

音声区間検出を用いた音響エコーキャンセラにおける音声歪み低減の試み*

○古賀健太郎 (富士通テン, 神戸大), 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, 車内で音楽がスピーカから出力されている状況下で音声認識を行うと, 音声認識マイクに認識対象の音声と音響エコーが混入し, 音声認識の妨げとなる. そこで, 音響エコーキャンセラによって音響エコーを除去し, SN 比を改善することによって認識率を確保する試みがなされている. しかし, 音響エコーキャンセラの適応フィルタ学習時に認識対象音声が含まれると, 音声歪む問題が存在している.

本稿では, 音響エコーキャンセラと音声区間検出を組み合わせることにより, 適応フィルタを制御し, 音声歪みを低減する方法について検討する.

2 音声歪み

マイク観測信号 $y(t)$ から音響エコーを除去するためには, Fig. 1 のように音響エコーキャンセラによって音楽の原信号 $x(t)$ を元に擬似エコー $x(t)*h(t)$ を推定して $y(t)$ から減算し, 推定誤差 $e(t)$ を基に適応フィルタ係数 $h(t)$ を更新し, 音響エコーの推定精度を高めることを繰り返す. 射影法において, フィルタ係数 $h(t)$ の更新式は式 (1) のようになる [1].

$$h(t+1) = h(t) + \mu\{a_1(t)x(t) + a_2(t)x(t-1)\} \quad (1)$$

$$\begin{bmatrix} a_1(t) \\ a_2(t) \end{bmatrix} = \begin{bmatrix} x(t)^T x(t) & x(t)^T x(t-1) \\ x(t-1)^T x(t) & x(t-1)^T x(t-1) \end{bmatrix}^{-1} \begin{bmatrix} e(t) \\ (1-\mu)e(t-1) \end{bmatrix}$$

しかし, 認識対象音声が発声された場合, マイク入力信号 $y(t)$ には音響エコー $n(t)$ に加え, 認識対象音声 $s(t)$ が含まれる. Fig. 1 の構成では, キャンセル時に音声 $s(t)$ がキャンセルされずに推定誤差 $e(t)$ に含まれるため, この $s(t)$ を含む推定誤差 $e(t)$ をそのまま適応フィルタに学習させてしまう. その結果, 音響エコーのみでなく音声もキャンセルするようにフィルタが更新され, 音響エコーキャンセル時に認識対象音声の欠落が生じる. この結果, 音声認識率が低下する.

本稿では, キャンセル後の推定誤差に対し, 音声区間検出 (VAD, Voice Activity Detection) を行うことでフィルタ更新を制御し, 音声歪みを低減する試みについて述べる. 先行研究として [2] が行われているが, [2] では推定誤差が音声を含む場合にフィルタ係数を過去の値に置き換える手法を用いているのに対し, 本稿では推定誤差に含まれる音声の尤度に応じて, ステップサイズ μ の制御を行うことにより適応フィルタを制御する.

