100%) の場合の SN，認識率を併記している．学習同定法では 2ch 参照信号を足し合わせて 1ch にした参照信号を入力としている．評価データの条件を Table. 1，アルゴリズムに関する条件は Table. 2 に示す．

尤度最大化基準による正しい環境の選択率 (入力 $y^{(o)}$ に対し，$\hat{s}^{(o)}$ を出力) の話者平均は 79.8% で，全ての話者に対して 100% 正しい環境を選択できているわけではないが，尤度最大化を用いた音響エコーキャンセラによるキャンセル効果は学習同定法と比較して 9.5dB の SN 改善，21.4% の音声認識率改善が見られる．選択率が 100% には至らない部分が理想の場合との SN0.1dB，認識率 4.5% の差になっている．

Table 1　評価データの条件

話者数	5
文章の数	100
信号の周波数	16kHz
インパルス応答を測定した実車	will CYPHA

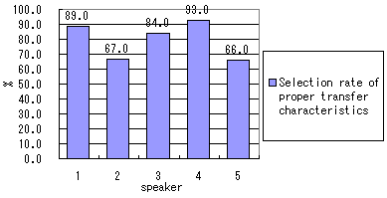

Fig. 4　正しい状況の選択率

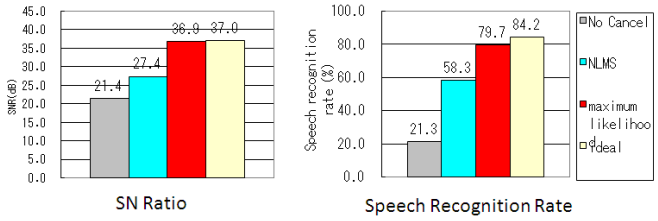

Fig. 5　SN と音声認識率

Fig. 6 に音声認識率と正しい結果の選択率の関係を示す．選択率が上昇すれば，音声認識率も上がることが言える．

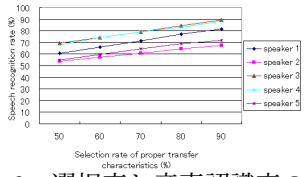

Fig. 6　選択率と音声認識率の関係

5　おわりに

今回，提案手法によって車内の音響エコーを除去し，従来の適応フィルタよりも観測信号の SN を改善し，音声認識率を改善できることを示した．

今後は，キャンセラ側で用意するモデルの削減や，車内の状況を増やしての検証を行いたい．

参考文献

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

Table 2　アルゴリズムに関する条件

フィルタのタップ長さ	h
GMM 学習の文章	1200 sentences
GMM の混合数	32
MFCC の次元数	16
MFCC 特徴抽出のフレーム幅	32ms
MFCC 特徴抽出のシフト幅	8ms