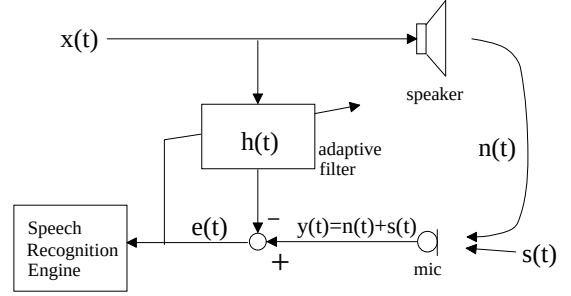


Fig. 1 音響エコーキャンセラ

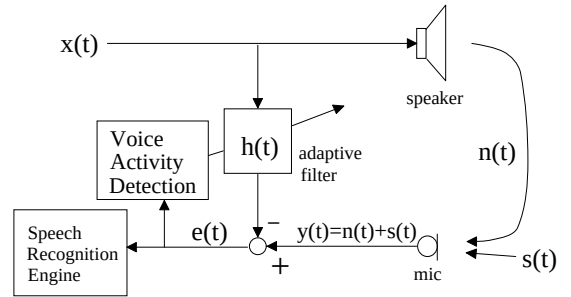


Fig. 2 音響エコーキャンセラと音声区間検出の組み合わせ

3 音声歪みを低減するための音響エコーキャンセラの構成

音声歪みを低減させるための音響エコーキャンセラの構成を Fig. 2 に示す. 推定誤差 $e(t)$ を用いて適応フィルタ $h(t)$ を更新する前に, 音声区間検出を行い, $e(t)$ の尤度を計算する. その結果, $e(t)$ の雑音尤度が低い (音声尤度が高い) と判定された場合には推定誤差 $e(t)$ の適応フィルタへの反映を抑制する.

4 提案手法

4.1 音声区間検出

音声区間検出は MFCC (Mel Frequency Cepstrum Coefficient) の比較により行う. 分析対象となる音響信号の MFCC を計算した上で, (あらかじめ作成していた) 音声の MFCC による GMM (Gaussian Mixture Model) と比較することで, 音声検出が可能になる.

Fig. 2 において, 音響エコーキャンセル前の観測信号 $y(t)$ は音声 $s(t)$ に音響エコー $n(t)$ が重畳した状態であり, $s(t)$ と $n(t)$ 両方の GMM を用意する必要がある. しかし, $n(t)$ の GMM はスピーカから出力される音楽によって変更する必要があるため, 実用的では無いと言える. キャンセルした後の推定誤差 $e(t)$ は音声に音響エコーの消し残しが重畳した状態であるが, この音響エコーの消し残しは Fig. 3 に示す通り, 定常なホワイトノイズに比較的近い波形になっている. また音声 $s(t)$ は Fig. 4 に示す通り, 非定常な

*A study on reduction of voice distortion in acoustic echo canceler with voice activity detection, by KOGA Kentaro (Fujitsu TEN and Kobe University), TAKIGUCHI Tetsuya, ARIKI Yasuo (Kobe University)

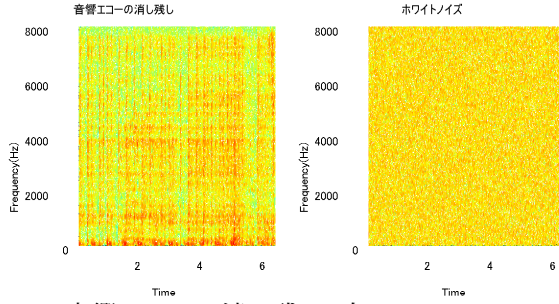


Fig. 3 音響エコーの消し残し (左) とホワイトノイズ (右) のスペクトログラム

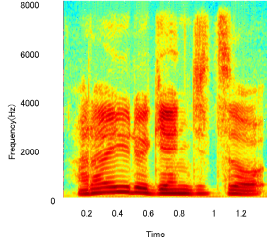


Fig. 4 音声のスペクトログラム

波形になっている。そこで定常なホワイトノイズを用いて GMM を学習し、推定誤差 $e(t)$ の定常性を調べることで、スピーカからどのような音楽が出力されていても精度の高い音声区間検出が可能と考える。

4.2 適応フィルタの制御方法

ホワイトノイズ GMM の尤度 θ に基づいて、適応フィルタ更新式のステップサイズ μ を調整することにより適応フィルタを制御する。 μ が大きければ誤差 $e(t)$ の適応フィルタへの反映度合いは大きくなり、 μ が小さければ誤差の反映度合いは小さくなる。 μ は次式 (2) に従って決定する。 $(\theta_2 < \theta < \theta_1)$ において、 μ は θ についての一次関数である

$$\begin{aligned} \mu &= \mu_{\max} & (\theta \geq \theta_1) \\ \mu &= \mu_{\min} & (\theta \leq \theta_2) \\ \mu &= \frac{\mu_{\max} - \mu_{\min}}{\theta_1 - \theta_2} (\theta - \theta_2) + \mu_{\min} & (\theta_2 < \theta < \theta_1) \end{aligned} \quad (2)$$

式 (2) において θ_1, θ_2 は $\theta_1 > \theta_2$ を満たす閾値で、 $\mu_{\max}, \mu_{\min}$ はそれぞれステップサイズの上限值、下限値である。

5 評価実験

5.1 実験条件

評価対象とする音響エコーキャンセラは

1. 学習同定法, 音声区間検出無し
2. 射影法, 音声区間検出無し
3. 射影法, 音声区間検出あり (音楽消し残しの GMM を使用)
4. 射影法, 音声区間検出あり (ホワイトノイズの GMM を使用)

の 4 通り (タップ数はそれぞれ 256, ステップサイズは $(\mu_{\max}, \mu_{\min}) = (1.0, 0.1)$, 閾値は音楽消し残しの GMM を用いた場合が $(\theta_1, \theta_2) = (-7.0, -11.0)$, ホワイトノイズの GMM を用いた場合が $(\theta_1, \theta_2) = (-13.0, -30.0)$) とし, Table. 1 に示す評価データを

Table 1 実験に用いた評価データ

発話データ	ASJ 男性話者 2 名, 女声話者 2 名, 各 50 発話
重畳音楽	3 曲
残響時間	120ms
原信号-マイクの遅延時間	3ms
サンプリング周波数	16kHz
発話までの時間	2.7s

Table 2 実験結果

評価対象	ϕ
学習同定法, 音声区間検出無し	16.86
射影法, 音声区間検出無し	13.29
射影法, 音声区間検出あり (音楽消し残しの GMM を使用)	12.42
射影法, 音声区間検出あり (ホワイトノイズの GMM を使用)	12.27
[理想値] 射影法, 音声区間検出あり (手動切り出し)	11.79

用いる。キャンセラによって生じた音声の歪みによって評価する。

歪みの度合いは音声のケプストラム距離を比較することによって行う。音響エコーが重畳していない音声の時間 t におけるケプストラム値を $c_1(t)$, 音響エコーが重畳している状態でエコーをキャンセルした後の音声の時間 t におけるケプストラム値を $c_2(t)$ とした場合, 次式 (3) によって求めた値 ϕ が大きいほど, 歪みが大きいものとして判断する。

$$\phi = \frac{\sum_{t=0}^{N-1} |c_1(t) - c_2(t)|}{N} \quad (3)$$

式 (3) において, $t = 0 \sim N - 1$ は音声区間である。

5.2 実験結果

実験結果を Table. 2 に示す。Table. 2 では, 音声区間の始点と終点を手動で与え, 音声区間で $\mu = \mu_{\min}$, 非音声区間で $\mu = \mu_{\max}$ とした場合の結果を理想値として示している。音響エコーキャンセラのアルゴリズムを学習同定法から射影法に変更することにより, ϕ の値 (ケプストラム距離) が 3.5 程度減少し, 歪みが改善されていることが分かる。また, 射影法に GMM による音声区間検出を追加することにより, ϕ の値がさらに 1.0 程度減少する。これは, 提案手法により音声歪みが改善できることを示している。

6 おわりに

今回, ホワイトノイズの GMM を用いた音声区間検出と音声尤度に応じた適応フィルタの制御方法を提案した。両手法を音響エコーキャンセラに組み込み, 評価実験により提案手法の有効性を示し, 音声歪みが改善できることを示した。今後は音響エコー重畳下において, 実際に音声認識率を改善できるか調査する。

参考文献

- [1] 大賀ら共著, “音響システムとデジタル処理,” 電子情報通信学会, 1995.
- [2] 庄境, 中村, 鹿野, 信学論, D81-D-, pp.1074-1083, 1998.